Hindawi
Security and Communication Networks
Volume 2020, Article ID 8849902, 8 pages
https://doi.org/10.1155/2020/8849902

Research Article

An Antiforensic Method against AMR Compression Detection

Diqun Yan ,[1,2] Xiaowen Li,[1] Li Dong,[1] and Rangding Wang[1]

[1]*Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, China*
[2]*Guangdong Key Laboratory of Intelligent Information Processing and Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen 518060, China*

Correspondence should be addressed to Diqun Yan; yandiqun@nbu.edu.cn

Received 24 May 2020; Revised 23 July 2020; Accepted 19 August 2020; Published 2 September 2020

Academic Editor: Zhihua Xia

Adaptive multirate (AMR) compression audio has been exploited as an effective forensic evidence to justify audio authenticity. Little consideration has been given, however, to antiforensic techniques capable of fooling AMR compression forensic algorithms. In this paper, we present an antiforensic method based on generative adversarial network (GAN) to attack AMR compression detectors. The GAN framework is utilized to modify double AMR compressed audio to have the underlying statistics of single compressed one. Three state-of-the-art detectors of AMR compression are selected as the targets to be attacked. The experimental results demonstrate that the proposed method is capable of removing the forensically detectable artifacts of AMR compression under various ratios with an average successful attack rate about 94.75%, which means the modified audios generated by our well-trained generator can treat the forensic detector effectively. Moreover, we show that the perceptual quality of the generated AMR audio is well preserved.

1. Introduction

AMR audio codec [1] is one of the most popular audio codec standards, which is optimized for speech signals and encodes narrowband (200–3400 Hz) signals, with sampling frequency of 8000 Hz [2]. As more and more AMR audio appears as evidence in the forensics scene, it is of extreme importance to verify their integrity [3]. Generally, to manipulate an AMR audio, attacker should decompress it into raw waveform first and then do the forgery operations and decompress it into AMR format. The double compressed audio becomes questionable because the manipulated audio is always through the double compression. In the past decade, many forensic techniques have been proposed to detect compression history of AMR audios based on traditional methods [3–5] and deep learning methods [2, 6, 7]. To represent the difference of single compressed audios and double compressed audios, traditional AMR compression detection techniques rely on low-level acoustic features such as sub-band energy and linear prediction coefficients (LPCs), which acquire professional acoustic knowledge. Recently, deep learning methods are gaining popularity in forensic research studies, which can capture the highly complex feature from a raw sample by training large-scale sample data with a neural network.

However, as many forensic techniques are proposed to detect the integrity of digital file, some antiforensic methods have also been proposed to expose the shortcomings and weakness of existing forensic techniques and thus help investigators better address the weaknesses and improve their forensic techniques. For example, Fontani et al. [8] firstly presented an antiforensic method of median filtering (MF), which made MF images undetectable by the MF detectors [9–11] while keeping the image quality in a good PSNR. Luo et al. [12] applied a GAN framework to improve the quality of JPEG images and fool the JPEG compression detectors successfully. Chen et al. [13] used the legacy traces of a designated camera to generate a forged image that can deceive the existing camera identification techniques successfully. Kim et al. [14] adopted a deep convolutional neural network (DCNN) to remove the forensic traces from MF images and effectively recover the MF images visually similar to the original image. Li et al. [15] modified the forensic traces using a data-driven manner to mislead the results of three advanced audio source identification techniques [16–18].

These antiforensic methods have a little consideration about exposing the weakness of the robustness of AMR compression detection. Generally speaking, as more and more AMR audio appears as evidence in forensics scene, it is important to help the investigators to address the weakness of AMR compression detectors. Therefore, in this paper, we propose an antiforensic method utilizing a GAN framework which comprised of two networks: a generator and a discriminator. The generated data can statistically model the distribution of real data [19]. To improve the perceptual quality of the double compressed audio and remove the artifacts introduced by AMR compression procedure, we adopt the GAN to modify the double compressed audios to avoid forensic detection. For building our antiforensic attack, we design the architecture of GAN and the loss functions. In particular, three state-of-the-art detectors of AMR compression have been selected as the attack target to evaluate the performance of our method.

The rest of this paper is organized as follows. In Section 2, we introduce the related work of forensic method of AMR compression and the GAN framework. The detail of our proposed GAN framework has been provided in Section 3. Section 4 presents the experimental settings and extensive experiments against three AMR compression detectors. Conclusions are given in Section 5.

2. Related Work

In this section, we briefly introduce three advanced detection methods, which are considered as attack targets. Additionally, the GAN framework is also briefly reviewed.

2.1. Detection of AMR Compression. In general, traditional detection of AMR compression consists of two primary steps: feature extraction and model classification.

As the first work of the detection of AMR compression, Shen et al. [3] used the traditional acoustic features including average sub-band frequency energy ratio, average low-frequency sub-band energy ratio, bispectrum features, and linear prediction spectrum to represent the difference caused by AMR compression. And a standard SVM modelling technique was employed for classification. They achieved an accuracy about 87% for detecting the single compressed audio from the double one.

In [2], Luo et al. adopted an autoencoder network for automatic feature extraction. They demonstrated that the deep features differ greatly between the single compressed audio and the double one which were extracted from a well-trained autoencoder. And they designed a majority voting strategy for classification.

In [6], the authors delved into the stack autoencoder (SAE) network for obtaining better deep features in the AMR compression forensic task. Then, they applied a universal background model-Gaussian mixture model (UBM-GMM) for the identification of compression history. They improved the classification accuracy to 98% on the TIMIT [20] database.

2.2. Generative Adversarial Network. The generative adversarial network (GAN) was firstly proposed by Goodfellow et al. [21] for generating realistic images. In GAN, two networks are training against each other in a min-max two-player game. In their iterative training, the purpose of generator G is to capture the distribution of real data and that of discriminator D is to classify a sample that came from the real database rather than generated by G. The generator G tries to maximize the probability of making the discriminator D mistakenly classify the generated data as real, while the discriminator guides the generator to produce a more realistic sample. Generally, the adversarial training process can be denoted as a min-max game and it will be optimized by the loss function L_{GAN} as follows:

$$L_{\text{GAN}} = \sum_{x} [\log (D(x))] + \sum_{z} [\log (1 - D(G(z)))], \quad (1)$$

where x denotes the real data and z denotes the random noise similar to x after the adversarial training of the generator G and discriminator D. In the training process, the purpose of G is to minimize the loss value while that of D is to maximize it.

Recently, GAN has gained growing popularity in various fields because of its effective generative capability. In this work, the GAN framework is assumed as the reverse procedure of AMR compression to improve the perceptual quality of double compressed audio and remove the forensic artifacts. Specifically, the generator and the discriminator can be regarded as an antiforensic model and AMR compression detector, respectively. Hence, the adversarial concept is suitable for antiforensic task in the AMR compression detection.

3. Proposed Antiforensic Framework

In this section, we briefly introduce three advanced detection methods, which are considered as attack targets. Additionally, the GAN framework is also briefly reviewed.

x_{db} is firstly sent into the generator to get a falsified audio x'_{db}. x'_{db} and x_{org} selected from the uncompressed audio are further fed into the discriminator for classification. Then, by freezing the parameters of discriminator, the loss from D will be fed back to G, which is represented by the dotted lines.

3.1. Overall Architecture. The overall goal of our attack is to remove the artifacts left by the AMR compression so that the resultant audio can fool the detectors. To deploy a successful attack, the generated audio should be decompressed back to AMR format because many investigators only accept the AMR file before the detection. Thus, the generated audio x'_{db} must statistically model the distribution of original audio x_{org} so that the decompressed ones will be similar to the single compressed audio x_{sg}.

As shown in Figure 1, the proposed framework consists of a generator G and a discriminator D. To remove the artifacts left by the compression, G is used to generate the falsified audio x'_{db} by adding a generated perturbation into x_{db}. The discriminator D is designed to distinguish an original audio x_{org}, which is never through compression from a falsified audio x'_{db}. In the adversarial training of G and D, G is encouraged to learn how to minimize the difference between x'_{db} and x_{org} and optimize the parameters to achieve a better performance in generating good perceptual quality of x'_{db}.

3.2. Architecture of Proposed Framework

3.2.1. Generator. Generator is used to generate the anti-forensic audios. In this framework, we use the SEGAN [22] as a reference architecture to design our adversarial network, which has been effectively applied in speech enhancement. As shown in Figure 2, the generator gets x_{db} (size = 1 × 8000) as the input and consists of 7 convolutional groups and 7 corresponding deconvolutional groups.

Each convolutional group includes a convolutional layer with 64 filters with 1 × 30 kernels and stride = 2, whereafter a batch normalization (BN) layer which can stabilize the training process makes the generated audios more realistic. And the Leaky-ReLU is chosen as the activation function. The deconvolutional group is constituted of a deconvolutional layer which is set up as the convolutional group, followed by a BN layer and ReLU as the activation function. To reconstruct the details of audio and diminish the loss when information flows through convolutional and deconvolutional groups, we apply the skip connection in the generator, which can make the convolutional group's output flow to its corresponding deconvolutional group. The skip connection can make the generator have a better performance, as the gradients can flow deeper through the skip connection without suffering much vanishing [23]. And the sigmoid activation is added to restrict the output for classification.

3.2.2. Discriminator. Since the key advantage of GAN is iterative training to obtain a better performance in generating samples, it seems that the architecture of D is a very important constraint to our framework. The discriminator D is intended to classify x_{org} and x'_{db} and force the generated audios to deceive the detector. Hence, the discriminator must perform well in distinguishing x_{org} and x'_{db}. Therefore, we build a CNN architecture for D. As shown in Figure 3, the discriminator is designed as a compression detector based on CNN. It comprises 6 convolutional groups and is followed by a group consisting of a global average pool layer. At the end of the network, a dense layer coupled with a softmax activation function is placed to output the categorical probability.

Before the iterative training, we firstly test the capability of the designed discriminator to distinguish the original audio x_{org} from double compressed audio x_{db}. Then, we test the capability of the designed discriminator in a sub-dataset including 6000 original audios selected from TIMIT database and its double compressed audios with a compression bit rate randomly selected from {4.75 kbps, 5.15 kbps, 5.9 kbps, 6.7 kbps, 7.4 kbps, 7.95 kbps, 10.2 kbps, 12.2 kbps}. The sub-dataset was then divided into training (70%) and validation (30%). The accuracy of the discriminator model is shown in Figure 4. It is observed that our designed discriminator achieves a good performance.

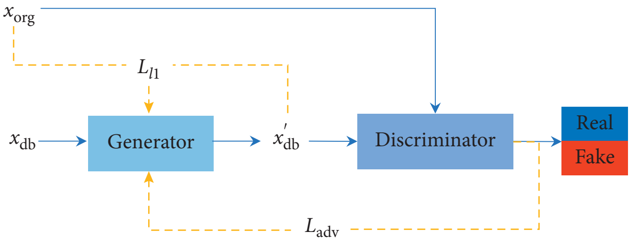

Figure 1: Overall structure of the proposed framework.

3.3. Loss Functions. In this section, we demonstrate the loss functions for the two networks. To achieve the goal of antiforensics, the generator G should be capable to learn how to minimize the difference of the modified double compressed audio x'_{db} and the original audio x_{org}, while maintaining an acceptable perceptual quality. In this work, we define the loss of generator L_G as

$$L_G = \alpha L_{l1} + \beta L_{\text{adv}}, \tag{2}$$

where L_{l1} represents the perceptual loss of x'_{db}, L_{adv} denotes the adversarial loss calculated from D, and α, β are the weights to balance the importance of L_{l1} and L_{adv}.

Considering that the attack needs to introduce lesser perceptual artifacts to improve the forensic undetectability, we employ the perceptual loss L_{l1} for improving the quality of x'_{db}. L_{l1} is defined as

$$L_{l1} = \sum_{i=1}^{N} \left(x_{\text{org}_i} - G\left(x_{\text{db}_i} \right) \right), \tag{3}$$

where $G(\cdot)$ presents the output of G, and N and i represent the batch size and the position of x in this batch, respectively.

Then, the adversarial loss L_{adv} is designed to force G to have a better performance in the iterative training. We define the L_{adv} as

$$L_{\text{adv}} = \sum_{i=1}^{N} \log\left(1 - D\left(G\left(x_{\text{db}_i} \right) \right) \right), \tag{4}$$

where $G(\cdot)$ denotes the class probabilities of the modified audio x'_{db} calculated by D.

In this adversarial task, for forcing G to modify x_{db} similar to x_{org}, D should have the ability to detect the original audio correctly from the decompressed x_{org} or the generated x'_{db}. Therefore, L_D is defined as follows:

$$L_D = \sum_{i=1}^{N} \left[\log\left(1 - D\left(G\left(x_{\text{db}_i} \right) \right) \right) - \log\left(D\left(x_{\text{org}_i} \right) \right) \right]. \tag{5}$$

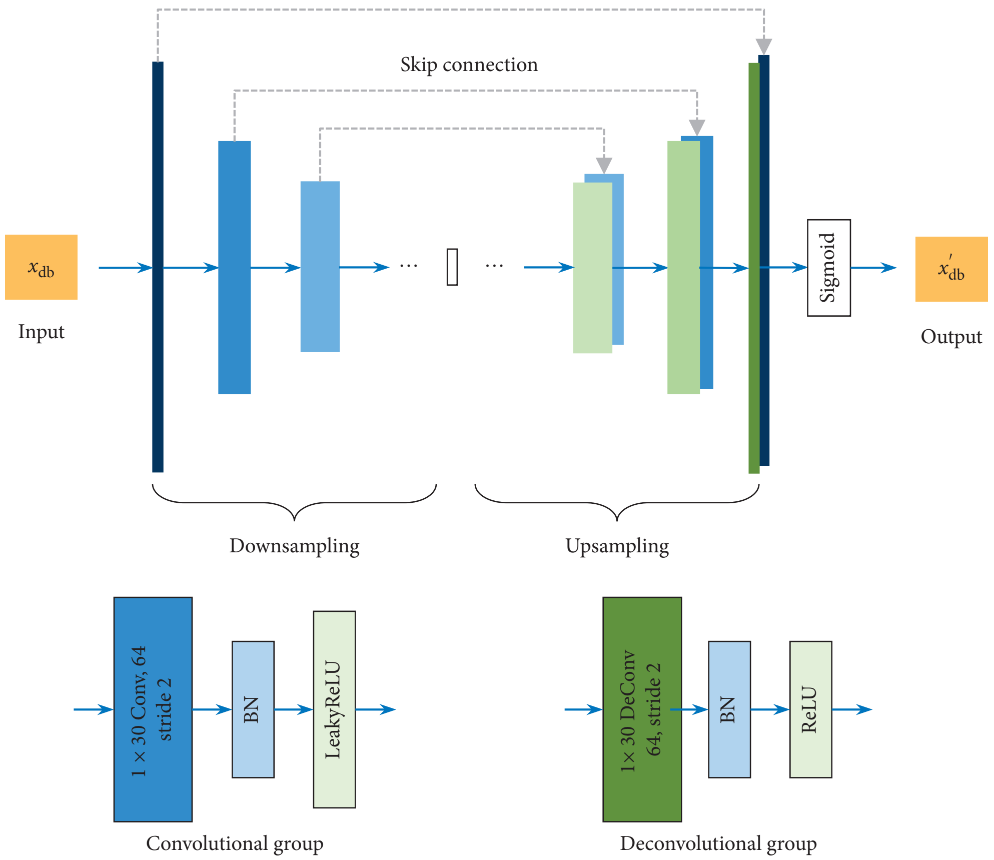

Figure 2: Architecture of the generator.

4. Experimental Results

In this section, we evaluate our antiforensic method against the three advanced forensic techniques [2, 3, 6]. First, we create an audio database designed especially for the experiment. Then, successful attack rate (SAR) is used to perform the forensic undetectability of our antiforensic audios and perceptual evaluation of speech quality (PESQ) [24] is adopted to present the quality of our antiforensic audios.

4.1. Database. TIMIT [20] is a typical speech database which consists of 630 speakers from different dialects of American English (192 females and 432 males) and each speaker reads ten sentences which are approximately three seconds. At first, to build the forensic database, we use the AMR codec to obtain the single compression audio from TIMIT database, with a random compression bit rate selected from {4.75 kbps, 5.15 kbps, 5.9 kbps, 6.7 kbps, 7.4 kbps, 7.95 kbps, 10.2 kbps, 12.2 kbps}. Then, we decode and recompress the AMR audios to get the double compressed AMR audio with random bit rates also selected from 4.75 to 12.2 kbps.

In the experiments, we first split those audios into 1 s clips and randomly divide those clips into the train set and test set. Therefore, we obtain 12000 1 s training audios and 6900 testing audios. Then, three detectors [2, 3, 6] are trained using the train set, and the average detection accuracies in test set are 87.52%, 92.60%, and 98.54%, respectively, which are essentially in agreement with the results reported in their works.

4.2. Experimental Setup and Evaluation Metrics

4.2.1. Experimental Setup. We train our network on patch sized audios with the pairs of sets: $\{x_{org}, x_{db}\}$. Considering the audio might be split into different sizes by the investigator before the detection, we stitch all 1 s audios x'_{db} to obtain more audios with difference sizes, including 13800 0.5 s clips, 6900 1 s clips, 3450 2 s clips, and 2300 3 s clips. Then, we compressed x'_{db} back to AMR format with random bit rates chosen from 4.75 to 12.2 kbps.

Adam [25] is adopted as the optimizer with a learning rate of 1×10^{-4} for G and 5×10^{-6} for D. Before the iterative training, we perform the generator training with batch size = 64 and weight terms of $\alpha = 1000$ and $\beta = 0$ for 5 epochs. Then, G and D are trained iteratively for 30 epochs with weight terms of $\alpha = 1000$ and $\beta = 1$, with an iteration ratio of 1 : 5, which gives the discriminator more iterations to get a better performance.

4.2.2. Evaluation Metrics. The successful attack rate (SAR) is used as the evaluation metric, which could well represent the forensic undetectability of our antiforensic audio. We define the SAR as

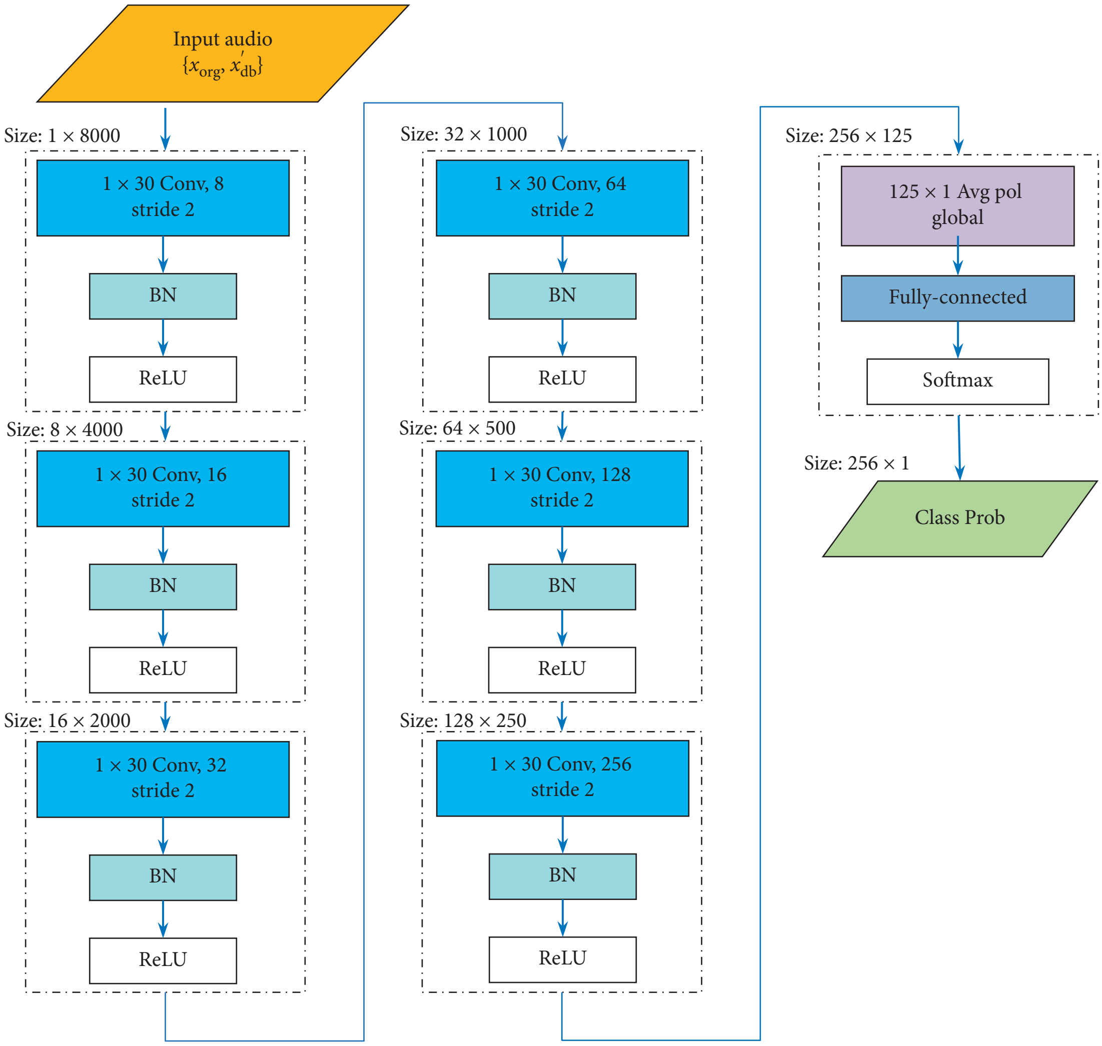

Figure 3: Architecture of the discriminator.

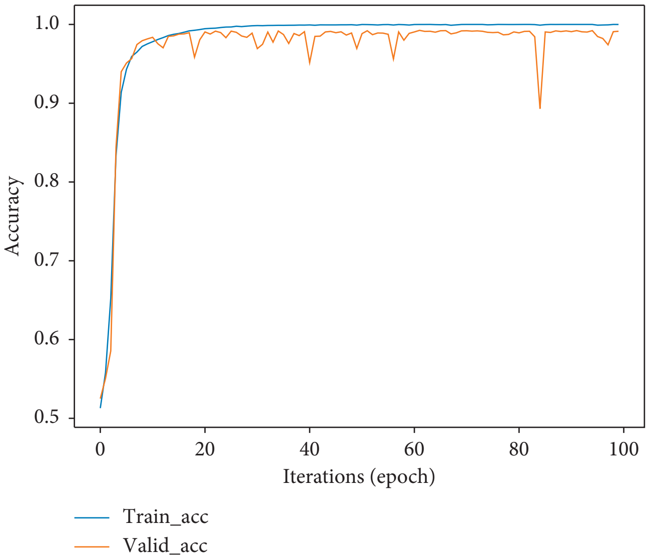

Figure 4: Accuracy in the training and validation performance of the designed discriminator.

Table 1: Successful attack rate (SAR) of antiforensic double AMR audio clips (%).

Size (s)	Method	Bit rates (kbps)								Average
		4.75	5.15	5.9	6.7	7.4	7.95	10.2	12.2	
0.5	[2]	97.07	93.96	98.41	95.28	96.52	97.40	96.88	97.59	96.63
	[3]	95.82	92.74	93.95	94.26	93.53	96.83	91.75	94.85	94.20
	[6]	94.95	96.49	92.60	97.25	96.55	91.64	98.28	93.75	95.69
1	[2]	97.66	98.05	98.15	91.72	95.78	98.42	92.17	94.82	95.82
	[3]	98.33	97.94	98.30	93.69	96.72	96.59	96.08	91.01	96.21
	[6]	94.05	96.88	92.20	90.63	86.08	91.26	91.73	90.23	91.63
2	[2]	96.12	90.95	97.17	91.56	95.86	98.49	98.12	97.32	95.70
	[3]	96.23	99.35	98.59	93.41	92.32	91.34	96.91	93.27	95.15
	[6]	94.77	92.88	95.26	89.14	89.00	95.25	90.19	96.85	92.91
3	[2]	98.86	96.06	93.32	98.05	95.35	93.36	85.95	96.74	94.71
	[3]	97.57	96.37	94.72	98.45	97.51	98.05	96.74	93.19	96.68
	[6]	93.36	96.61	94.72	92.20	93.86	91.56	94.14	92.53	93.62

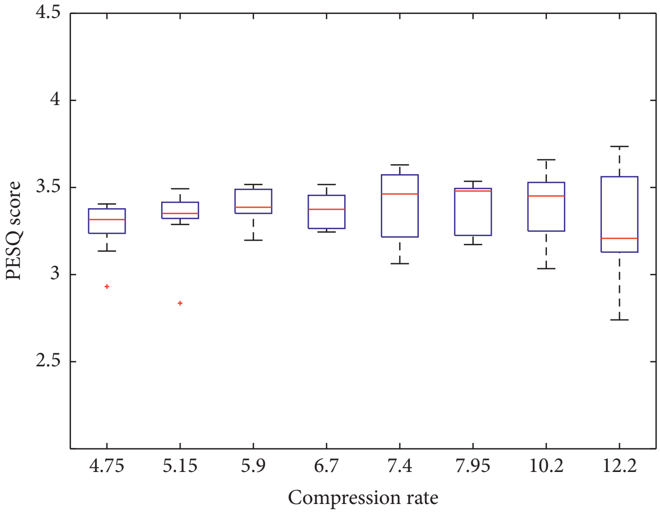

Figure 5: PESQ score of box plots calculated by antiforensic audios and single compressed audios via same compression bit rate.

$$L_D = \frac{1}{N}\sum_{i=1}^{N}\left(F\left(\mathrm{AF}_{x_{\mathrm{db}_i}}\right)\right), \tag{6}$$

where AF_x_{db} represents the audio decompressed with each bit rate selected from 4.75 to 12.2 kbps and $F(\cdot)$ is the classification result of forensic detector, that is, $F(\mathrm{AF}_{x_{\mathrm{db}}}) = 1$ while AF_x_{db} has been misclassified as x_{sg} and 0 otherwise.

Meanwhile, we apply the PESQ to test the perceptual quality of the antiforensic audio AF_x_{db}. PESQ is an industry-standard methodology for the assessment of speech quality. The range from −0.5 to 4.5 is the default PESQ score range, and higher score means better perceptual quality.

4.3. Experimental Performance and Analysis. We perform our attack on three advanced forensic methods [2, 3, 6]. Specifically, for each clip in the testing set, we generate a copy of it with the well-trained generator and then decompress the copy with eight different bit rates.AF_x_{db}. Finally, three trained detectors are used to classify our antiforensic audios.

As shown in Table 1, the experimental results are in line with expectations. The antiforensic audios AF_x_{db} can significantly deceive the three advanced AMR compression detectors, and the SAR of AF_x_{db} is significantly achieved with an average rate about 94.71% which means the forensic techniques cannot distinguish the antiforensic audios correctly. Obviously, our method can significantly make x_{db} undetectable by the forensic techniques.

To measure the quality of our antiforensic audios, we compute the PESQ score of AF_x_{db} comparing the original audios. As shown in Figure 5, it is obvious that our antiforensic audios can retain good perceptual quality and the PESQ values of most AF_x_{db} are over 3.3 compared with x_{sg}, which means that our method can improve the perceptual quality of x_{db} while achieving the antiforensic purpose. Figure 6 shows the spectrograms of an original audio x_{org} from the test set, its x_{sg} and x_{db}, and its antiforensic audio AF_x_{db} compressed with random bit rates. Compared with x_{sg}, x_{db} presents fewer losses of content in the high frequency than AF_x_{db}, and AF_x_{db} is similar to x_{sg}.

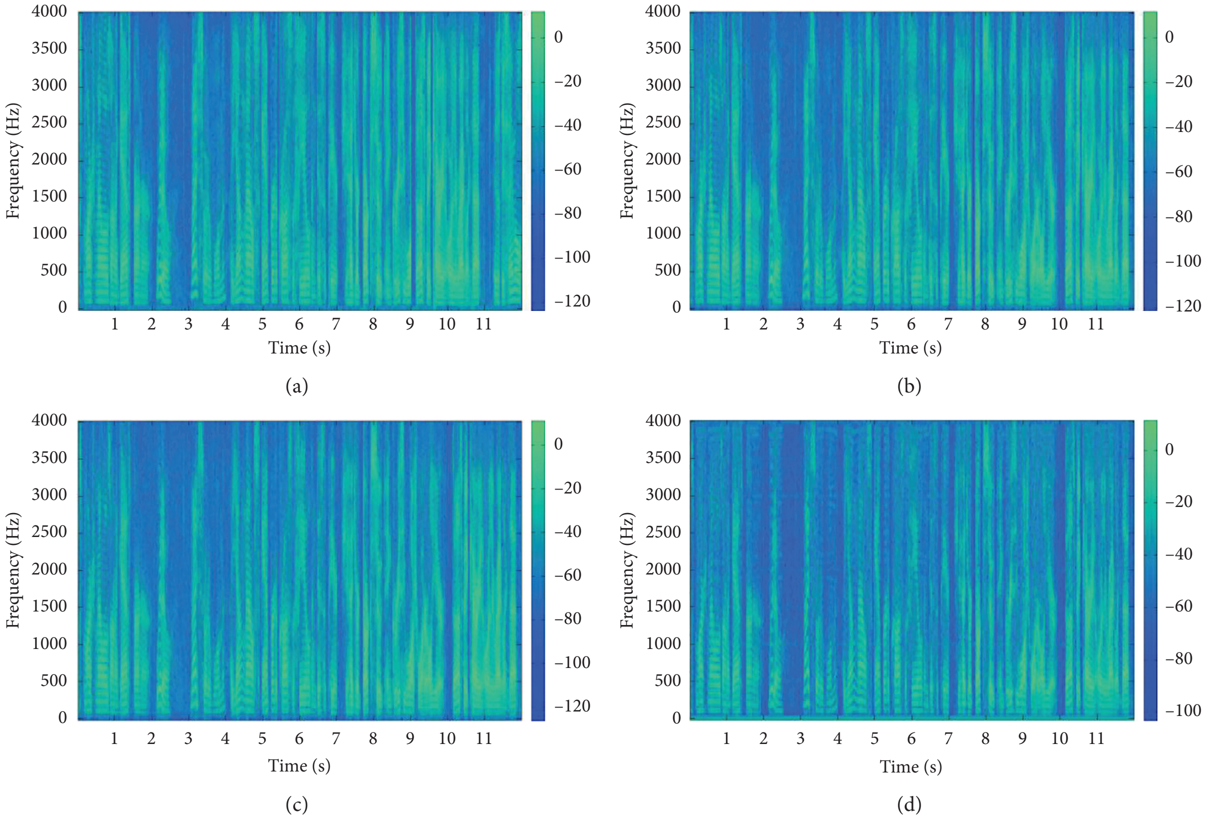

Figure 6: (a) Spectrogram of original audio. (b) Spectrogram of single compressed audio. (c) Spectrogram of double compressed audio. (d) Spectrogram of antiforensic audio generated by a well-trained generator.

5. Conclusion and Future Work

In this paper, we have proposed a new method to prove the weakness of the forensic detectors of AMR compression. To do this, we developed a GAN framework for the removal of AMR compression artifacts. Unlike the conventional antiforensic methods, our method can retain good perceptual quality with a better antiforensic capability in a data-driven manner. Through extensive experiments, the results demonstrate that the antiforensic double compressed audio can effectively avoid the detection of existing AMR compression methods with an average SAR about 94.75%, while retaining good perceptual quality.

However, there are still many remaining problems in the competition of forensics and antiforensics. In the future, we plan to consider the robustness of the forensic approach of AMR compression, i.e., whether adversarial framework could obtain a robust discriminator which can detect the antiforensic audios correctly by a well-trained generator or other attack strategy while distinguishing the double compressed audios from single compressed audios successfully.

Data Availability

The TIMIT dataset used to support the findings of the study is public and available at https://catalog.ldc.upenn.edu/LDC93S1.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study was supported in part by the National Natural Science Foundation of China under grant nos. U1736215, 61672302, and 61901237, in part by the Zhejiang Natural Science Foundation under grant nos. LY20F020010 and LY17F020010, and in part by the K. C. Wong Magna Fund of Ningbo University.

References

[1] B. Bessette, R. Salami, R. Lefebvre et al., "The adaptive multirate wideband speech codec (AMR-WB)," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 8, pp. 620–636, 2002.

[2] D. Luo, R. Yang, and J. Huang, "Detecting double compressed AMR audio using deeplearning," in *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2669–2673, Florence, Italy, May 2014.

[3] Y. Shen, J. Jia, and L. Cai, "Detecting double compressed AMR-format audio recordings," in *Proceedings of the 10th Phonetics Conference. China (PCC)*, pp. 1–5, Shanghai China, April 2012.

[4] J. Sampaio and F. Nascimento, “Double compressed AMR audio detection using linear prediction coefficients and support vector machine,” in *Proceedings of the 22th Brazilian Conference on Automation*, João Pessoa, Brazil, September 2018.

[5] J. F. P. Sampaio and F. A. D. O. Nascimento, “Detection of AMR double compression using compressed-domain speech features,” *Forensic Science International: Digital Investigation*, vol. 33, Article ID 200907, 2020.

[6] D. Luo, R. Yang, B. Li, and J. Huang, “Detection of double compressed AMR audio using stacked autoencoder,” *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 2, pp. 432–444, 2016.

[7] K. Valanchery, “Analysis of different classifier for the detection of double compressed AMR audio,” *International Journal of Advance Research, Ideas and Innovations in Technology*, vol. 4, pp. 98–107, 2018.

[8] M. Fontani and M. Barni, “Hiding traces of median filtering in digital images,” in *Proceedings of the 2012 20th European Signal Processing Conference (EUSIPCO)*, IEEE, Bucharest, Romania, pp. 1239–1243, August 2012.

[9] M. Kirchner and J. Fridrich, “On detection of median filtering in digital images,”in Proceedings of the SPIE Electronic Image, Security, Steganography, vol. 7541, pp. 1–6, Watermarking on Multimedia Contents, San Jose, CA, USA, August 2010..

[10] G. Cao, Y. Zhao, R. Ni, L. Yu, and H. Tian, “Forensic detection of median filtering in digital images,” in *Proceedings of the 2010 IEEE International Conference on Multimedia and Expo*, pp. 89–94, Singapore, July 2010.

[11] H.-D. Yuan, “Blind forensics of median filtering in digital imagesfiltering in digital images,” *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 4, pp. 1335–1345, 2011.

[12] Y. Luo, H. Zi, Q. Zhang, and X. Kang, “Anti-forensics of JPEG compression using generative adversarial networks,” in *Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 952–956, IEEE, Rome, Italy, September 2018.

[13] C. Chen, X. Zhao, and M. C. Stamm, “Mislgan: an anti-forensic camera model falsification framework using a generative adversarial network,” in *Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 535–539, Athens, Greece, October 2018.

[14] D. Kim, H.-U. Jang, S.-M. Mun, S. Choi, and H.-K. Lee, “Median filtered image restoration and anti-forensics using adversarial networks filtered image restoration and anti-forensics using adversarial networks,” *IEEE Signal Processing Letters*, vol. 25, no. 2, pp. 278–282, 2018.

[15] X. Li, D. Yan, L. Dong, and R. Wang, “Anti-forensics of audio source identification using generative adversarial networkfication using generative adversarial network,” *IEEE Access*, vol. 7, pp. 184332–184339, 2019.

[16] C. Hanilci, F. Ertas, T. Ertas, and O. Eskidere, “Recognition of brand and models of cell-phones from recorded speech signals,” *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 625–634, 2012.

[17] C. Kotropoulos and S. Samaras, “Mobile phone identification using recorded speech signals,” in *Proceedings of the 2014 19th International Conference on Digital Signal Processing*, pp. 586–591, IEEE, Hong Kong, China, August 2014.

[18] D. Luo, P. Korus, and J. Huang, “Band energy difference for source attribution in audio forensics,” *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 9, pp. 2179–2189, 2018.

[19] S. Chintala, E. Denton, M. Arjovsky, and M. Mathieu, How to train a GAN? tips and tricks to make GANs work (2017), 2016.

[20] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, *Getting Started with the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database*, Vol. 107, National Institute of Standards and Technology (NIST), Gaithersburgh, MD, USA, 1988.

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., “Generative adversarial nets,” in *Advances in Neural Information Processing Systems*, pp. 2672–2680, Université De Montréal, Montreal, Canada, 2014.

[22] S. Pascual, A. Bonafonte, and J. Serra, “SEGAN: speech enhancement generative adversarial network,” 2017, https://arxiv.org/abs/1703.09452.

[23] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[24] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “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, vol. 2, pp. 749–752, IEEE, Salt Lake City, UT, USA, February 2001..

[25] D. P. Kingma and J. Ba, “Adam: a method for stochastic optimization,” 2014, https://arxiv.org/abs/1412.6980.