Improving Audio Steganalysis Using Deep Residual Networks

Zhenyu Zhang[1,2], Xiaowei Yi[1,2], and Xianfeng Zhao[1,2](✉)

[1] State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
zhaoxianfeng@iie.ac.cn

[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100093, China

Abstract. In this paper, we propose an effective audio steganalysis scheme based on deep residual convolutional networks in the temporal domain. Firstly, considering the weak difference between cover and stego, a high pass filter is adopted in the proposed network which is used to calculate the residual map of the audio signal. Then, comparing with convolutional neural networks (CNNs) based audio steganalysis in recent studies, the deeper network structure and complicated convolutional modules are considered to capture the complex statistical characteristic of steganography. Finally, batch normalization layers and shortcut connections are applied to decrease the dangers of over-fitting and accelerate the convergence of back-propagation. In the experiments, we compared the proposed scheme with CNNs based and hand-crafted features based audio steganalysis methods to detect the various steganographic algorithms on speech and music audio clips respectively. The experimental results demonstrate that the proposed scheme is able to detect multiple state-of-the-art audio steganographic schemes with different payloads effectively and outperforms several recently proposed audio steganalysis methods.

Keywords: Audio steganalysis · Deep residual network · Adaptive steganography · Temporal domain

1 Introduction

Modern steganography is a science and art of covert communication that slightly changes the original digital media in order to hide secret messages without drawing suspicions from the steganalyzers [4,10]. Corresponding to the development of steganographic techniques, the steganalysis with the aim of revealing the presence of hidden messages in digital media has also been made considerable progress. Recently, many researches about steganalysis have been reported. However, most existing steganalytic methods are mainly dependent on the high-dimensional steganalysis features and supervised classifiers, which can not adapt itself to the various steganographic algorithms. In this paper, we introduce a deep

H. Wang et al. (Eds.): IWDW 2019, LNCS 12022, pp. 57–70, 2020.
https://doi.org/10.1007/978-3-030-43575-2_5

learning method over audio steganalysis, which achieve better detection accuracy than several recent works.

In the past decade, many statistical steganalysis features have been investigated for detecting audio steganographic algorithms. For instance, the Mel-frequency based features were introduced for audio steganalysis in [12]. Liu et al. proposed an approach based on Fourier spectrum statistics and Mel-cepstrum coefficients to detect the audio steganography [14]. In [15], the authors employed the Mel-cepstrum coefficients and Markov transition features from the second derivative of the audio signal to make steganalysis. In [6], the authors tried to build a linear basis to capture certain statistical properties of audio signal. In addition, there are some effective features to detect audio steganography in the frequency domain, such as MPEG-1 audio layer III (MP3) based steganalysis [9], and advanced audio coding (AAC) based steganalysis [17]. The above researches are based on hand-crafted high-dimensional features, and the performances of these works for the steganalysis of audio in the temporal domain are still far from satisfactory.

Recently, various deep learning architectures are proposed successively, which have achieved state-of-the-art results in many areas, especially in speech recognition and computer vision. However, very few deep learning based methods have been applied for audio steganalysis. Chen et al. [2] first proposed a sample convolutional neural network (CNN) to detect ±1 least significant bit (LSB) matching steganography in the temporal domain and achieved better results than the hand-crafted features. Lin et al. [13] proposed an improved CNN-based method for boosting the detection performance for the low embedding-rate steganography by adopting parameter transfer strategy. Wang et al. [22] presented an effective steganalytic scheme based on CNN for detecting MP3 steganography in the entropy code domain by using the quantified modified DCT (QMDCT) coefficients.

In this paper, we propose a modified deep residual convolutional network model for steganalysis of audio in the temporal domain. The deep residual network introduced in [7] has achieved promising performance on computer vision tasks, and also has been used for steganalysis of digital images [1,23,24]. Compared with the traditional convolutional neural network, the residual network introduces shortcut connections that directly pass the data flow to later layers, thus effectively avoids the vanishing gradient problem caused by multiple stacked non-linear transformations. As a consequence, deeper network constructed with residual block generally gets better performance in comparison with networks that consist of simply stack layers. The proposed network model is empirically designed with shortcut connections and a series of proven propositions, such as tanh activation function and high pass filter. The main contributions of this work are summarized as follows: (1) According to [2,18,25], employing the residual filters before inputting the original signal to neural networks usually results in better performance for steganalytic scheme. Inspired by this, a high pass filter module is implemented in the proposed scheme. (2) From previous researches [19] and [20], it is observed that the network model with larger depth can extract complex optimal functions more efficiently. Following the notions of residual network [7], we repeat the convolutional module twice in each residual block and design the network model with shortcut components. Our experimental results demonstrate that the proposed scheme obtains considerable improvement in terms of detection performance compared with several existing steganalytic methods.

The remaining parts of the paper is organized as follows. In Sect. 2, we describe the structure of the proposed steganalytic network and discuss the details for each component of the architecture. Next, the experimental setup and the overall performance of the proposed scheme for different scenarios are described in Sect. 3 and Sect. 4 respectively. Finally, we conclude the paper and state some directions for future works in Sect. 5.

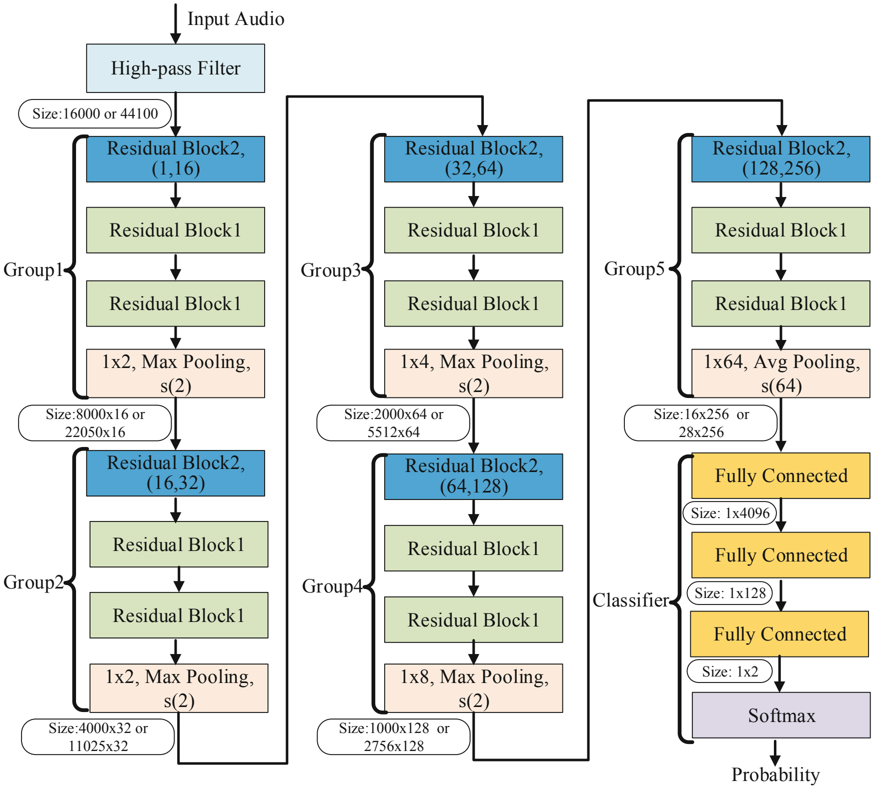


Fig. 1. The architecture of WavSResNet. The parameters in Residual Block2 are the input and output channels respectively. For example, "(m, n)" denotes input with m channels and n output channels. In Max and Avg Pooling layers, "1×2" denotes that pool size 1×2 and s(2) denotes that the stride of slide window is 2. The "Size 8000×16" represents the output dimension of the block, which is the shape of feature maps.

2 Proposed Method

The proposed network architecture is called “WavSResNet”, which represents residual network for waveform audio steganalysis. The word “residual” refers to the residual blocks with shortcut connections from deep learning. The shortcut connections help propagate gradients to upper layers and encourage feature reuse in the training process. In addition, residual blocks improve the performance of network because the vanishing gradient phenomenon that often negatively affects the convergence of deep network architectures [5,8]. The remainder of this section is divided into several subsections. Firstly, we describe the architecture of WavSResNet, and then analyze each part of the network in details. Most of the explorations focused on components in each residual block, the activation function, pooling layers and shortcut components.

2.1 WavSResNet Architecture

As schematically depicted in Fig. 1, the proposed WavSResNet is divided into a concatenation of several segments. The Residual Block1 and Residual Block2 represent identity shortcut (Fig. 2) and projection shortcut (Fig. 3) respectively. A convolutional layer with a fixed kernel (-1, 2, -1) is placed at the beginning of the network to transform the input audio data into residual signal, which act as the high pass filter. Then five groups are stacked one after another and each group contains one Residual Block2, two Residual Block1 and one max pooling layer, expect for Group5 with average pooling layer. Each residual block consecutively consists of convolutional layers, batch normalization (BN), Tanh activation function and the shortcut component. Among them, the convolutional layers followed by BN and tanh activation function extract features from different perspectives. The “shortcut” allows the gradient to be directly back-propagated to the earlier layer [8]. Tanh activation function help the network decide whether the information that the neuron received is relevant for the given information or should it be ignored. The max pooling layers keep the texture information of the sliding window and reduce the number of parameters from the previous convolutional layer. Having been processed by five groups of blocks, the input clips data with size 1×16000 or 1×44100 (see the experimental setups in Sect. 3.1) are finally transformed to a 256 feature maps with the size of 1×16 or 1×28. In the last segment, three standard fully connected layers and followed by a soft-max function, which act as a role of “the linear classifier” and map the features to the label space.

2.2 Convolutional Layers

The convolutional layers are the main components in CNN, which use one or several filters to convolve the input data and generate different feature maps for subsequent processing. In the proposed network, there are two kinds of convolutional layers which are the fixed convolutional layer and the common convolutional layer. Fixed convolutional layer is used to capture the minor modification introduced by the data hiding methods through reducing the impact of content information which can be seen as a high pass filter (HPF). The common convolutional layers with parameters are used to generate feature maps. In each convolutional layer, the convolutional kernel with the shape of 1×3 is used with the stride of (1, 1) and the padding is "SAME", which is followed by BN step to speed up training.

2.3 Activation Function

In the proposed network, we choose the nonlinear activation function to introduce nonlinear factors. The Rectified Linear Unit (ReLu) is the most commonly activation function, which seems maybe faster than Tanh for many of the given examples [3]. However, considering the property of audio steganalysis task, the saturation region of the tanh activation function limit the range of data value which can improve the performance of our model (refer to Sect. 4.2, compared with network #3). So we choose Tanh as the activation function instead of the Relu.

2.4 Pooling Layers

In order to extract robust invariant features and reduce the number of parameters from previous convolutional layer, it is very often to insert a pooling layer right after the convolutional layer in neural network. The pooling layer can be regarded as a kind of fixed convolutional layer and realized by average pooling or max pooling. The difference between average pooling and max pooling is the output of pooling. The max pooling keeps the texture information of the sliding window and outputs the maximum value of the sliding window, while the average pooling outputs the average value of the sliding window.

Considering characteristics of audio steganalysis, the max pooling layer and the average pooling layer are used in our proposed network. The size of max pooling layer is 1×2, 1×4, 1×8 with stride (1, 2). The average pooling layer has the size of 1×64 with the stride of (1, 64). In order to keep enough parameters, we just use five pooling layers and each one follows the second Residual Block1 in each group.

2.5 Shortcut Components

The proposed WavSResNet contains two types of shortcut connections because convolutional layers require different shortcut connections. Two main types of shortcuts are the projection shortcut and the identity shortcut, depending mainly on whether the input and output dimensions are same or different. The shortcut connection allows the gradient to be directly back-propagated to earlier layers by skipping over layers and helps deep residual networks from vanishing gradients.

We insert shortcut connections which turn the network into its counterpart residual version. The identity shortcuts can be directly used when the input and

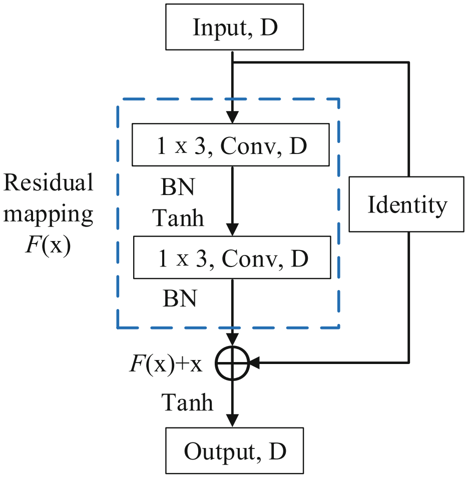


Fig. 2. Identity shortcut with skipping over 2 convolutional layers. The parameters inside the boxes represent kernel size, layer type, and the number of channels respectively. For example, "1 × 3, Conv, D" means a convolutional layer with 1 × 3 kernel size and D channels. "BN" and "Tanh" represent batch normalization layer and activation function respectively.

output are of the same dimensions as described in Fig. 2. When the dimension increases, to make the shortcut still performs identity mapping, the projection shortcut is used to match dimensions as sketched in Fig. 3 (done by 1×3 convolution). The difference between projection shortcut with the identity shortcut is that there is a convolutional layer without any non-linear activation function in the shortcut path, which is used to resize the input feature map to a different dimension. In this paper, the shortcut connections skip over 2 layers.

3 Experimental Setup

This section describes the common elements of all experiments that appear in Sect. 4, including the dataset, evaluation metric, training and testing of WavS-ResNet.

3.1 Dataset

The experiments will be carried out on three datasets which are SpeechData1, SpeechData2 and MusicData. SpeechData1 is the dataset used in [2], which includes the 40,000 cover-stego speech pairs. SpeechData2 and MusicData consist of 40,000 speech clips and 40,000 music clips respectively which were downloaded from the public data set [21]. Each clip was recorded with resolution of 16 bits per sample, duration of 1 s and stored in the uncompressed wave audio files (WAV). The SpeechData1 and SpeechData2 include mono speech corpus with

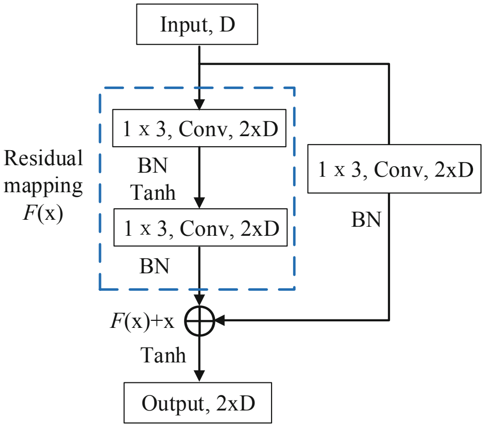


Fig. 3. Projection shortcut with skipping over 2 convolutional layers. The parameters inside the boxes represent the same meaning in Fig. 2.

a sampling rate of 16 kHz, and MusicData is mono audio clips with a sampling rate of 44.1 kHz. Three steganographic algorithms for WAV audio, which are ±1 LSB matching method (LSBM), modification of the amplitude value of sampling (Amplitude Modification) [26] and the adaptive steganographic algorithms (Luo Adaptive) [16], are implemented on original audio clips to generate the corresponding stego audio dataset. The stego clips in SpeechData1 were made by LSBM algorithm with 0.50 bit per sample (bps). The stego clips in SpeechData2 and MusicData were made by another two steganographic algorithms, which are Amplitude Modification with payloads 0.5 bps and Luo Adaptive with payloads 0.3 bps. Totally, each dataset contains 40,000 cover-stego pairs. Half of the pairs are used for training, and the rest are used for testing. In the training stage, the 16,000 pairs are used to train the network, and the rest 4,000 pairs are set aside for validation to chose the best trained model. In Sect. 4.1, we will introduce the experimental results on three datasets respectively.

3.2 Training Part

The WavSResNet has been experimented on speech clips dataset and music clips dataset with the same hyperparameters. The stochastic gradient descend (SGD) optimizer Adamax [11] was used with mini-batches of 32 cover-stego pairs and the training database was shuffled after each epoch. The batch normalization parameters were learned via an exponential moving average with decay rate 0.99. At the beginning of the training, the filter weights were initialized with random numbers generated from zero mean truncated Gaussian distribution with standard deviation of 0.1, and L^2 regularization. The filter biases were initialized to zero and no regularization. For the fully connected classifier layer, we initialized the weights with a zero mean Gaussian and standard deviation 0.01 and no bias.

On our dataset, the training was run for 100k iterations with an initial learning rate of $r_1 = 0.0001$. The snapshot achieving the best validation accuracy in the last 40k iterations was taken as the result of training. This training strategy was applied for three audio steganographic algorithms at three kinds of WAV audio datasets.

3.3 Testing Part

For comparison with the current state-of-the-art steganalytic methods of WAV format audio, the WavSResNet was compared with CNN based method introduced in [2], which we called in this paper ChenNet to distinguish it from the proposed network. Furthermore, another two hand-crafted features based steganalytic methods called as Liu1 [14] and Liu2 [15] were also used to be compared. The ChenNet was trained on exactly the same dataset as the WavSResNet and implemented in TensorFlow.

After choosing the bast trained model, we made the test with mini-batches of 64 audio clips which were randomly sampling from the test dataset without replacement until all data were recycled, and calculated the average detection accuracy. For another two hand-crafted features based steganalytic methods, after training support vector machine classifiers on the training dataset, the classifier was carried on the test dataset to calculate the detection accuracy.

3.4 Evaluation Metric

To evaluate the performance of the proposed scheme and state-of-the-art steganographic algorithms on WAV format audio dataset, the performance is measured by the detection accuracy of audio clips in the testing dataset. The detection accuracy is calculated as the number of correctly detection examples over the total number of the selected audio clips. Mathematically, the detection accuracy can be stated as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP, TN, FP and FN represent the number of true positives, true negatives, false positives and false negatives respectively. To obtain convincing results, all the experiments are repeated 10 times by randomly splitting the training and testing datasets.

4 Experiments

4.1 Experiments on Three Datasets

To show the efficiency of the proposed scheme in improving the performance of audio steganalysis in the temporal domain, we use WavSResNet and ChenNet [2] to make the detection on SpeechData1. Then we make experiments on

Table 1. The average detection accuracy (%) for LSBM algorithm in SpeechData1.

Steganalysis method	Accuracy
ChenNet	88.81
WavSResNet	89.70

Table 2. The average detection accuracy (%) of WavSResNet for different steganographic schemes.

Steganographic scheme	Dataset	
	SpeechData2	MusicData
Amplitude Modification	99.80	99.39
Luo Adaptive	54.69	52.25

SpeechData2 and MusicData to discuss the influence of sampling frequency to the WavSResNet. The stego examples are made as described in Sect. 3.1. We train the WavSResNet for 100,000 iterations with the parameters description in Sect. 3.2, and choose the best trained model to calculate the average detection accuracy on the test dataset. The average detection accuracy of WavSResNet and ChenNet [2] for LSBM on SpeechData1 are shown in Table 1, and the detection on SpeechData2 and MusicData for another two steganographic algorithms are shown in Table 2.

From Table 1, the WavSResNet's detection accuracy has achieved 89.70% on SpeechData1. It can be seen that the WavSResNet has better performance than ChenNet, which demonstrate improvement of WavSResNet for making audio steganalysis. From Table 2, we can see the overall performance on speech is better than that on the music, which may be interpreted as the samples' values of music are more complex than that of the speech clips and the network could not learn the rules efficiently. In addition, we can see that the steganographic algorithms have a heavily efficiency for the detection accuracy. The best detection accuracy is achieved when detects Amplitude Modification in SpeechData2, which reaches 99.80%. However, for the Luo Adaptive algorithm, the WavSResNet can not get the well results. This loss of performance is due to the fact that adaptive steganographic algorithms modify less samples and choose the complex area to embed messages, which makes the WavSResNet difficult to capture the discipline of this modification.

4.2 Comparison with the Variants

In this experiment, we try to show the effective of the WavSResNet by comparing it with its several variants. As listed in Table 3, three variant networks are used in this experiment, indexing from #2 to #4. The components of these variant networks are slightly different from the WavSResNet #1. All the networks have the same hyper-parameters described in Sect. 2 and are analysised on SpeechData1. In order to compare the fluctuation of 50 times experiments' detection accuracy for different networks, we record each experiment's detection accuracy and show the box plot for each network in Fig. 4.

Table 3. Network indices and descriptions.

#1	The proposed WavSResNet
#2	Remove the high pass filter layer
#3	Replace the activation function "Tanh" with "Relu"
#4	Remove all of shortcut connections from Residual Block1 and Residual Block2

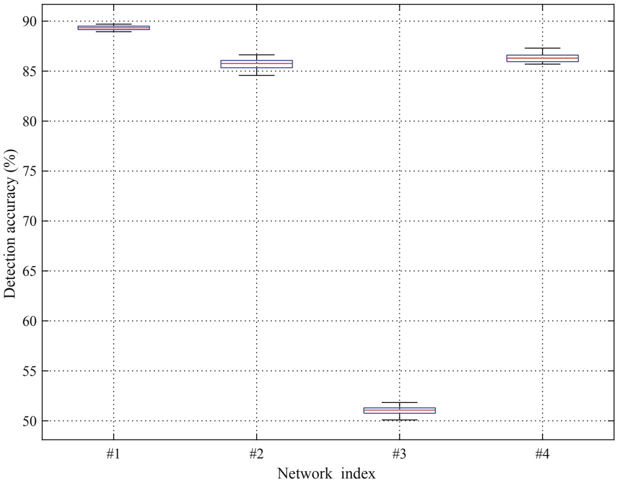


Fig. 4. Box plots of the detection accuracy obtained by different networks for detection LSBM with 0.5 bps.

From Fig. 4, it is observed that network #1 achieves the highest average accuracy of 89.70% and has a very stable performance over different times of testing. Most of detection accuracies of network #2 and #4 are between 84.01% and 87.72%, which are slight lower than that of network #1. The average detection accuracy of the network #3 is 51.02% because of the vanishing gradient problem, which are much lower than that of the proposed network. Moreover, we can observe from Fig. 4 that the accuracies of network #2 and #4 spread out in wide ranges, which indicates that these networks are not stable enough. In addition, we have additionally tested some other variants of the WavSResNet, which always obtain lower classify accuracy compared with the network #1, and we do not report them in the paper because of the limitation of space. As a result, the proposed WavSResNet (#1) converges relatively fast and achieves the best detection accuracy, so it is the most effective network compared to other variants.

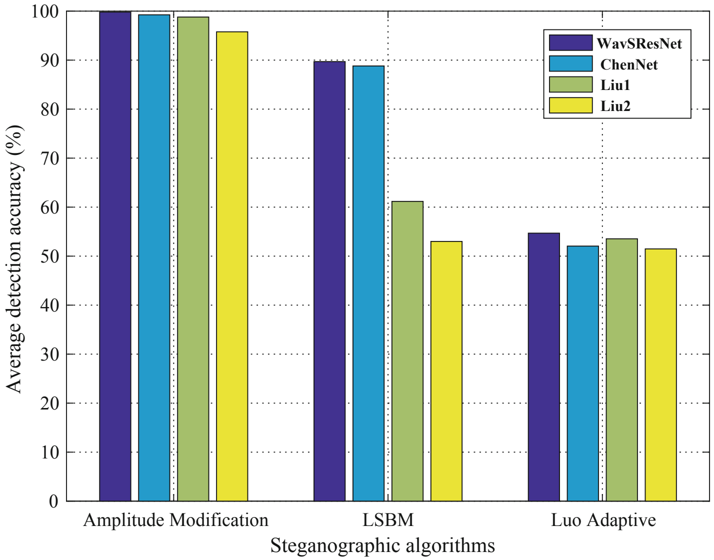


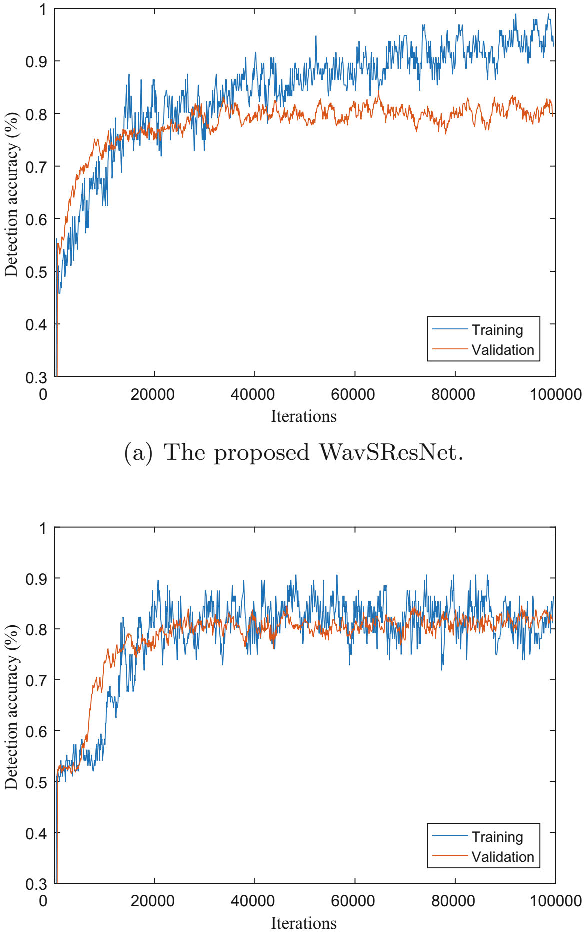
Fig. 5. The average detection accuracy of tree steganalytic algorithms for different steganographic methods.

4.3 Comparison with Previous Methods

In this experiment, we try to show the effectiveness of the WavSResNet by comparing it with three state-of-the-art steganalytic methods, which are ChenNet [2], Liu1 [14] and Liu2 [15]. The average detection accuracy of these steganalytic methods for three steganographic algorithms are shown in Fig. 5. The LSBM algorithm is tested on the SpeechData1 and another two steganographic algorithms are tested on SpeechData2.

As we can see in Fig. 5, the proposed WavSResNet has better detection performance for different steganographic algorithms than ChenNet. The average detection accuracy of WavSResNet improves upon ChenNet by up to 2.32%. The biggest improvement is typically observed for Luo Adaptive steganography. In addiction, we show that both network based detectors clearly outperform the traditional hand-crafted features based steganalytic paradigms. As a rule, the overall detection accuracy of Luo Adaptive algorithms is not very high. The reason is that the adaptive steganographic algorithms modify less samples, which makes the steganalytic algorithms hard to detect audio steganography in temporal domain effectively.

To further evaluate the WavSResNet and ChenNet with relatively good performance, we draw the curves of their training and validation accuracy during the training stage in Fig. 6. It is observed that the training accuracy of both networks steadily increased with more iterations before the convergence of networks. The validation accuracy of the ChenNet almost doesn't increase after about 30,000 iterations, which means that network converges in this case and more iterations



(a) The proposed WavSResNet.

(b) ChenNet.

Fig. 6. The detection accuracy curves of two networks. (a) Our proposed method WavSResNet. (b) Chen's method ChenNet [2].

could not improve its validation accuracy. The WavSResNet, which doesn't converge even after 60,000 iterations, can eventually converge after more 80,000 iterations and its final detection accuracy is higher than that of ChenNet. On the whole, the residual blocks and more layers make WavSResNet need more iterations to be converged, but also make it achieve the better results than the other CNNs based audio steganalysis.

5 Conclusion

In this paper, a novel audio steganalytic method based on residual convolutional neural network is proposed as WavSResNet. Compared with existing CNNs based audio steganalytic methods, the deeper network structure and the shortcut connections are utilized to the WavSResNet. Experimental results demonstrate that the WavSResNet obtains considerable improvement in terms of detection performance compared with several existing steganalytic methods. However, although the proposed method can achieve the state-of-the-art performance, it still has a long way to improve the accuracy of audio steganalysis in the temporal domain to a high level. All of our source codes and datasets will be available via GitHub: https://github.com/Amforever/WavSteganalysis.

Further works will be focus on two directions. On the one hand, the higher quality features of the audio signal can be developed before inputting the original audio data to the convolutional networks. On the other hand, some more powerful neural networks can be adopted to improve the accuracy of audio steganalysis.

Acknowledgments. This work was supported by NSFC under U1736214, 61902391 and 61972390, and National Key Technology R&D Program under 2019QY0700 and 2016QY15Z2500.

References

1. Boroumand, M., Chen, M., Fridrich, J.: Deep residual network for steganalysis of digital images. IEEE Trans. Inf. Forensics Secur. **14**(5), 1181–1193 (2018)
2. Chen, B., Luo, W., Li, H.: Audio steganalysis with convolutional neural network. In: Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, pp. 85–90. ACM (2017)
3. Eger, S., Youssef, P., Gurevych, I.: Is it time to swish, comparing deep learning activation functions across NLP tasks. arXiv preprint arXiv:1901.02671 (2019)
4. Fridrich, J.: Steganography in Digital Media: Principles, Algorithms, and Applications. Cambridge University Press, Cambridge (2009)
5. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256 (2010)
6. Han, C., Xue, R., Zhang, R., Wang, X.: A new audio steganalysis method based on linear prediction. Multimedia Tools Appl. **77**(12), 15431–15455 (2017). https://doi.org/10.1007/s11042-017-5123-x
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. IEEE (2016)
8. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016, Part IV. LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_38

9. Jin, C., Wang, R., Yan, D.: Steganalysis of MP3Stego with low embedding-rate using Markov feature. Multimed. Tools Appl. **76**(5), 6143–6158 (2016). https://doi.org/10.1007/s11042-016-3264-y
10. Ker, A.D.: The square root law of steganography: Bringing theory closer to practice. In: Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, pp. 33–44. ACM (2017)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Kraetzer, C., Dittmann, J.: Mel-cepstrum-based steganalysis for VoIP steganography. In: Proceedings of SPIE conference on the Security, Steganography and Watermarking of Multimedia. pp. 5–12. SPIE (2007)
13. Lin, Y., Wang, R., Yan, D., Dong, L., Zhang, X.: Audio steganalysis with improved convolutional neural network. In: Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, pp. 210–215. ACM (2019)
14. Liu, Q., Sung, A.H., Qiao, M.: Temporal derivative-based spectrum and Mel-cepstrum audio steganalysis. IEEE Trans. Inf. Forensics Secur. **4**(3), 359–368 (2009)
15. Liu, Q., Sung, A.H., Qiao, M.: Derivative-based audio steganalysis. ACM Trans. Multimed. Comput. Commun. Appl. **7**(3), 1–19 (2011)
16. Luo, W., Zhang, Y., Li, H.: Adaptive audio steganography based on advanced audio coding and syndrome-trellis coding. In: Kraetzer, C., Shi, Y.-Q., Dittmann, J., Kim, H.J. (eds.) IWDW 2017. LNCS, vol. 10431, pp. 177–186. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-64185-0_14
17. Ren, Y., Xiong, Q., Wang, L.: A steganalysis scheme for AAC audio based on MDCT difference between intra and inter frame. In: Kraetzer, C., Shi, Y.-Q., Dittmann, J., Kim, H.J. (eds.) IWDW 2017. LNCS, vol. 10431, pp. 217–231. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-64185-0_17
18. Shi, X., Li, B., Tan, S.: Preprocessing layer in spatial steganalysis based on deep learning. J. Appl. Sci. **36**(2), 309–320 (2018)
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
20. Sun, S., Chen, W., Wang, L., Liu, X., Liu, T.Y.: On the depth of deep neural networks: a theoretical view. In: Thirtieth AAAI Conference on Artificial Intelligence, pp. 2066–2072 (2016)
21. Wang, Y., Yang, K., Yang, Y., Zhang, Z., Yi, X., Zhao, X.: Audio steganalysis dataset (2019). https://ieee-dataport.org/documents/audio-steganalysis-dataset
22. Wang, Y., Yang, K., Yi, X., Zhao, X., Xu, Z.: CNN-based steganalysis of MP3 Steganography in the entropy code domain. In: Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security, pp. 55–65. ACM (2018)
23. Wu, S., Zhong, S.H., Liu, Y.: Steganalysis via deep residual network. In: 2016 IEEE 22nd International Conference on Parallel and Distributed Systems, pp. 1233–1236. IEEE (2016)
24. Wu, S., Zhong, S., Liu, Y.: Deep residual learning for image steganalysis. Multimed. Tools Appl. **77**(9), 10437–10453 (2017). https://doi.org/10.1007/s11042-017-4440-4
25. Ye, J., Ni, J., Yi, Y.: Deep learning hierarchical representations for image steganalysis. IEEE Trans. Inf. Forensics Secur. **12**(11), 2545–2557 (2017)
26. Zou, M., Li, Z.: A wav-audio steganography algorithm based on amplitude modifying. In: Tenth International Conference on Computational Intelligence and Security, pp. 489–493. IEEE (2014)