



Detection of Various Speech Forgery Operations Based on Recurrent Neural Network

Diqun Yan[✉]  and Tingting Wu 

College of Information Science and Engineering, Ningbo University, Ningbo 315211, China
yandiqun@nbu.edu.cn

Abstract. Most existed algorithms of speech forensics have been proposed to detect specific forgery operations. In realistic scenes, however, it is difficult to predict the type of the forgery. Since the suspicious speech might have been processed by some unknown forgery operation, it will give a confusing result based on a classifier for a specific forgery operation. To this end, a forensic algorithm based on recurrent neural network (RNN) and linear frequency cepstrum coefficients (LFCC) is proposed to detect four common forgery operations. The LFCC with its derivative coefficients is determined as the forensic feature. An RNN frame with two-layer LSTM is designed with preliminary experiments. Extensive experiments on TIMIT and UME databases show that the detection accuracy for the intra-database evaluation can achieve about 99%, and the detection accuracy for the cross-database can achieve higher than 88%. Finally, compared with the previous algorithm, better performance is obtained by the proposed algorithm.

Keywords: Forensics · Forgery operations · Recurrent neural network

1 Introduction

Nowadays, speech recording can be easily forged by some audio software. It will cause a huge threat if we cannot make sure the speech is natural or maliciously modified. Specifically, it will bring an inestimable impact on society when the forged speech is used for news report, court evidence and other fields.

In the past decades, digital speech forensics plays a crucial role on identifying the authenticity and integrity of speech recordings. Lots of works have been proposed. In order to detect the compression history of AMR audio, Luo [1] proposed a Stack Autoencoder (SAE) network for extracting the deep representations to classify the double compressed audios with a UBM-GMM classifier. Jing [2] present a detection method based on adaptive least squares and periodicity in the second derivative of an audio signal as a classification feature. For protecting text-dependent speaker verification systems from the spoofing attacks, Jakub [3] proposed an algorithm for detecting the replay attack audio. In [4], Galina use a high-level feature with a GMM classifier to against the synthetize audio in ASVspoof challenge. To detect the electronic disguised speech, Huang [5] proposed a forensic algorithm that adopted the SVM model with the Mel-frequency Cepstral Coefficients (MFCC) statistical vectors as acoustic features, including the MFCC and its mean value and correlation coefficients. The experimental results

show that their algorithm can achieve a high detection accuracy about 90%. In [6], Wang combined Linear Frequency Cepstrum Coefficient (LFCC) statistical moment and formant statistical moment as input features to detect electronic disguised audio in adding different SNR and different types of background noise.

Most of those forensic methods have achieved a good performance on detecting the modified speech with a specific forgery operation. However, they will be failed to detect the unknown forgery operation. For example, the electronic disguised classifier can identify whether the testing speech has undergone disguising processing. If the testing speech was only processed with noise-adding, the classifier will not give the correct result.

In recent years, some researchers start to focus on the forensics of various forgery operations. In [7], Jeong proposed a method to detect various image operations by a statistical feature. Luo [8] used the statistical features derived from image residuals to build an identifications of various image operations. The traditional features like MFCC are adopted as the acoustic feature in most existing forensic methods. However, with the fast development of deep learning, we can obtain a more powerful discrimination ability classifier based on deep learning techniques such as CNN and RNN [10–13]. In [9], Chen designed a convolutional neural network (CNN) with a fixed prior layer to classify different audio operations. The result shows that the CNN based method can achieved a better accuracy than the traditional forensic methods.

In this paper, we present an RNN to detect various speech forgery operations with the traditional feature MFCC and LFCC. We have made extensive experiments to verify the suitable feature and the architecture of RNN. The results show that the proposed method can detect the kinds of forgery operations, and outperforms better than the other detection works.

The rest of the paper is organized as follows. Section 2 introduces input of the network and feature extraction. Section 3 describes the proposed network architecture and some important hyper parameters. Section 4 presents comparative results for the detection of various forgery operation. Finally, the concluding remarks of this paper are given in Sect. 5.

2 Feature Extraction

The cepstrum coefficients which are the representation of the spectrum of speech signal in the setting window frame, have been commonly applied as a classificational feature to present the difference between original speech and the forged speech. The experimental results show that the forgery operations will cause the cepstrum coefficients of operated speech different from the original speech. In this section, we will give a briefly introduction of MFCC and LFCC, which are two of most used cepstrum coefficients.

2.1 Mel-Frequency Cepstrum Coefficient

MFCC is a beneficial speech feature based on human auditory perception characteristics, which are widely used for speech recognition [14]. Figure 1 shows the procedure for extracting the MFCC statistical moments.

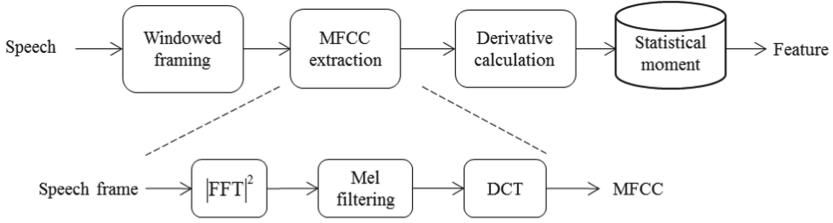


Fig. 1. Extraction procedure of MFCC statistical moment.

The MFCC focuses on the non-linear frequency characteristic and the size of Mel frequency corresponds to the relation of the logarithmic distribution of linear frequency and accords with the human ears' characteristic. The relationship between Mel frequency and linear frequency is shown as,

$$Mel(f) = 2595 \lg(1 + f/700) \quad (1)$$

where f is linear frequency.

At first, the speech signal $x(n)$ is divided into N frames, and the Hamming window $H(n)$ is adopted to obtain the windowed frame from the raw speech signal, as shown,

$$H(n) = 0.54 - 0.46 \cos \frac{2\pi n}{Z-1}, n = 0, 1, \dots, Z-1 \quad (2)$$

where Z is the total number of the frames in a speech sample.

Then the frequency spectrum $F(\omega)$ of the i -th frame $x_i(n)$ is calculated through a Fast Fourier Transform (FFT). The power spectrum $|F(\omega)|^2$ is process by a Mel-filter bank B_{Mel} which consist of M triangular band-pass filters. Then the power P_m of the m -th Mel-filter $B_m(\omega)$ is denoted as,

$$P_m = \int_{f_{lm}}^{f_{um}} B_m(\omega) |F(\omega)|^2 d\omega, m = 1, 2, \dots, M \quad (3)$$

where f_{um} and f_{lm} present the upper and lower cut-off frequencies of $B_m(\omega)$.

Then pre-emphasize the i -th frame $x_i(n)$ and transform it through Fast Fourier Transform and gain the L -dimensional MFCC of $x_i(n)$ through discrete cosine transform. The calculative formula is defined as,

$$C_l = \sum_{m=1}^M \left[\log P_m \cdot \cos \frac{l(m-0.5)\pi}{M} \right], l = 1, 2, \dots, L \quad (4)$$

where C_l is the l -th MFCC composition, L is less than the number of Mel filters.

We also calculate the dynamic cepstrum coefficients derivatives ($\Delta MFCC$ and $\Delta MFCC$). Assume that v_{ij} is the j -th component of the MFCC vector of the i -th frame, and V_j is the set of all j -th components. The average value E_j of each component set V_j and the correlation coefficient $CR_{jj'}$ between different component sets V_j and $V_{j'}$ are obtained by Eq. 5 and Eq. 6, respectively.

$$E_j = E(V_j) = E(\{v_{1j}, v_{2j}, \dots, v_{Nj}\}), j = 1, 2, \dots, L \quad (5)$$

$$CR_{jj'} = \frac{cov(V_j, V_{j'})}{\sqrt{VAR(V_j)}\sqrt{VAR(V_{j'})}}, 1 \leq j \leq j' \leq L \tag{6}$$

$$W_{MFCC} = [E_1, E_2, \dots, E_L, CR_{12}, CR_{13}, \dots, CR_{L-1L}] \tag{7}$$

The E_j and $CR_{jj'}$ are combined to form the statistical moment W_{MFCC} of the L -dimensional MFCC vector by Eq. 7. In this way, the statistical moment $W_{\Delta MFCC}$ of the $\Delta MFCC$ vector and the statistical moment $W_{\Delta\Delta MFCC}$ of the $\Delta\Delta MFCC$ vector will also be obtained.

2.2 Linear Frequency Cepstral Coefficients

LFCC is an average distribution from low frequency to high frequency bandpass filters [14]. The extraction procedure of LFCC statistical moment is shown in Fig. 2.

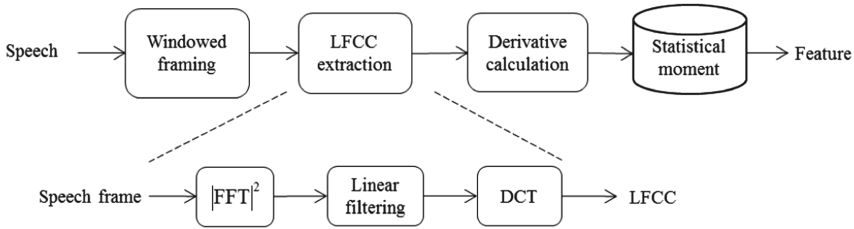


Fig. 2. Extraction procedure of LFCC statistical moment.

As shown in Fig. 2, the speech will firstly through the pre-process, and then the spectral energy can be obtained through the FFT, the calculative formula is shown as,

$$X_i(k) = \sum_{n=0}^{N-1} x_i(m)e^{-j2\pi n/N}, 0 \leq k \leq N \tag{8}$$

$$E(i, k) = [X_i(k)]^2 \tag{9}$$

where $x_i(m)$ is the speech signal data of the i -th frame, N is the number of Fourier.

Then the spectral energy will be processed through the bank filter group which including L bank filters with the center frequency f , $m = 1, 2, \dots, L$. The frequency response of triangular band-pass filter is shown as,

$$H_l(k) = \begin{cases} 0, & k < f(l-1) \\ \frac{k-f(l-1)}{f(l)-f(l-1)}, & f(l-1) \leq k \leq f(l) \\ \frac{f(l+1)-k}{f(l+1)-f(l)}, & f(l) \leq k \leq f(l+1) \\ 0, & k > f(l+1) \end{cases} \tag{10}$$

And the filtering spectral energy processed by bank filter group is denoted as,

$$S(i, l) = \sum_{k=0}^{N-1} [X_i(k)]^2 H_l(k), 0 \leq l \leq L \quad (11)$$

where l denote the i -th triangular band-pass filter.

Then the DCT is applied to calculate the cepstrum coefficients of the output of the bank filters, the calculated formula is denoted as,

$$lfcc(i, n) = \sqrt{\frac{2}{L}} \sum_{l=0}^{L-1} \ln[S(i, l)] \cos\left(\frac{\pi n(2l-1)}{2L}\right) \quad (12)$$

where n represents the spectrum after the DCT of the i -th frame,

As the same process of MFCC, we also calculate the first-order difference Δ LFCC of LFCC and second-order difference $\Delta\Delta$ LFCC. The concrete calculative formula is shown as,

$$LFCC = \begin{vmatrix} x_{1,1} & \cdots & x_{1,n} \\ \cdots & \cdots & \cdots \\ x_{s,1} & \cdots & x_{s,n} \end{vmatrix} \quad (13)$$

$$\Delta x_{i,j} = \frac{1}{3} \sum_{u=-2}^2 u x_{i+u,j}, 3 \leq i \leq s-2, 1 \leq j \leq s \quad (14)$$

3 Detection Method Based on RNN

In this section, we will give a general description of the proposed framework for detecting four forgery operations based on RNN.

3.1 Framework

Recently, many deep learning approaches have been applied as the classifier especially the CNN which can capture the highly complex feature from a raw sample significantly [15]. It is obvious that, the CNN structure can effectively extract deep high-level features and obtain a good detection result in image forensics. However, it is not suitable for speech forensic task because the CNN structure cannot capture the sequential connection well. Recently, RNN have been widely used for applications processing temporal sequences such as speech recognition, which can capture the correlation between the frames [16]. Hence, we apply the RNN model in our task of classify the various forgery operations.

The proposed framework is shown in Fig. 3. The traditional feature is extracted from raw waveform, then fed into the RNN. In this work, we choose the statistical moments of MFCC and LFCC cepstrum coefficients as the features mentioned in Sect. 2.1.

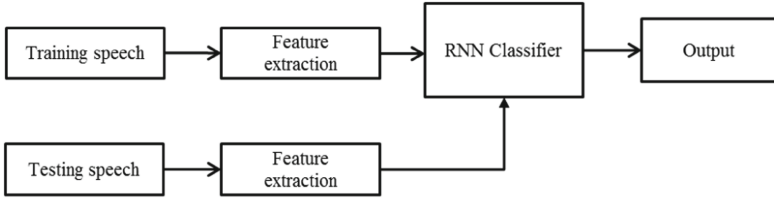


Fig. 3. Proposed classification framework.

Due to the gradient vanishing and exploding issues in training a single-layer RNN, most of the existing RNN architectures only consist of several layers (1, 2 or 3), although the deeper network will capture more useful information. Hence, in this work, to find the better architecture of RNN, three networks have been designed. The network configurations are shown in Fig. 4. Meanwhile, we set the *tanh* activation function to improve the performance of the model, and set the value of the Dropout function to 0.5, which can help the network reduce the overfitting in training procedure. And a Softmax layer is followed to output the probability.

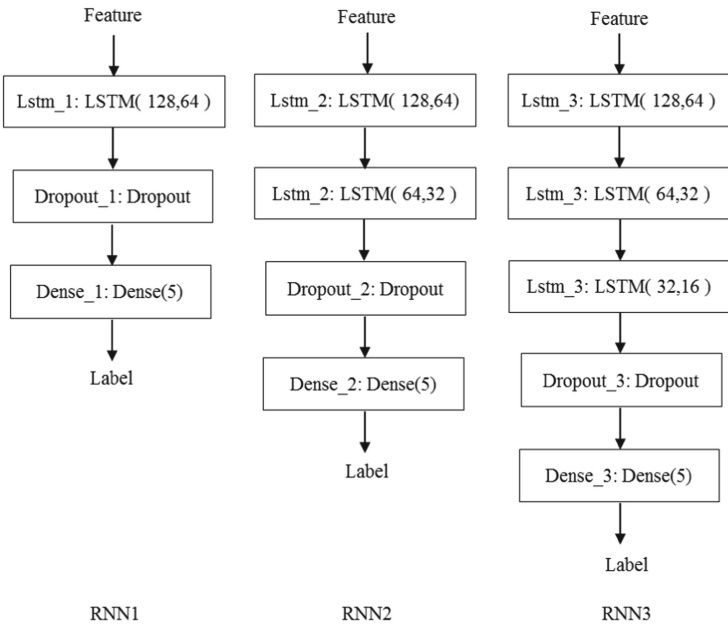


Fig. 4. Three proposed recurrent neural networks.

In the experimental stage, the RNN with two-layers of LSTM layers temporarily selected as the baseline network to find the best features among MFCC and LFCC for detect forgery operations. Then the selected features are used to determine the architecture of RNN.

3.2 Training Strategy

The training strategy of the proposed method includes two stages: training and testing. Before the training, we process the original speech by selecting a parameter for each forgery operation randomly. The training procedure is performed according to the process shown in Fig. 3. The classification feature will be firstly extracted from the original speech and the forged speech which through the disguising, noise-adding, high-pass filtering and low-pass filtering, and then the features will be used for training the RNN. In the testing, we frozen the parameters of RNN model, and choose a part of the original speech and forged speech as the test database, then the final detection result from the output of the Softmax layer will be obtained. Finally, the accuracy is taken as the evaluation metric, and we perform the confusion matrix by making a comparison of the predict labels of testing database and its true labels.

4 Experimental Results and Analysis

In this section, we first present the experimental data and then compare the proposed method with other existing methods.

4.1 Experiment Setup

We create four forgery databases based on the TIMIT speech database [17] and the UME speech database [18], including disguising, low-pass filtering, high-pass filtering and noise-adding. Specifically, we use the Audition CS 6 to build the electric disguised database, and the MATLAB is applied to build the other three forgery database. As shown in Table 1, for each forgery operation, we choose four different operational parameters. And we use the Gaussian white noise as the added noise. And the sample splicing of train setting and test setting of TIMIT and UME are shown in Table 2.

Table 1. Parameters processed by different forgery operations.

Operation	Parameter
Noise-adding	SNR (dB): 5, 10, 15, 20
Disguising	Modification degree: +4, +8, -4, -8
Low-pass filtering	Pass-band cut-off frequency (Hz): 500, 750, 1000, 1250
High-pass filtering	Pass-band cut-off frequency (Hz): 1750, 2000, 2250, 2500

Forged speech databases are built by selecting the forgery speech from those forgery databases. Then, a 4 NVIDIA GTX1080Ti GPUs with 11 GB graphic memory is used for the RNN training.

Table 2. Specific database for multiple operations (Natural/Operated).

Database	TIMIT		UME	
	Training	Testing	Training	Testing
TIMIT	4000/64000	2300/36800	6300/100800	4040/64640
UME	4040/64640	6300/100800	3200/51200	840/13400

4.2 Experimental Results

First, we choose a two-layer RNN architecture for selecting a suitable forensic feature from the acoustic features, including MFCC, LFCC and its first and second derivative called Δ MFCC, $\Delta\Delta$ MFCC and Δ LFCC, $\Delta\Delta$ LFCC. Then 6 well-trained two-layer RNN models are obtained for each feature and the sample for testing on TIMIT and UME are fed into the 6 models to compare the forensic capability of 6 acoustic features. Table 3 shows the detection accuracy of 6 traditional acoustic features. The MFCC with its first and second derivative features Δ MFCC, $\Delta\Delta$ MFCC is better than the other features for classifying the various forged samples in the intra-database, the average accuracy is about 99%. But it is perform a lower accuracy in cross-database (testing the UME samples while the model was trained by TIMIT database), approximately 80%, which means the MFCC features may not be universal and robust.

Table 3. Average detection accuracy of six features in a two-layer RNN (%).

Training	Feature	Testing	
		TIMIT	UME
TIMIT	LFCC	99.17	98.43
	LFCC + Δ LFCC	99.69	87.90
	LFCC + Δ LFCC + $\Delta\Delta$ LFCC	95.14	90.25
	MFCC	99.95	100
	MFCC + Δ MFCC	99.93	100
	MFCC + Δ MFCC + $\Delta\Delta$ MFCC	100	100
UME	LFCC	88.95	97.92
	LFCC + Δ LFCC	87.90	98.28
	LFCC + Δ LFCC + $\Delta\Delta$ LFCC	78.56	89.78
	MFCC	82.67	99.9
	MFCC + Δ MFCC	79.47	99.9
	MFCC + Δ MFCC + $\Delta\Delta$ MFCC	77.67	99.92

Different from MFCC, the LFCC feature have a better performance in the forensic task of detect the various operations. As shown in Table 3, the LFCC and its first and second derivative features Δ LFCC, $\Delta\Delta$ LFCC have achieved a detection accuracy about 88% in cross-database while maintaining a good performance in intra-database.

Compared with the results shown in Table 3, the MFCC features is not well performance the difference between the original samples and the four forged samples. It indicates that MFCC features is not robustness enough. Although the LFCC and its first and second derivate features is slightly reduced in the intra-database, it is also still in the acceptable range better than the MFCC in cross-database. Hence, the LFCCs is selected as the suitable acoustic feature considering.

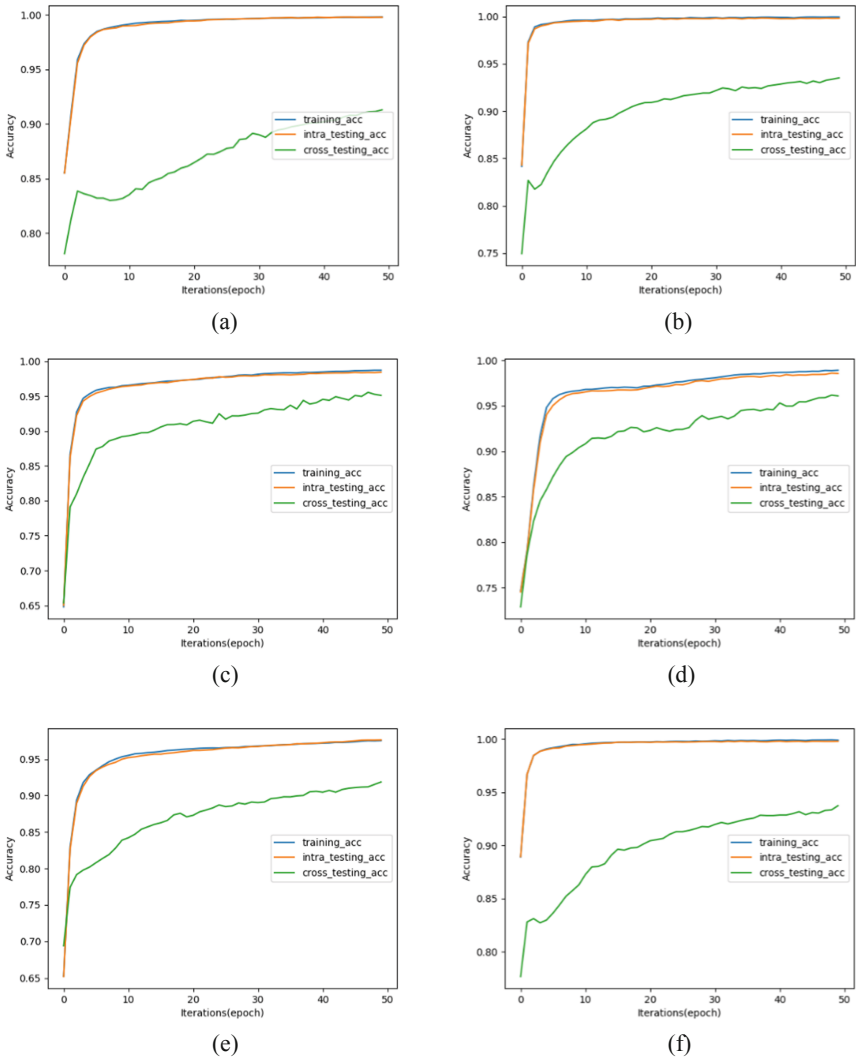


Fig. 5. Detection accuracy of three RNN networks among training process in TIMIT and UME database. (a) and (b) are the detection performance of RNN1 model. (c) and (d) are the detection performance of RNN2 model. (e) and (f) are the detection performance of RNN3 model. (a) (c) (e) are trained by TIMIT database and (b) (d) (f) are trained by UME database.

The structure of the RNN has play an important role in affecting the classification result. We design three structures for RNN in Fig. 4 to explore the impact of the specific network structures. Then, the selected features LFCCs are extracted from the original database and forgery databases for training the three RNN models. Finally, the classification probability will be obtained by the Softmax layer. The detection accuracy of three models based on TIMIT and UME databases in training process are shown in Fig. 5 (a–f). And the comparison results of three models are shown in Table 4.

Table 4. Average detection accuracy of different structures of RNN (%).

Structure	Testing					
	RNN1		RNN2		RNN3	
Training	TIMIT	UME	TIMIT	UME	TIMIT	UME
TIMIT	98.63	97.85	99.11	98.43	99.08	98.68
UME	87.07	96.93	88.95	97.92	88.90	97.85

As shown in the second and third rows, the testing results have an excellent accuracy (above 97%). And the RNN2 model achieved a better detection accuracy about 88% in detecting the cross-database. Results show that the detection ability of the RNN2 structure and the RNN3 network structure are similar. Hence, we choose the RNN2 model as the final structure for the detection of various forgery operations considering the complexity of the experiment.

4.3 Comparative Experiment

We make a comparative experiment of the detection performance between this work based on RNN and our previous work based on CNN [19]. In our previous work, we proposed a forensic method for identifying the four kinds of forgery operations. First, a fixed convolutional layer is used to obtain the residuals of the speech sample, and then the residual signals are classified by a set of convolutional layer group. The comparative experiment shows that the method proposed in this paper has greatly improved the classification accuracy.

As shown in Sect. 4.2, the RNN2 is determined as the final recurrent neural network with LFCCs as the acoustic feature in this work. Results show that the average detection accuracy of its classification result is about 90%. In order to compare with the existing work, we repeated the experiments in [19] with the original and forged databases, and the experimental results are shown in Table 5. As shown in the second and third rows, the test results all have excellent accuracy (above 96%) in the intra-database. Even the results of the CNN are slightly better than RNN. However, the test results all have a certain decline in the cross-database, and the detection rate of RNN can be maintained above 87%. Some of the multiclassification results given in this paper are comparable with the CNN model in [19], and some detection accuracy are significantly better than the detection method based on CNN.

Table 5. Classification capability of the proposed RNN compared with CNN model (%).

Classifier	Training	Testing	
		TIMIT	UME
CNN [19]	TIMIT	99.77	84.33
	UME	76.59	99.82
RNN	TIMIT	98.63	97.85
	UME	87.07	96.93

5 Conclusion

In this paper, we carefully design a speech forensic method based on RNN for the detection of various forgery operations, and provide extensive results to show that the proposed method can effectively identify forgery operations. In the future, we will extend the proposed model and explore the deep features extracted by the neural network to identify unknown forgery operations.

References

1. Luo, D., Yang, R., Li, B., et al.: Detection of double compressed AMR audio using stacked autoencoder. *IEEE Trans. Inf. Forensics Secur.* **12**(2), 432–444 (2017)
2. Jing, X.U., Xia, J.: Digital audio resampling detection based on sparse representation classifier and periodicity of second derivative. *J. Digit. Inf. Manag.* **13**(2), 101–109 (2015)
3. Gaka, J., Grzywacz, M., Samborski, R.: Playback attack detection for text-dependent speaker verification over telephone channels. *Speech Commun.* **67**, 143–153 (2015)
4. Lavrentyeva, G., Novoselov, S., Malykh, E., Kozlov, A., Kudashev, O., Shchemelinin, V.: Audio-replay attack detection countermeasures. In: Karpov, A., Potapova, R., Mporas, I. (eds.) *SPECOM 2017. LNCS (LNAI)*, vol. 10458, pp. 171–181. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66429-3_16
5. Wu, H., Wang, Y., Huang, J.: Identification of electronic disguised speech. *IEEE Trans. Inf. Forensics Secur.* **9**(3), 489–500 (2014)
6. Cao, W., Wang, H., Zhao, H., Qian, Q., Abdullahi, S.M.: Identification of electronic disguised voices in the noisy environment. In: Shi, Y.Q., Kim, H.J., Perez-Gonzalez, F., Liu, F. (eds.) *IWDW 2016. LNCS*, vol. 10082, pp. 75–87. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53465-7_6
7. Jeong, B.G., Moon, Y.H., Eom, I.K.: Blind identification of image manipulation type using mixed statistical moments. *J. Electron. Imaging* **24**(1), 013029 (2015)
8. Li, H., Luo, W., Qiu, X., et al.: Identification of various image operations using residual-based features. *IEEE Trans. Circuits Syst. Video Technol.* **28**(1), 31–45 (2018)
9. Chen, Q., Luo, W., Luo, D.: Identification of audio processing operations based on convolutional neural network. In: *ACM Workshop on Information Hiding and Multimedia Security*, Innsbruck, pp. 73–77 (2018)
10. Szegedy, C., Liu, W., Jia, Y., et al.: Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, pp. 1–9. IEEE (2015)
11. Liu, Y., Qian, Y., Chen, N., et al.: Deep feature for text-dependent speaker verification. *Speech Commun.* **73**, 1–13 (2015)

12. Tian, X., Wu, Z., Xiao, X., et al.: Spoofing detection from a feature representation perspective. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, Shanghai, pp. 2119–2123. IEEE (2016)
13. Variani, E., Lei, X., Mcdermott, E., et al.: Deep neural networks for small footprint text-dependent speaker verification. In: IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, pp. 4052–4056. IEEE (2014)
14. Rana, M., Miglani, S.: Performance analysis of MFCC and LPCC techniques in automatic speech recognition. *Int. J. Eng. Comput. Sci.* **3**(8), 7727–7732 (2014)
15. Chen, B., Luo, W., Li, H.: Audio steganalysis with convolutional neural network. In: Conference: the 5th ACM Workshop, Philadelphia, pp. 85–90 (2017)
16. Sak, H., Senior, A., Rao, K., et al.: Learning acoustic frame labeling for speech recognition with recurrent neural networks. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, Brisbane, pp. 4280–4284. IEEE (2015)
17. Timit Acoustic-Phonetic Continuous Speech Corpus. <https://catalog.ldc.upenn.edu/LDC93S1>. Accessed 20 Feb 2017
18. Advanced Utilization of Multimedia to Promote Higher Education Reform Speech Database. <http://research.nii.ac.jp/src/en/UME-ERJ.html>. Accessed 27 Feb 2017
19. Wu, T.: Digital speech forensics algorithm for multiple forgery operations. *Wirel. Commun. Technol.* **28**(3), 37–44 (2019). (in Chinese)