A ROBUST DEEP AUDIO SPLICING DETECTION METHOD VIA SINGULARITY DETECTION FEATURE

Kanghao Zhang[1,*], *Shan Liang*[2,*], *Shuai Nie*[2], *Shulin He*[1], *Jiahui Pan*[1], *Xueliang Zhang*[1], *Haoxin Ma*[2], *Jiangyan Yi*[2]

[1]College of Computer Science, Inner Mongolia University, China
[2]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
imu.koer@mail.imu.edu.cn, {sliang,shuai.nie}@nlpr.ia.ac.cn

ABSTRACT

There are many methods for detecting forged audio produced by conversion and synthesis. However, as a simpler method of forgery, splicing has not attracted widespread attention. Based on the characteristic that the tampering operation will cause singularities at high-frequency components, we propose a high-frequency singularity detection feature obtained by wavelet transform. The proposed feature can explicitly show the location of the tampering operation on the waveform. Moreover, the long short-term memory (LSTM) is introduced to the CNN-architecture LCNN to ensure that the sequence information can be fully learned. The proposed feature is sent to the improved RNN-architecture LCNN together with the widely used linear frequency cepstral coefficients (LFCC) to learn forgery characteristics where the LFCC is used as a supplement. Systematic evaluation and comparison show that the proposed method has greatly improved the accuracy and generalization.

***Index Terms*—** forged audio, tampering, high frequency, singularity detection feature

1. INTRODUCTION

Audio forgery detecting receives widespread attention in recent years since it is related to social security, privacy security, and property security [1–3]. Most of the previous works focus on the detecting task for playback, conversion, and synthesis [4]. Traditional methods have studied many well-performing features such as Cochlear Filter Cepstral Coefficient and Instantaneous Frequency (CFCCIF) [5], Linear Frequency Cepstral Coefficients (LFCC), and Constant-Q Cepstral Coefficients (CQCC) [6]. Meanwhile, much progress has also been made in terms of the back-end classifier. [7, 8] explored the end-to-end systems with light convolution neural networks (LCNN). [9] proposed a deep residual network (ResNet) with temporal pooling. And [10] proposed a feature genuinization based LCNN system for synthesis attack.

Compared to other attack methods, splicing is simpler and more difficult to detect. Splicing refers to the attack method of replacing or inserting a forged clip into pure audio. This kind of spliced audio tampers with the semantics while most information of original audio is retained, which brought a great challenge to the forgery detection system. Many works try to detect splicing through traditional algorithms. [11, 12] used the effect of forgery on frame offset to detect the forgeries in the time domain for MP3. [13, 14] extracted and analyzed the electric network frequency (ENF) from the signal to detect forgery and locate the tampering position. [15] used the local noise estimation as the detecting method for splicing audio. These traditional algorithms usually need to design many hyperparameters to determine the forgery boundary, while the deep learning methods learn it automatically. However, there are not many deep learning methods for splicing detection.

As described in [16], the splicing operation will produce singularities on the waveform. Therefore, we propose a high-frequency singularity detection feature based on wavelet transform which shows different magnitudes of forged and original segments on the waveform. In order to learn the difference between adjacent segments of the feature, we insert an LSTM after the last convolutional layer to enhance the long-term modeling ability. Note that much less energy of speech is presented in higher frequency bands. Therefore, the conventional used LFCC is also used as a kind of supplementary. Finally, we conduct some effectiveness and robustness experiments on the Half-truth Audio Detection (HAD) dataset [17]. The experimental results show that a great improvement is obtained in accuracy and generalization performance. We reckon that the ability of the high-frequency singularity detection feature to locate the tampering position provides great help for forgery detection. In addition, according to our test, the high-frequency singularity detection feature owns a high degree of noise immunity which makes the system more robust.

The rest of the article is organized as follows. Section 2 introduces the wavelet transform and the proposed feature. Section 3 describes the used network architecture and the

[*] These authors contributed equally to this work.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

training method in detail. Section 4 describes our experiments and shows the results, and gives some discussions. Finally, the conclusions are given in Section 6.

2. WAVELET TRANSFORM AND SINGULARITY DETECTION FEATURE (SDF)

Wavelet transform is an important tool for signal analysis. Different from the fixed window length of the short-time Fourier transform (STFT), the window length used by wavelet transform is variable which allows the different strategies to obtain low-frequency and high-frequency information [18]. For the high-frequency components of signal which are also called details in wavelet analysis, we can obtain better time-resolution using the contracted version of the wavelet. In contrast, better frequency resolution can be obtained at the low-frequency range using the dilated version of the wavelet. Therefore, this kind of analysis is particularly suitable for some stages where better time-resolution is required at high-frequency to detect rapidly changing singularities.

The continuous wavelet transform can be described by the following formulas:

$$CWT(c,\tau) = \int f(x)\psi_{c,\tau}(t)\mathrm{d}t \tag{1}$$

$$\psi_{c,\tau}(t) = \frac{1}{\sqrt{c}}\Psi(\frac{t-\tau}{c}) \tag{2}$$

where c and τ are real numbers, and $\Psi(t)$ is the mother wavelet function. The wavelets are dilated (c<1) or contracted (c>1) with the value of c and moved over the signal to be analyzed by the shift time τ. Contraction and expansion scale the frequency response so that the set of wavelets is able to span the desired frequency range [19].

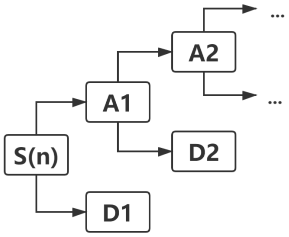


Fig. 1. The decomposition process of the 2-level wavelet transform.

Fig. 1 shows the specific process of 2-level wavelet decomposition. Each decomposition generates a high-scale, low-frequency component (An) which presents the overview and changing trend of the signal, and a low-scale, high-frequency component (Dn) which presents the details of the signal. Both of them are downsampled by the factor of 2. Figure 2 shows the frequency range analyzed by every process of the 2-level wavelet transform. The fm represents the highest frequency of the input signal. Each wavelet transform divides the frequency into low frequency and high frequency components. By multiple iterations of decomposition, we can obtain several low-frequency and high-frequency components.

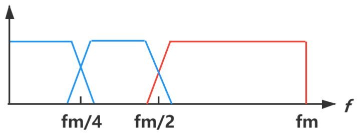


Fig. 2. Spectral characteristics for 2-level decomposition.

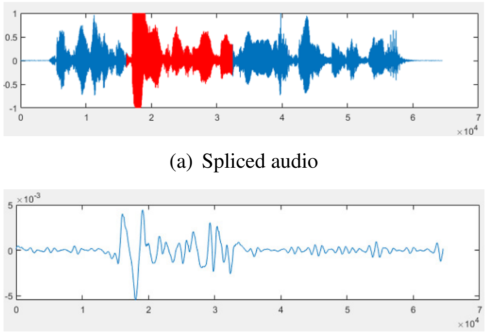
We perform an N-level wavelet decomposition to generate our singularity detection feature (SDF) and the process can be described as the following formulas:

$$A_n, D_i = \varphi(S,n), \qquad i \in [1,n] \tag{3}$$

$$Z_j = 0 \times D_j, \qquad j \in [1,k] \tag{4}$$

$$\hat{S} = \phi(A_n, Z_j, D_m, n), m \in [k,n] \tag{5}$$

where n is the level of decomposition, and k is the lowest level in the series of desired components. S and $\hat{S}$ represent the input signal and the reconstructed signal, respectively. $\varphi(\cdot),\phi(\cdot)$ represent the decomposition and reconstruction, respectively. To put it simply, we iterate multiple wavelet decomposition and set all components to 0 except the desired n-k high-frequency components, and then the modified components are used to reconstruct the high-frequency singularity detection feature.



(a) Spliced audio

(b) SDF reconstructed from a high-frequency component

Fig. 3. The original spliced audio and corresponding feature.

As mentioned earlier, tampering operations (including insertion, replacement, or splicing) will produce sharply changing singularities. However, the high-frequency components in wavelet analysis are particularly useful for detecting these sharply changing singularities. And the signal reconstructed from high-frequency components can clearly show the location of the tampering operation on the waveform. Fig. 3(a)

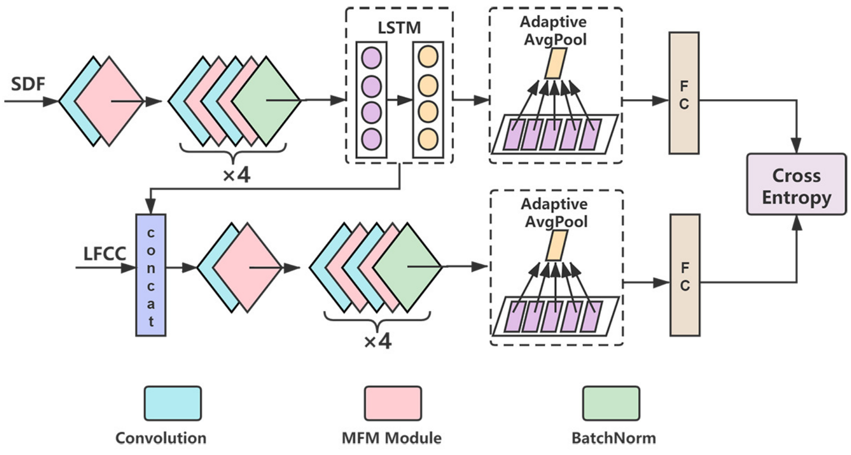


Fig. 4. Diagram of the proposed joint model

shows a spliced speech and Fig 3(b) shows the corresponding feature reconstructed from a high-frequency component. The segment depicted in red is the fake clip, which replaces the corresponding segment of the original audio. The corresponding segment in the reconstructed feature shows significantly higher magnitudes than other segments, which indicates the location of the splicing. Note that the feature and the original signal have the same length.

3. MODEL DESCRIPTION

3.1. Light Convolution Neural Network

LCNN is a lightweight framework widely used for forgery detection. It consists of convolutions, max-feature-map (MFM) modules, a pooling layer, and a linear layer. The MFM is a module that only retains the maximum value of the two channels to reduce CNN architecture. A convolution and an MFM module are combined to form a group. Several groups are arranged consecutively to encode the input feature, and then the encoded feature is pooled and sent to the linear layer to obtain scores. More details about LCNN can be found in [20].

3.2. Proposed System

As mentioned earlier, the proposed feature contains obvious sequence characteristics, and CNN architecture LCNN is not sufficient. Therefore, we introduce two layers of LSTM to learn the sequence information of SDF. Considering the SDF is quite different from the LFCC, we propose a joint model that includes two independent branches. The overall structure is shown in Fig. 4. The branch with the LSTM in the upper is used for preliminary learning of the proposed feature. We select three layers of the convolutions and take 2 as their frequency stride to transform the feature into a lower-dimensional space and make the output closer to the LFCC. Then the output of LSTM is concatenated with LFCC along the feature axis. Finally, the combined feature is sent to a basic LCNN. In addition, the outputs of the two branches are both constrained by the cross-entropy loss.

4. EXPERIMENTS

4.1. Dataset

In our experiments, we use the Half-truth Audio Detection (HAD) dataset [17]. The HAD is generated based on the AISHELL-3 corpus [21] which consists of 88035 utterances about 85 hours from 218 native Chinese mandarin speakers. The authors use fake clips generated by TTS and vocoder to replace named entities or attitudinal words to generate fake audios. HAD consists of two subsets of Partial and Full, where the Full set contains the real speech, and the Partial set contains the fake speech. The training set, validation set, and test set of each subset are partitioned with 6:2:2, and there is no overlap between them. In addition, it also contains an out-of-domain (OOD) test, which means the TTS model is different from the one used in the training set. The Partial set and the Full set are combined to form our experimental data. All utterances are resampled to 16 kHz and there are about 40K sentences for training in total.

4.2. Baseline Model and System Settings

In this study, we train three baselines that have the same architecture as the benchmark systems used in [17]. Firstly, we train two basic LCNN models which take LFCC as input. The number of parameters of the first baseline is 200K and the second one is 814K which is closer to the proposed model with 874K parameters. Secondly, we train a proposed

RNN architecture LCNN model which takes the proposed high-frequency singularity detection feature as its input. The number of parameters is 710K. The models of utterance-level binary classification are evaluated in terms of equal error rate (EER). A lower percent of EER means the better performance of the model. More details of metrics can be found in [22].

The signals are divided into frames with a frame length of 20ms and 30ms to obtain LFCC and SDF, and the overlap between adjacent frames is 10ms. The Adam optimizer with a learning rate of 0.0001 is used with a batch size of 16 utterances. Note that a segment with 750 frames will be intercepted from the feature if it has more frames than 750 while smaller ones are zero-padded to match the size of the longest utterance in the batch. The decomposition level n is 12 and the desired level k ranges from 10 to 12 in our experiments. And the db4 is used as the mother wavelet.

4.3. Results and Analysis

Firstly, we conduct a series of experiments to verify the effectiveness of the proposed method for forgery detection. The experimental results are shown in Table 1 and the best results are marked in bold. The baselines are expressed in the form of LCNN-N, where LCNN-1 refers to the CNN architecture LCNN with 200K parameters, and LCNN-3 is the proposed LCNN with LSTM. The Pro. refers to the proposed joint model. For the results of the test set, the proposed method outperforms the other three baseline methods. The LCNN-1 and LCNN-3 reach the EERs of 1.8% and 9.3%, respectively. Compared with these two baselines, the proposed method improves EER by 0.8 and 8.3 percentage point, respectively. To provide a further fair comparison, we increase the parameters of the LCNN-2 to 814K. The proposed method still improves the EER by 0.4 percentage point of EER compared to LCNN-2. In addition, we can find that the increase of parameters slightly improves the performance in the unseen test set. In contrast, the EER is significantly improved from 8.4% to 4.8% if the proposed feature is used, which demonstrates the excellent generalization ability of the proposed method.

Table 1. Equal Error Rate Objective Evaluation (EER) on test set, OOD test set and noisy test set.

Model	Feature	Test EER	OOD EER	Noisy EER
LCNN-1	LFCC	1.838	9.219	44.93
LCNN-2	LFCC	1.434	8.461	36.84
LCNN-3	SDF	9.331	13.261	**20.91**
Pro.	**BOTH**	**1.057**	**4.826**	20.95

Secondly, we add white noise at 15dB to the audio and evaluate the trained model on the noisy data. It can be found that the EER drops to 44.9% for the baseline LCNN-1, which implies the LCNN almost loses the discrimination ability if only LFCC feature is used. However, the baseline LCNN-3 and the proposed method using the SDF keep the EER of 20.9%. These results show that SDF can adapt to the noise environment, which improves the generalization of the model.

Table 2. The result of precision (P), recall (R) and F1-score (F1) in the localization experiments.

Model	Feature	Test			OOD Test		
		P	R	F1	P	R	F1
LCNN-1	LFCC	97.6	90.4	93.8	93.3	47.7	63.1
LCNN-2	LFCC	96.7	97.9	97.3	92.1	63.1	74.8
LCNN-3	SDF	90.4	75.6	82.3	86.4	58.2	69.5
Pro.	**BOTH**	**98.4**	**98.1**	**98.3**	**94.4**	**76.9**	**84.8**

Lastly, we perform localization experiments as supplements. In this experiment, the model judges each frame of audio. It can be seen from Table 2 that the proposed method improves the ability to locate forged clips. This further proves that the proposed SDF has a great effect on localization.

Based on our experimental results, we reckon that several reasons explain the excellent performance of the proposed method. First of all, the proposed high-frequency singularity detection feature can detect the tampered position of the forged audio and shows the different magnitudes of the forged segment and real segment on the waveform explicitly. Secondly, we introduce LFCC as a supplement, which can compensate for the missing energy information of the proposed feature. Thirdly, the RNN architecture LCNN has strong long-term modeling capability, which ensures that the network can record the difference between the spliced and the original segment. Finally, the adopted joint model retains the advantages of the two features. LFCC is used as a supplement of energy information to the proposed feature while the proposed feature makes up for the poor noise immunity.

5. CONCLUSIONS

In this research, we propose a novel feature with clear physical meaning called high-frequency singularity detection feature. We introduce LFCC as a kind of supplementary information with the proposed singularity detection feature and use a dual-branch model to learn the representation. A series of experiments is conducted on the HAD dataset. The experiments illustrate that the proposed method outperforms other baselines in terms of accuracy and generalization, which proves the effectiveness of the proposed method. Some reasons for the excellent performance are given and analyzed. For future work, we would like to further optimize the SDF and extend our method to fake audio localization.

6. ACKNOWLEDGMENTS

This research was partly supported by the China National Nature Science Foundation (No.62006223, No.61971419 and No.61876214).

7. REFERENCES

[1] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D.Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Odyssey*, 2018, pp. 195–202.

[2] T. Kinnunen, J. Lorenzo-True.Ba, J. Yamagishi, T. Toda, and Z. Ling, "A spoofing benchmark for the 2018 voice conversion challenge: Leveraging from spoofing countermeasures for speech artifact assessment," in *Odyssey*, 2018, pp. 187–194.

[3] J. Lorenzo-Trueba, F. Fang, X. Wang, I. Echizen, J. Yamagishi, and T. Kinnunen, "Can we steal your vocal identity from the internet?: Initial investigation of cloning obama's voice using gan, wavenet and low-quality found data," in *Odyssey*, 2018, pp. 240–247.

[4] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[5] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *Interspeech*, 2015, pp. 2062–2066.

[6] M. Todisco. H. Delgado and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients," in *Odyssey*, 2016, pp. 283–290.

[7] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "Stc antispoofing systems for the asvspoof2019 challenge," in *Interspeech*, 2019, pp. 1033–1037.

[8] Yexin Yang, Hongji Wang, Heinrich Dinkel, Zhengyang Chen, Shuai Wang, Yanmin Qian, and Kai Yu, "The sjtu robust anti-spoofing system for the asvspoof 2019 challenge," in *Interspeech*, 2019, pp. 1038–1042.

[9] B Jma, B Jaa, and A Thf, "Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers," *Computer Speech and Language*, vol. 63, pp. 101096, 2020.

[10] Z. Wu, R. K. Das, J. Yang, and H. Li, "Light convolutional neural network with feature genuinization for detection of synthetic speech attacks," in *Interspeech*, 2020, pp. 1101–1105.

[11] Rui Yang, Zhenhua Qu, and Jiwu Huang, "Detecting digital audio forgeries by checking frame offsets," in *Proc. of the 10th ACM Multimedia and Security Workshop*, 2008, pp. 21–26.

[12] Rui Yang, Zhenhua Qu, and Jiwu Huang, "Exposing mp3 audio forgeries using frame offsets," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 8, no. 2, pp. 1–20, 2012.

[13] Catalin Grigoras, "Digital audio recording analysis: The electric network frequency (enf) criterion," *International Journal of Speech Language and The Law*, vol. 12, pp. 63–76, 2005.

[14] Daniel Patricio Nicolalde Rodriguez, José Antonio Apolinario, and Luiz Wagner Pereira Biscainho, "Audio authenticity: Detecting enf discontinuity with high precision phase analysis," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 3, pp. 534–543, 2010.

[15] Xunyu Pan, Xing Zhang, and Siwei Lyu, "Detecting splicing in digital audios using local noise level estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 1841–1844.

[16] Jiaorong Chen, Shijun Xiang, Weiping Liu, and Hongbin Huang, "Exposing digital audio forgeries in time domain by using singularity analysis with wavelets," in *Proceedings of the 2013 ACM Information Hiding and Multimedia Security Workshop*, 2013, pp. 149–158.

[17] Jiangyan Yi, Ye Bai, Jianhua Tao, Zhengkun Tian, Chenglong Wang, Tao Wang, and Ruibo Fu, "Half-truth: A partially fake audio detection dataset," *arXiv eprint arvix:2104.03617*, 2021.

[18] Y. Ghanbari and M. R. Karami-Mollaei, "A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets," *Speech Communication*, vol. 48, no. 8, pp. 927–940, 2006.

[19] J. W. Seok and K. S. Bae, "Speech enhancement with reduction of noise components in the wavelet domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1997, pp. 1323–1326.

[20] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," in *Interspeech*, 2017.

[21] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li, "Aishell-3: A multi-speaker mandarin tts corpus and the baselines," *arXiv eprint arxiv:2010.11567*, 2020.

[22] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, and Msa Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Interspeech*, 2015.