AUDIO REPLAY SPOOF ATTACK DETECTION USING SEGMENT-BASED HYBRID FEATURE AND DENSENET-LSTM NETWORK

Lian Huang, Chi-Man Pun

Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macau 999078, China (e-mail: yb77454@umac.mo; cmpun@umac.mo)

ABSTRACT

At present, most automatic speaker verification (ASV) systems are vulnerable to replay spoof attacks. Therefore, this paper proposes a new approach for the detection of audio replay spoof attacks. Here, a segment-based hybrid feature extraction method is used, which includes the Mel-frequency cepstral coefficient (MFCC) features and Constant-Q cepstral coefficients (CQCC) features. Then, hybrid features are trained using a variety of deep learning networks, including DenseNet, LSTM, and DenseNet-LSTM hybrid architectures. Experiments using the DenseNet-LSTM model with mixed features framework achieves the best performance. Compared to the baseline system built on the CQCC and Gaussian mixture model (GMM), the proposed method achieved 64.31% relative improvement.

Index Terms— Speaker verification, replay spoof attack detection, DenseNet, LSTM, ASV spoof

1. INTRODUCTION

In recent years, the products based on automatic speaker verification (ASV) system have experienced explosive growth. It brings a lot of convenience to people's lives, but it also brings risks [1]. The highest risk is that spoofed speech may gain unauthorized access. The ASV system is vulnerable to spoofing attacks, especially the replay spoof attacks. Therefore, the genuine and spoofing discriminative ability is one of the key issues in multimedia information security. Replay spoof attack can be performed without any professional knowledge or additional device but just a mobile phone. In addition, the detection of audio replay spoof attack is more difficult than the other attacks and it poses the greatest threat to the ASV system [2]. So, in this paper, we only focus on detecting audio replay spoof attack.

Due to the many researcher's great efforts the performance of spoofing detection has improved significantly. In the 2015 'Automatic Speaker Verification Spoofing and Countermeasures Challenge' (ASV spoof 2015), the best system achieved Equal Error Rate (EER) of 1.211% [3]. Unfortunately, the replay attack is not included in the ASV spoof 2015 dataset. On the other hand, the ASVspoof Challenge 2017 [4] was only focused on audio replay spoof attack detection task for the ASV system. In this challenge, the baseline system which is based on CQCC feature and Gaussian mixture model (GMM) [5] just got 30.60% EER on average in test data after trained by only training data alone. Shaik at al [6] proposed a new approach for replay spoof detection and got a better result of 19.77% EER with training and development data in the same dataset. In particular, the rapid development of deep learning networks in recent years has benefited the field of multimedia information security. By applying CNN and RNN networks, a low EER are also achieved under noisy conditions [7]. Even if the literature shows that great progress has been made in this field, the problem is far from being solved, ASV systems remain vulnerable to replay spoofing.

To address the issue, we explore specifically the intervals between the genuine and replay speech. And then propose a novel feature extraction method and a deep learning architecture as a classifier in this paper. The contributions of this paper are as follows:

1) A segment-based hybrid feature extraction method is proposed, which can be used to get more promising results in detecting the audio replay spoof attack.

2) Compared to traditional constant Q transform (CQT), MFCC, CQCC as one type of speech feature, the segment-based hybrid feature is much better for distinguishing between the genuine speech and the replay spoof speech.

3) A novel DenseNet-LSTM architecture, which is used to better learn the essential features and to differentiate between the genuine and replay spoof speech, is designed with the advantage for accelerating the convergence speed and preventing over-fitting.

The rest of the paper is organized as follows. In Section 2 we propose a segment-based hybrid feature extraction method. Section 3 introduces the main concept of the backend classifier with deep learning architecture. The dataset and baseline system with the experimental set-up and results are described in Section 4. Finally, we get the conclusions in Section 5.

2. SEGMENT-BASED HYBRID FEATURE EXTRACTION

The difference between the genuine and replay speech may be determined by the different implementation process. As

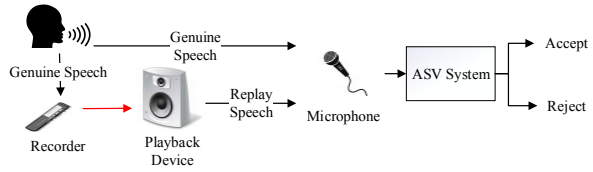

Fig.1 The Progress of Replay Attack

shown in Fig. 1 From this process, it can be clearly seen that replay speech has two more steps than genuine speech, that is, recording and playback. Although there are many different types of recording and playback devices, replay speech certainly carries the characteristics of both. Therefore, unlike speech recognition, replay spoof detection should focus not on the content of the speech or the timbre characteristics of the speaker, but on whether it has the sound characteristics of the recording and playback device. By the only one type of traditional features, such as CQT [8], MFCC, CQCC, the performance of detection for audio replay spoof attack is not convinced [9]. It is not enough to represent the replay spoofing speech. So, we propose a novel feature extraction method in this section.

At first, the speech signal is divided into two parts, some of them are parts with high zero-crossing rate and low energy, which are considered to be the relatively silent frames and others are low-interference and high-energy part, which are approximately considered as speech frames. The calculation obtains short-term zero-crossing rate and short-term energy and uses a variable threshold to perform similar endpoint detection processing. Secondly, in the relatively silent frames, the genuine and replay speech have a large difference, so the MFCC feature is extracted. Simultaneously, in the speech frames, relatively high frequency(3k-8kHz) CQCC features are extracted. Finally, combining these two characteristics, as the hybrid feature of speech, prepare for the next analysis. The specific process is shown in Fig. 2.

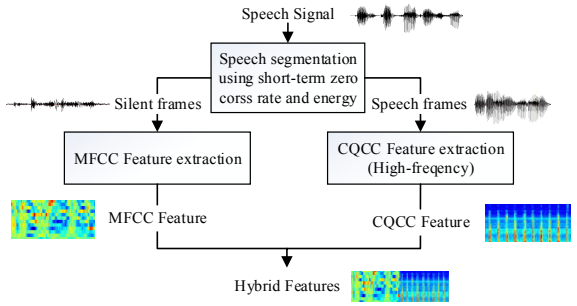

Fig.2 Segment-based Hybrid Feature Extraction

2.1. Speech segmentation using short-term zero cross rate and energy

In the experiment, the human ear can easily distinguish between genuine and replay speech. It is mainly because of that the silent segment of replayed speech has very noticeable background noise. The massive difference between the two in the silent segment can also be easily recognized on the waveform. Fig.3 is the waveform of the genuine and replay spoof speech with the same sentence that the same person said. So, if all the silent segments can be found accurately, it will be easy to distinguish the spoof speech from genuine speech. Unfortunately, accurate detection of silent segments is still difficult. However, an approximate silent segment can be found instead. Inspired by the endpoint detection method in speech processing, the speech segment is segmented by a combination of short-term zero-crossing rate and energy.

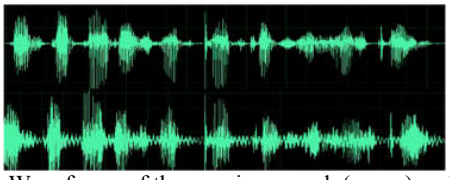
Fig.3 Waveforms of the genuine speech (upper) and the replayed speech (lower)

Short-Term Zero Crossing Rate (ZCR) is the number of times a speech signal passes through the zero points (from positive to negative or from negative to positive) in each frame.

$$st_{zcr} = \frac{1}{T-1}\sum_{t=1}^{T-1}\pi\{S_t S_{t-1} < 0\} \quad (1)$$

Where S is the value of the sample point, T is the frame length, and the function $\pi\{A\}$ is 1 when A is true, otherwise 0.

The energy of a speech signal changes with time and short-term energy refers to the average energy over a short period of time. Defining the short-term average energy of a speech signal at time n is :

$$E_n = \sum_{m=-\infty}^{+\infty}[S(m)W(n-m)]^2 = \sum_{m=n-(N-1)}^{n}[S(m)W(n-m)]^2 \quad (2)$$

Where N is the window length, and W may be a rectangular window, a Hamming window, or the like.

Judging from the zero-crossing rate, the zero-crossing rate of the unvoiced segment and the background segment is large. In another aspect, the energy is the lowest in the silent segment, while the energy is the highest in the voiced segment and the average energy of the unvoiced segment is in the first two. Therefore, the short-time energy threshold and the zero-crossing rate threshold can be set, and the speech segment and the relatively silent signal segment can be comprehensively judged. If the length of the silent segment or the speech segment is less than 100ms, adjust the threshold until a signal of sufficient length is found. As shown in Figure 4, blue is the speech signal, red is the short-time zero-crossing rate, and pink is the short-term energy. After setting the threshold, the green area is approximated as the silent segment, and the non-green area is the speech segment.

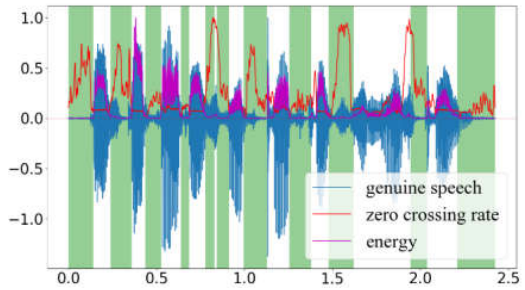

Fig. 4 Short-term zero-crossing rate and energy segmentation

2.2. MFCC and CQCC Features extraction

Mel Frequency Cepstral Coefficients (MFCCs) is a feature widely used in automatic speech and speaker recognition [10], [11]. The extraction process mainly includes pre-emphasis (sub-frame and adding window), Fast Fourier Transform, absolute value and square operation, and Mel-scaled triangle filters, a Logarithmic operation, and Discrete cosine transform. Here, MFCC feature is extracted in the approximate silent segment.

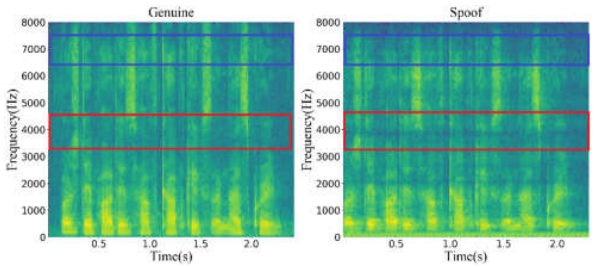

Fig. 5. The spectrograms of the genuine and replayed spoof speech

CQCC analyzes and processes speech signals based on CQT and traditional cepstrum. More details about CQCC can be found in [12]. Fig.5 presents an example of spectrograms of a genuine speech and its corresponding replay speech. From the observation of the spectrograms, the genuine and spoof speech are very close in the low band, with only some differences in the high band (the red box and the blue box in Fig. 5). Therefore, the voice segment can be extracted the CQCC feature at high frequency, i.e. 3kHz to 8kHz.

3. THE DENSENET-LSTM ARCHITECTURE CLASSIFIER

Deep learning structure has achieved great progress in many applications including audio spoof detection [13]. But the direct use of CNN, DNN or RNN architecture in this area does not yield convincing results [14], [15]. So, this paper proposes a new classification model in the form of deep learning.

When using a deep network, it is often necessary to have enough samples, and in order to improve the performance of the deep network, the number of layers of the network structure will increase accordingly. However, the problem of gradient disappearance limits the increase in the number of network layers [16]. DenseNet uses a subtler structure, which not only greatly reduces the possibility of gradient disappearance, but also reduces the network parameter size [17]. More importantly, DenseNet can still have good resistance to overfitting without enough samples, which is especially important in the replay spoof attack detection.

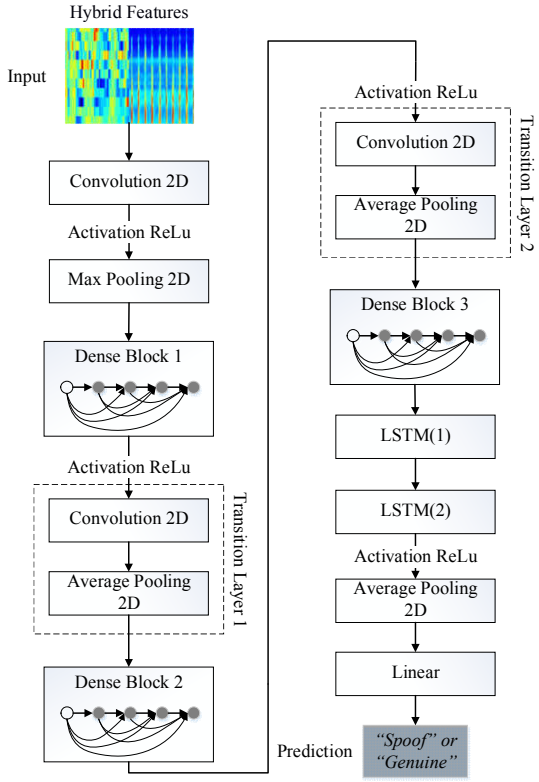

Fig. 6. Architecture of DenseNet-LSTM

we use DenseNet instead of CNN as classifier directly. Since the DenseNet has more layers (40 layers and above), its convergence speed is faster. To improve the performance of the classifier, based on DenseNet, the Long Short-Term Memory (LSTM) network was added. The specific method is to modify the last layer of DenseNet and add two more LSTM layers before the linear layer. We call this architecture as DenseNet-LSTM. The detailed architecture is shown in fig.6. In DenseNet-LSTM, Hybrid features (Extracted in section 2) are as input. Then follow a convolutional layer and max pooling layer. After that, there is a dense block and transition layer. After repeat dense block and transition layer, there is

another dense block. Then we add two LSTM layers. Finally, add a pooling layer and liner output. After this classifier, we will get the result "genuine" or "spoof" to identify it is a genuine speech or a replay spoof speech.

4. EXPERIMENTAL EVALUATION

4.1. Dataset and Baseline system

This paper uses ASVspoof 2017 dataset [18] in the experiment. This dataset is mainly for replay spoof attack. All audio signals have a resolution of 16 bits and a sampling rate of 16 kHz, and the utterance lasts approximately 3 to 5 seconds. Its summary is as follows:

Table 1. Description of ASVspoof 2017 Dataset

subset	#speakers	#utterances	
		#genuine	#spoofed
Training	10	1508	1508
Devel.	8	760	950
Eval.	24	1298	12008
Total	42	3566	14466

The baseline system is a reference implementation based on the CQCC feature and the mixed Gaussian model proposed by the ASVspoof organizer in 2017 [4]. The CQCC feature is 90-dimensional, while the GMMS is a standard class 2 classifier for classifying genuine and replay spoof speech. For each utterance, a log-likelihood score is obtained from both models, and the final system score is calculated as a log-likelihood ratio.

4.2. Experiment setup and results

We need to use the Hamming window with a size of 10ms to calculate the short-time zero-crossing rate and short-term energy and then set the threshold to find the silent segment and the speech segment. The length of the silent segment extraction is 512ms, and the length of the speech segment is 1525ms. The segment with insufficient length will be duplicated. The CQCC feature is then extracted from the silent segment and after setting at a frequency of 3k-8kHz, it will get 60*78 features. The speech segment extracts the MFCC feature and selects the Hamming window. The size of the window is 50ms, and the step size is 25ms, that is, there is a 25ms overlap. Select 48 coefficients.

After extracting the CQCC feature and the MFCC feature, the two features are combined to obtain a 60×126 feature, where 60 is the length and 126 is the feature dimension. Use this feature as input to the backend classifier such as DenseNet-LSTM. Details of DenseNet-LSTM architectures are shown in Table 2.

We train the system from Training subset and evaluate in Dev subset. Then Eval subset is as test data, and different systems are trained by Training subset (T) or Training and Dev subset (T+D). The results are shown in Table 3. We evaluate the hybrid feature and classifier with two group. One group is with the same hybrid feature and different classifier and the other group is with different features and classifiers. the segment-based hybrid feature with DenseNet-LSTM classifier gets the best performance.

Table 2. Details of DenseNet-LSTM architectures

Layers	Output Size	Layer config
Convolution	30×63	7×7 conv, stride 2
Pooling	15×32	3×3 max pool, stride 2
Dense Block 1	15×32	$\begin{bmatrix} 1\times 1\ conv \\ 3\times 3\ conv \end{bmatrix}\times 6$
Transition Layer 1	15×32	1×1 conv, stride 1
	7×16	2×2 average pool, stride 2
Dense Block 2	7×16	$\begin{bmatrix} 1\times 1\ conv \\ 3\times 3\ conv \end{bmatrix}\times 6$
Transition Layer 2	7×16	1×1 conv, stride 1
	3×8	2×2 average pool, stride 2
Dense Block 3	3×8	$\begin{bmatrix} 1\times 1\ conv \\ 3\times 3\ conv \end{bmatrix}\times 6$
LSTM	1×48	2 Layers
Classification Layer	1×1	1×48 global average pool
	(None,1)	Linear

Table 3. Evaluation results

Individual System	EER(%)		
	Dev set	Eval set(T)	Eval set(T+D)
Baseline (CQCC+GMM)	10.83	30.60	24.77
Hybrid Feature+GMM	8.67	25.63	18.11
Hybrid Feature+DenseNet	5.62	12.39	11.08
Hybrid Feature+LSTM	9.45	15.64	14.78
Hybrid Feature+DenseNet-LSTM	**3.62**	**9.56**	**8.84**
CQCC+DenseNet	7.65	17.73	15.27
MFCC+DenseNet	6.77	15.86	13.45
CQCC+DenseNet-LSTM	3.87	12.64	11.67

5. CONCLUSIONS

In this study, we proposed a new approach to detect the audio replay spoof attack. The experimental results show that it can achieve the best performance. For ASV systems, this is a very promising system performance and closer to practical. However, the method heavily depends on the segment-base hybrid feature and this feature is based on the background sound of different record and playback device. If there is high-quality hardware it may be hard to distinguish the genuine and replay speech. Under this condition, we may need to explore further in future work.

6. REFERENCES

[1] S. Kucur Ergunay, E. Khoury, A. Lazaridis, S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing". In *Biometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference* on (pp. 1-6). IEEE.

[2] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, H. Li, "Spoofing and counter measures for speaker verification: A survey," *Speech Comm. (Elsevier)*, vol. 66, pp. 130–153, Feb. 2015.

[3] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Interspeech 2015*, 2015, pp. 2037–2041.

[4] Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Kong Aik Lee, Md Sahidullah, Massimiliano Todisco, and Héctor Delgado, "Asvspoof 2017: automatic speaker verification spoofing and countermeasures challenge evaluation plan," *Training*, vol. 10, no. 1508, pp. 1508, 2017.

[5] D. Reynolds, T. Quatieri, and R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, nos. 1-3, pp. 19-41, 2000.

[6] B. S. M. Rafi, K. S. R. Murty, and S. Nayak, "A new approach for robust replay spoof detection in ASV systems," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Montreal, Canada, 2017, pp. 51–55.

[7] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition", *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 24, pp. 2263–2276, Aug. 2016.

[8] J. C. Brown, M. S. Puckette. (1992). An efficient algorithm for the calculation of a constant Q transform. *The Journal of the Acoustical Society of America*, 92(5), 2698-2701.

[9] M. Singh, J. Mishra, & D. Pati, (2016, December). Replay attack: Its effect on GMM-UBM based text-independent speaker verification system. In *Electrical, Computer and Electronics Engineering (UPCON), 2016 IEEE Uttar Pradesh Section International Conference on* (pp. 619-623). IEEE.

[10] R. Vergin, D. O'shaughnessy, & A. Farhat, (1999). Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition. *IEEE Transactions on speech and audio processing*, 7(5), 525-532.

[11] C. Ittichaichareon, S. Suksri, & T. Yingthawornsuk, (2012, July). Speech recognition using MFCC. In *International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012)* July (pp. 28-29).

[12] Massimiliano Todisco, Héctor Delgado, and Nicholas Evans, "A new feature for automatic speaker verification anti-
spoofing: Constant q cepstral coefficients," in *Speaker Odyssey Workshop*, Bilbao, Spain, 2016, vol. 25, pp. 249–252.

[13] S. Scardapane, L. Stoffl, F. Röhrbein, & A. Uncini, (2017, May). On the use of deep recurrent neural networks for detecting audio spoofing attacks. In Neural Networks (IJCNN), 2017 International Joint Conference on (pp. 3483-3490). IEEE.

[14] Chen, Z., Xie, Z., Zhang, W., & Xu, X. (2017, August). Resnet and model fusion for automatic spoofing detection. In Proc. *Interspeech* (pp. 102-106).

[15] Dinkel, H., Qian, Y., & Yu, K. (2018). Investigating Raw Wave Deep Neural Networks for End-to-End Speaker Spoofing Detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(11), 2002-2014.

[16] K. He, X. Zhang, S. Ren, & J. Sun, (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[17] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. arXiv preprint arXiv:1608.06993, 2016.

[18] Tomi Kinnunen, Md Sahidullah, Mauro Falcone, Luca Costantini, Rosa González Hautamäki, Dennis Thomsen, Achintya Sarkar, Zheng-Hua Tan, Héctor Delgado, Massimiliano Todisco, et al., "Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2017*.