Multimedia Tools and Applications (2019) 78: 8383–8396
https://doi.org/10.1007/s11042-018-6834-3

Replay attack detection based on distortion by loudspeaker for voice authentication

Yanzhen Ren[1] · Zhong Fang[2] · Dengkai Liu[1] · Changwen Chen[3]

Received: 16 July 2018 / Revised: 21 September 2018 / Accepted: 30 October 2018 /
Published online: 7 November 2018

Abstract

Identity authentication based on Automatic Speaker Verification (ASV) has attracted extensive attention. Voice can be used as a substitute of password in many applications. However, the security of current ASV systems has been seriously challenged by many malicious spoofing attacks. Among all those attacks, replay attack is one of the biggest threats to the ASV System, where an adversary can use a pre-recorded speech sample of the legal user to access the ASV system. In this paper, we present a replay attack detection (RAD) scheme to distinguish normal speech and replayed speech. We focus on the distortion caused by loudspeaker: low-frequency attenuation and high-frequency harmonics, and present a suite of RAD features DL-RAD, including Harmonic Energy Ratio (HER), Low Spectral Ratio (LSR), Low Spectral Variance (LSV), and Low Spectral Difference Variance (LSDV), to describe the different characteristics between the normal speech signal and replay speech signal. SVM is adopted as a classifier to evaluate the performance of these features. Experiment results show that the True Positive Rate (TPR), True Negative Rate (TNR) of the proposed method are about 98.15% and 98.75% respectively, which are significantly better than the existing scheme. The proposed scheme can be applied to both text-dependent and text-independent ASV systems.

This work is supported by the Natural Science Foundation of China (NSFC) under the grant NO. U1536114, NO. 61872275, NO.U1536204, and China Scholarship Council.

✉ Yanzhen Ren
renyz@whu.edu.cn

Zhong Fang
fangzhong@ict.ac.cn

Dengkai Liu
dengkailiu@whu.edu.cn

Changwen Chen
chencw@buffalo.edu

[1] Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan, China

[2] Chinese Academy of Sciences, Institute of Computing Technology, Beijing, China

[3] University at Buffalo, the State University of New York, Buffalo, NY 14260, USA

Keywords Automatic Speaker Verification (ASV) · Replay Attack Detection (RAD) · Loudspeaker · Low-frequency attenuation · Spoofing attack

1 Introduction

Voice, as one of the most important means of communication for human communication and human-computer interaction, has been used as an important biometric features. Automatic Speaker Verification (ASV) is to determine whether a person is who he or she claims from the personal characteristics of voices [10]. Due the portability, stability and privacy of the voice features, ASV has attracted extensive attention and application in recent years, including the smart phone logical access, e-commerce, and authentication of telephone banking, etc. For example, Google has provided a "Trusted voice" function into Android operator system to allow the owners to unlock their phone with their voice [9]. SayPay Technologies, Inc. has integrated its mPayment platform with voice biometrics platform, allowing the customers to use a voice password to conduct mPayment transactions in card-not-present environments [3]. UK high-street bank(HSBC) announced that it would offer biometric banking software to access online and phone accounts using their fingerprint or voice [6].

However, the security of current ASV systems has been seriously challenged by many malicious spoofing attacks [20]. There are four kinds of spoofing attacks to ASV system: impersonation, speech synthesis, voice conversion, and replay attack [20]. Impersonation is a method that the attacker tries to mimic the target genuine speakers voice by himself. Speech synthesis is the artificial production of human speech, create the target genuine speakers speech based on the model by the computer. Voice conversion is to modify the attackers voice to sound as the target genuine speakers. Replay attack is to collect and record the target genuine speakers speech, then play back this pre-recorder speech to pass through the ASV system. Among all those spoofing attacks, replay attck is the easiest and most effective one. There are two benefits for a replay attacker: At first, it is so easy to realize. There are too many low-cost recording devices, such as smartphone, recording pen, etc., to capture the legal user's voice and play it back. Even if you need to cut and stitch the audios to access the random digital ASV system, there are many free tools to edit them. The second is due to the high similarity on the spectrogram and formant tracks between the original and replay speech, which are the basic features used for speaker verification. There are two types of voice authentication systems: text-dependent and text-independent. Text-dependent systems used fixed or prompted password phrases, which are usually the same for enrolment and verification. Text-independent systems verify arbitrary password phrases [12]. Text-dependent system is easier to be integrated into commercial application, but it's more vulnerable to spoofing attacks. The first replay attack method had been proposed in [7] to evaluate the risk of ASV system. In [14, 17], the influences of replay attack on ASV system are also evaluated. All the studies [7, 14, 17] are found to be consistent in their findings: regardless the ASV system tested, replay attacks provoke significant increases in FAR (False Alarm Rate) [20]. That is, the replay speech has been mistakenly judged as the original speech. Even Google prompts users for the risk of trusted voice"... a recording of your voice could unlock your device" [9].

There are three types of replay attack detection (RAD) methods [4, 5, 13, 15–18, 21, 22, 24]: similarity comparison, channel characteristic analysis, and liveness voice detection. The main idea of the similarity comparison is that for the normal situation, it is impossible to generate two identical speech signals at different time. If the similarity between the detected speech sample and the legal users pre-recorded speech sample is too high, it must be a replayed one. How to calculate the similarity score is the key. In [5], Shang and Maryhelen used peakmap to calculate the similarity score. In [21], Wu et al. used spectrogram bitmaps as features. Inspired by the music recognition technology, Gaka et al. used peaks per frame, pairs per peak, etc. as features [4]. Although RAD methods based on similarity comparison have good performance on text-dependent ASV system, they are based on the assumption that each passphrase spoken by the legal user has been recorded in the ASV system. The attacker must have recorded the passphrase which the legal user had spoken to the ASV system. This is too restrictive for the application, and they cannot be used in text-independent ASV system. RAD methods based on channel characteristics focus on the signal distortions introduced by the recording and speaker devices. In [24], Zhang et al. proposed to extract the channel features introduced by the high-fidelity recording equipment from mute voice. In [17, 18], He et al. extracted 6-order Legendre coefficients and 6 statistical features from channel pattern noise, which were introduced from recording and replay devices. In [15, 16], Villalba and Eduardo assume far-field recording will enhance the noise and reverberation level of speech signal, and loudspeaker will flat the spectrum and reduce the modulation indexes of speech signal, they proposed four features: Spectral Ratio, Low Frequency Ratio, Modulation Index, and Sub-band Modulation Index. RAD methods based on channel characteristics only need to detect the tested speech signal, not dependent on the stored speech, so its application will be extensive, and can be used in both of text-dependent and text-independent ASV system. In [13], Shiota et al. proposed a RAD method based on liveness voice detection to verify whether the detected speech are originated by a real human being based on the liveness evidence of the pop noise, which is introduced by the human breath. In [23], Linghan Zhang et al. proposed a RAD method based on the user’s unique physical vocal system and the stereo recording of smartphones. In a sequence of phoneme sounds, the time-difference-of-arrival (TDoA) changes to the two microphones in the smartphone was captured. This feature is unique for the user and couldn’t exist in the speech signal of replay attacks. RAD methods based on liveness detection are novel approaches, and have good performance, but there are still some limitations for these methods. For example, in [23], the scheme requires the user to hold the phone close to her/his mouth with the same pose in both enrollment and authentication processes [23], and is used in a text-dependent ASV systems. If the scheme need to be extended into the text-independent ASV systems, the TDoA value of each phoneme should be extracted, and the performance could not be predicted.

In the procedure of replay attack, the original speech signal needs to be recorded, processed(edited or synthesized), and replayed back into the ASV system as the replayed speech signal. No matter whether the original speech signal has been processed or not, and how it has been processed, the loudspeaker, as the last link of replay attack chain, plays a decisive role on the difference between original and replayed speech signal. There are many multimedia steganalysis and forensics schemes which extract features to train the detection model [8, 19, 25]. In this paper, a RAD scheme which is based on the distortion caused by Loudspeaker was proposed. The proposed scheme belongs to channel characteristics based RAD scheme, and can be applied into both of the text-dependent and text-independent ASV systems.

The rest of the paper is organized as follows. In Section 2, the procedure of replay attack is reviewed, and the influence of loudspeaker on speech signals is analyzed in detail. In Section 3, the proposed RAD features are presented and the effectiveness of each sub-features are analyzed. The experiments are carried out to evaluate and compare the performance of the proposed RAD scheme with the existing channel characteristics based scheme. The results are presented in Section 4. Finally, the conclusions are drawn and future works are discussed in Section 5.

2 The analysis of replay attack

2.1 The influence on speech signal from replay attack

The target of RAD method is to find the difference between the original and replay signals. Figure 1 shows the schematic of replay attack chain. Each link will introduce distortion on the original signal. The original speech signal is sent from human being's vocal system directly. The replay signal is reproduced from the original signal, under the processing of microphone, speech editor, speech synthesizer, and loudspeaker. Loudspeaker is the necessary processing and final link in replay chain. Whether or not the speech signal is distorted by the previous processing, loudspeaker will play a major decision role on the difference of the original and replay signal.

The regular distortions caused by loudspeaker include linear and non-linear distortions. They are directly related to the geometry and properties of the material used in loudspeaker design [1]. Linear distortion is due to the linear components in the loudspeakers circuit, characterized by the frequency response, which shows the range of audio frequencies a loudspeaker can reproduce. Figure 2 is the frequency response curve of iPhone 5 loudspeaker, it show that the low-frequency is attenuated. The lower the frequency is, the greater the attenuation will be. The phenomenon is common and determined by the physical structure of the loudspeaker. The generation of nonlinear distortion depends on the properties of the stimulus, harmonic distortion is one main kind of nonlinear distortion, which generates new spectral components at multiples of the fundamental frequency and thus increase the energy of the high frequency. In Fig. 3, there is an obvious harmonic energy on the high frequency of replay signal.

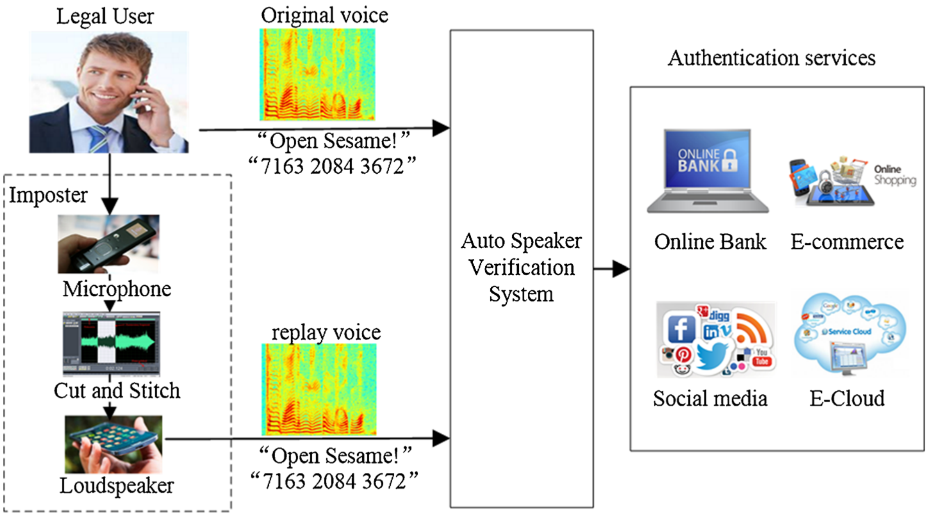


Fig. 1 The schematic of replay attack

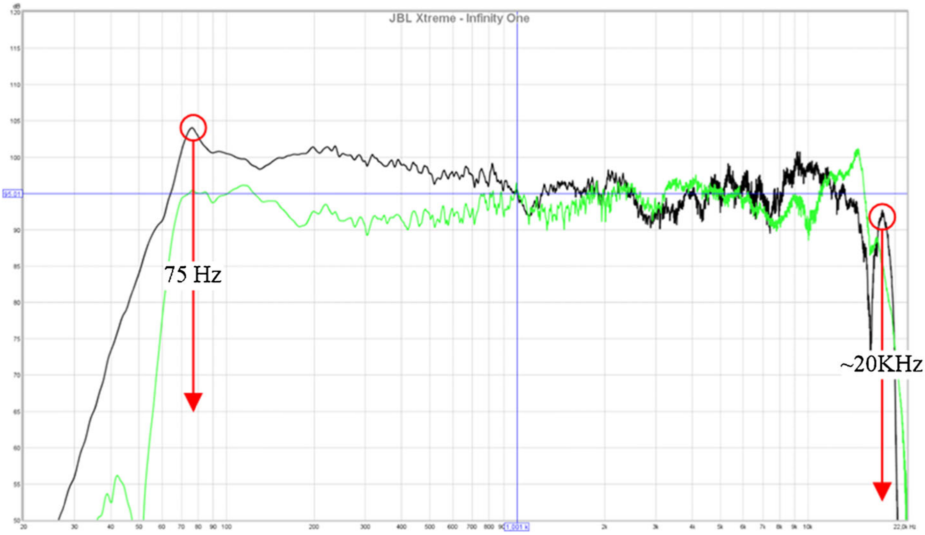


Fig. 2 The Frequency response curve of the iPhone 5 loudspeaker [11]: The black curve

2.2 The difference between original and replay signals

In an ASV system, there are three types of features used to recognize a speaker: short-term spectral and voice source features, spectra-temporal and prosodic features, and high-level linguistic features [20]. Those characteristics of replay speech signal are very much similar to the original signals. This is the reason why replay attack is easy to success. Figure 3 is an example of the spectrogram and formant tracks (F1-F4) of original (left) and replay (right) speech, a speaker pronounces "Jian Jia Cang Cang". It is obvious that the spectrogram and

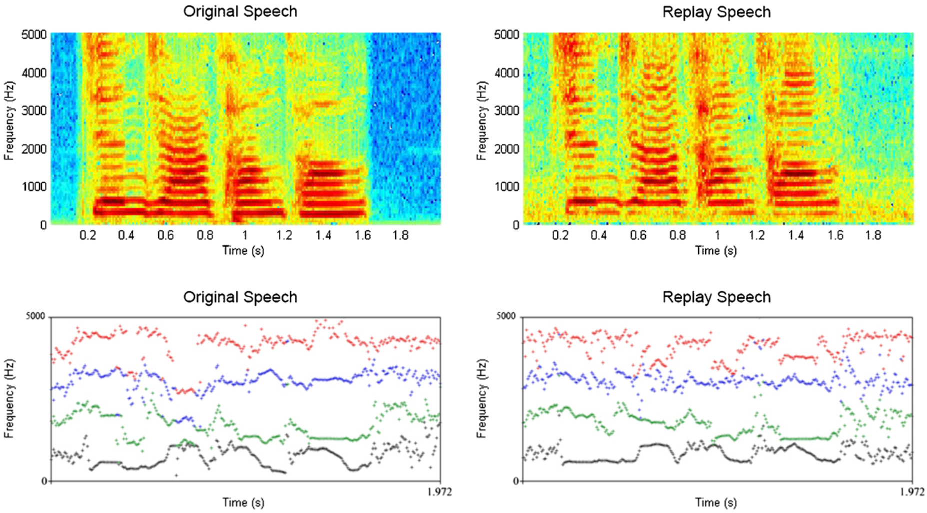


Fig. 3 Spectrogram of original speech and replay speech

formant tracks of the replay speech are very much similar to the original one. However, there are two noticeable differences in the details. First, the low frequency energy is attenuated. In Fig. 3, the color of the lowest horizon bar in original speech is darker than that of the replay signal. The lowest horizon bar in spectrogram represents the energy near the speech signals fundamental frequency, and the other horizon bars represent the energy near the signals at multiple frequencies. This implies that the energy of the low frequency zone of replay signal is attenuated, much weaker than the original signal. Figure 4 shows the spectrum of the original and the replay signal in the same time slice, the frequency power in replay signal is lower than the original signal from 0-500Hz. The lower the frequency is, the greater the attenuation will be. This difference can be explained with the frequency response curve in Fig. 2. The second difference is that there are many high-frequency harmonic in replay signal. This is due to the non-linear distortion caused by loudspeaker.

3 RAD scheme based on the distortion of loudspeaker (DL-RAD)

In this section, a new RAD scheme based on the low-frequency attenuation and high-frequency harmonics introduced by loudspeaker is proposed. Actually, in most microphones, the signals less than 40Hz will be filtered out, but it will cause little influence to the difference between original signal and replay signal in low-frequency. In this scheme, the signal under 500 Hz is analyzed, four discriminating feature are designed: Harmonic energy

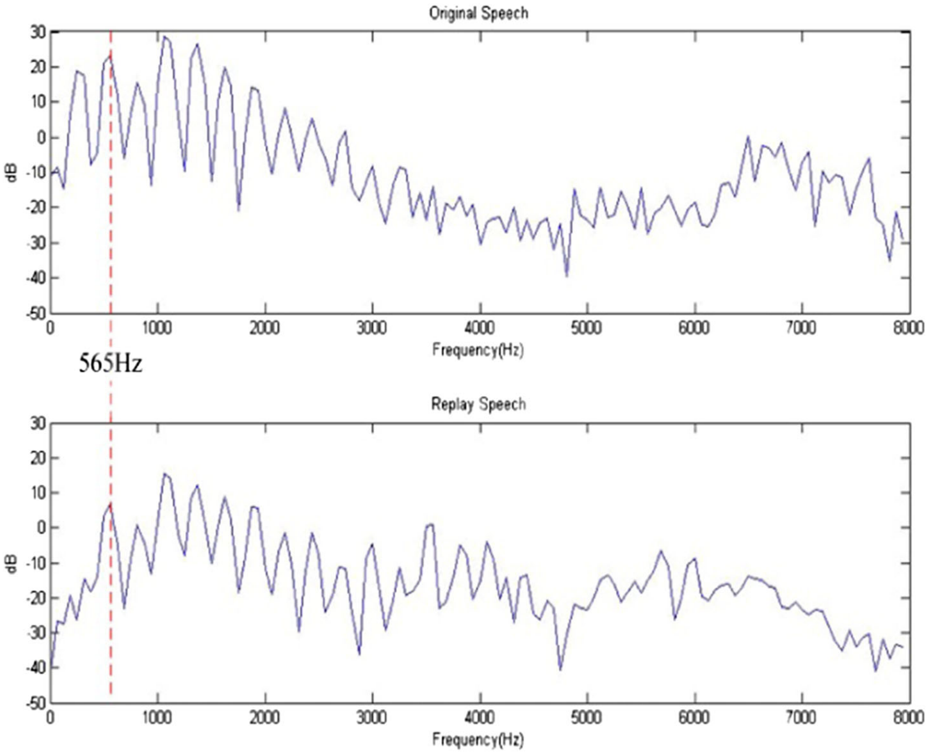


Fig. 4 Spectrum of original speech and replay speech

ratio (HER), Low Spectral Ratio (LSR), Low Spectral Variance (LSV), and Low Spectral Difference Variance (LDV). The effectiveness of these features is analyzed in the following sub-sections.

3.1 The proposed RAD features

3.1.1 Harmonic energy ratio (HER)

HER features are designed to measure the low-frequency attenuation and high-frequency harmonics introduced by loudspeaker. They actually measure the voiced signals energy proportion of the fundamental frequency and the harmonic frequencies. The energy of voiced speech signal is concentrated on a small range around the fundamental frequency and harmonic frequencies. Since that the different speakers have different fundamental frequencies, so we adopt the energy ratio based on the fundamental frequency instead of frequency zone, in order to reduce possible negative influence on the detection accuracy resulting from diversity of the speakers.

HER features are calculated as (1), where $1 \leq i \leq 15$, f_1 is the fundamental frequency and $f_2 ... f_{15}$ are the multiple harmonic frequencies, $f_i = f_1 * i$. $E(f_i) = \sum_{f_i - \triangle_f}^{f_i + \triangle_f} (X(f))^2$ is the energy of the fundamental and harmonic frequencies. $X(f)$ is the power of frequency f after FFT (Fast Fourier transform), $\triangle_f$ is a range designed to avoid the prediction error of the fundamental frequency. In our experiment, $\triangle_f$ is chosen as 10Hz. $ER(f_i) = E(f_i)/E(f_3)$ is energy ratio between each frequency and $E(f_3)$. This is because the energy of f_3 is relatively stable in both original and replay signals. $HER(f_i)$ is the normalized form of $ER(f_i)$, while $min(f)$ is the minimum value of all f_i, $max(f)$ is the maximum value of all f_i.

$$HER(f_i) = \frac{ER(f_i) - min(ER(f))}{max(ER(f)) - min(ER(f))} \tag{1}$$

3.1.2 Low Spectral Ratio (LSR)

LSR feature is designed to measure the low frequency attenuation based on the frequency range. Based on the statistics we derived a large number of speech signal, we found that, for replay speech signal, the energy in the range between 250Hz and 350Hz is clearly lower than the original speech signal, while the total power below 500 Hz is relatively stable. LSR is the ratio of energy between 250Hz and 350Hz over the total energy of 0˜500Hz as defined in (2).

$$LSR = \frac{\sum_{f=250}^{f=350} |X(f)|}{\sum_{f=0}^{f=500} |X(f)|} \tag{2}$$

3.1.3 Low Spectral Variance (LSV)

Variance has been used to describe the degree of fluctuation in a signal. Due to the attenuation of low-frequencies, the spectral signal of replay speech is flatter than the original speech below 500Hz. Low Spectral Variance (LSV) is the variance of the spectral signal below 500Hz, calculated as in (3). While EP is the average frequency power below 500Hz.

$$LSV = E((|X(f)| - EP)^2) \tag{3}$$

3.1.4 Low Spectral Difference Variance (LSDV)

The first-order difference is used to describe the continuity of adjacent data. Due to the flatness of low-frequency part in replay speech, the degree of change for adjacent frequency will be flat. The first-order difference is obtained from 15 samples within 0 ˜ 500Hz under FFT transformation. LSDV is calculated as in (4).

$$LSDV = \frac{1}{N_s}\sum_{i=I}^{N_s}(D(f_i) - |\bar{D}|)^2 \tag{4}$$

Where $D(f_i)$ is the first-order difference sequence of the signals FFT sequence under 500Hz, $0 \leq f_i \leq 500$. $\bar{D}$ is the average of $D(f_i)$. N_s is the maximum index of the samples while $f = 500$. In this research, the frame is 30ms, and N_s is 15.

3.2 The effectiveness of each features

To evaluate the effectiveness of each DL-RAD features, an audio sample database was created, with detailed description given in Section 4. Figure 5 shows the box-plot of HER features of original and replay signals. In low-frequency range, $HER(f_1)$ of the replay speech is significantly lower than that of the original signal. This is due to the low frequency attenuation introduced by loudspeaker. In high-frequency range, the energy ratio of the replay speech is higher than that of the orignal signals. The reason is due to the spurious high frequency harmonics caused by the loudspeaker. Figure 6 shows the distribution of LSR, LSV and LSDV in original and replay signals respectively. The value of LSR, LSV and LSDV of the original signal is obviously higher than that of the replay signals, These results verify the influence of the low-frequency attenuation in the replay signal.

3.3 The proposed DL-RAD scheme based on distortion of loudspeaker

The framework of the proposed DL-RAD scheme is considered as supervised machine learning. At first, the audio sample should be pre-processed by framing, windowing, and voiced frame detecting to find the voiced audio segment. Then, DL-RAD features are extracted from the trained database. Supervised classifier is used to train the classifier model. The trained model is then used to classify the original and the replay signals. In this paper, LibSVM [2] with its default parameters and RBF kernel is used as a classifier to evaluate the performance of the proposed features.

4 Experiments

To evaluate the performance of the proposed features, three experiments have been carried out. The speech databases is created from 771 WAV speech samples. There is no public speech database claims that their audio sample is original audio, so we record all the audio samples by CoolEdit. Each WAV audio is mono, 8 kHz, 16 bit quantization, 3 seconds, from different peoples, include English and Chinese speakers. All audio samples are replayed by four brand of smart phones to build four replay audio databases. The smartphones are all widely used now, include iPhone 6, Samsung Galaxy S6, Meizu MX4, and ZTE N798. The total number of original audio samples is 771. The total number of replay audio samples is 771*4 = 3084.

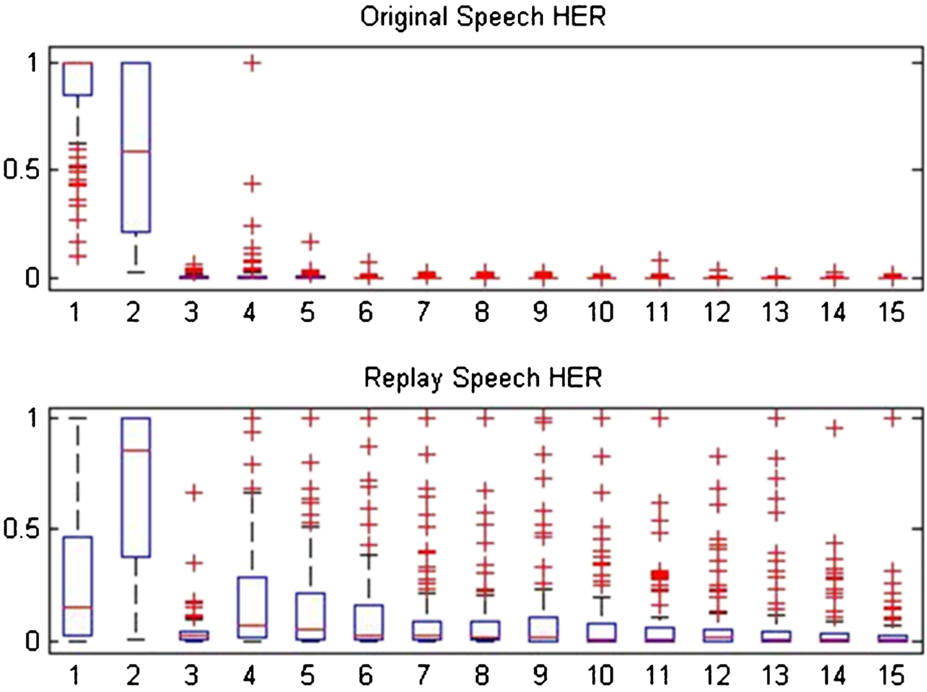


Fig. 5 The distribution of HER

Three metrics are used to evaluate the performance of the RAD methods: TPR (True Positive Rate), TNR (True Negative Rate), and ACC (Accuracy). Original speech is considered as positive sample. TPR represents the probability that original signal is judged as original. TNR represents the probability that replay signal is judged as replayed. ACC is the probability that all signals are judged correctly.

Experiment I is designed to analyze the performance of each sub-features in DL-RAD feature set. Experiment II is designed to compare the performance of DL-RAD with the existing methods. Experiment III is designed to analyze the generalization ability of DL-RAD method under different loudspeakers and compare it with the existing methods. The details and the results are shown as follows:

Experiment I To evaluate the classification ability of sub-features separately, we choose 500 positive samples from original database and 500 negative samples from replay database of iPhone randomly. Four sub-features, namely HER, LSR, LSV, and LSDV, are extracted to train the classifiers separately. These four classifiers are then used to detect the remaining audio samples in the original database and the replay database of the iPhone. TPR, TNR and ACC of all classifiers are shown in Fig. 7. The ROC (Receiver Operating Characteristic) Curves are shown in Fig. 8. These results show that TPR, TNR, and ACC of all sub-features are higher than 80%. In all sub-features, LSRs contribution is the smallest, while LSDV's contribution is the largest.

Experiment II This experiment is designed to compare DL-RAD with the existing channel characteristic based RAD schemes. Villalba and Eduardo [15, 16] proposed a RAD scheme

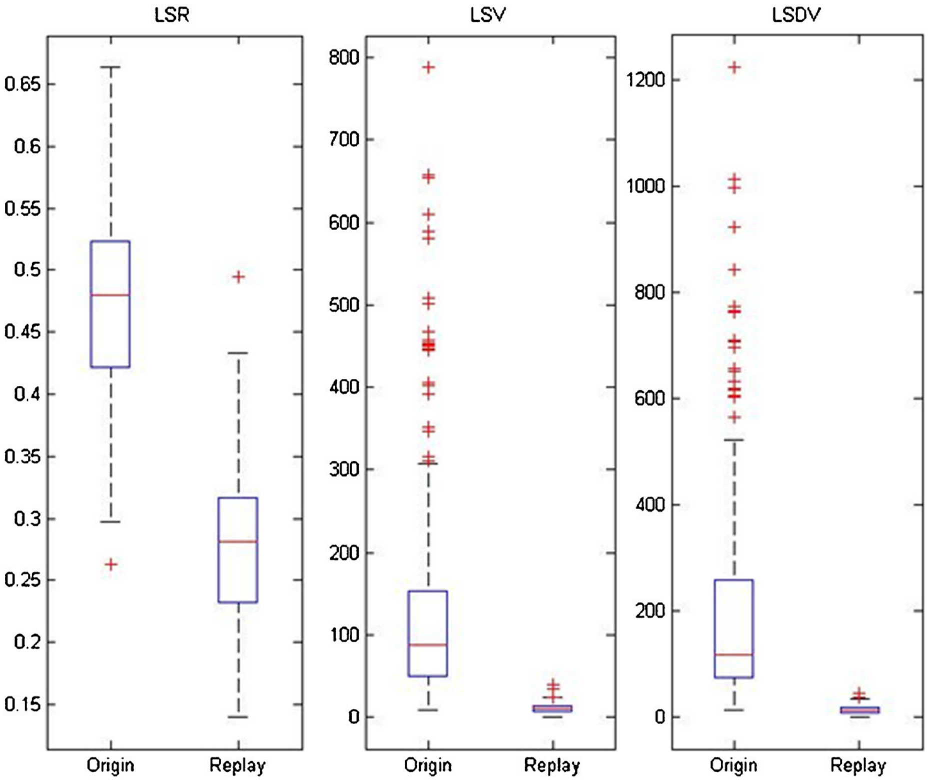


Fig. 6 The distribution of LSR, LSV, and LSDV

FFR-RAD based on the distortion of recording equipment and loudspeaker. This is the most effective channel characteristics-based scheme [20]. Other RAD schemes can not be used in text-independent ASV sytem, so we compare DL-RAD scheme with FFR-RAD scheme. 500 original audio samples and 500 replay audio samples from iPhone are selected randomly to build training set, and the remaining 271 original and 271 replay audio samples from iPhone are tested. The TPR, TNR and ACC of DL-RAD and FFR-RAD scheme are shown in Fig. 7. The ROC Curves of DL-RAD and FFR-RAD scheme are shown in Fig. 8. The results show that the TPR, TNR and ACC of DL-RAD scheme is 97.78%, 99.62%, and 98.89% respectively, all of them are higher than that of the FFR-RAD scheme, which is 84.01%, 95.86%, and 89.91% respectively. The ROC curves show that the DL-RAD scheme demonstrate much better comprehensive classification ability than that of the FFR-RAD scheme.

Experiment III Since DL-RAD scheme is based on the distortion caused by the loudspeaker, the performance will be influenced by different loudspeakers. To evaluate the generalization ability of DL-RAD and FFR-RAD scheme under different loudspeakers, four replay databases from iPhone, Samsung, Meizu, and ZTE are created. 500 original audio samples and 500 replay audio samples from each mobile-phones replay database are selected randomly to train the classifier model of M_iPh, M_Sam, M_Mei, and M_ZTE respectively. Then the trained models are used to detect the remaining 271 original audios

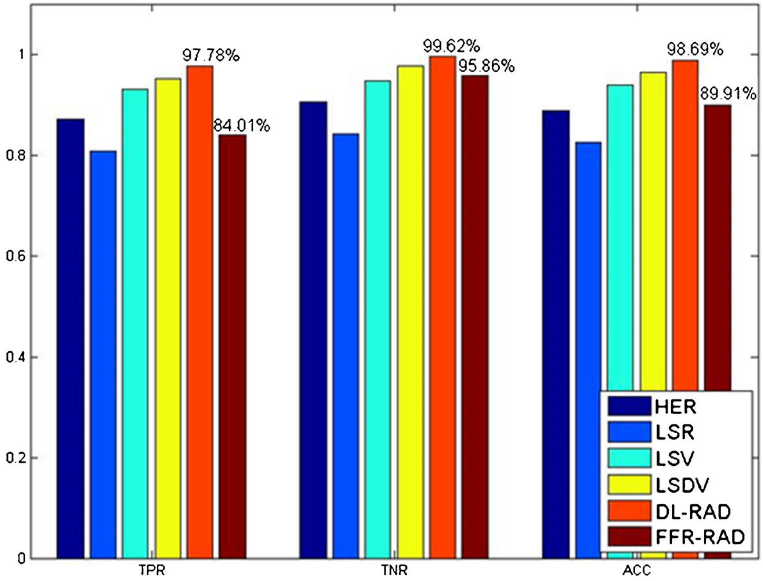


Fig. 7 The TPR, TNR and ACC of DL-RAD, sub-features and FFR-RAD

and 271 replay audios from each of the mobile-phones replay databases. In addition, we choose 125 replay audio samples from each mobile-phones replay database to build a mixture replay audio database of 500, with the 500 original audios, train a mixture classifier

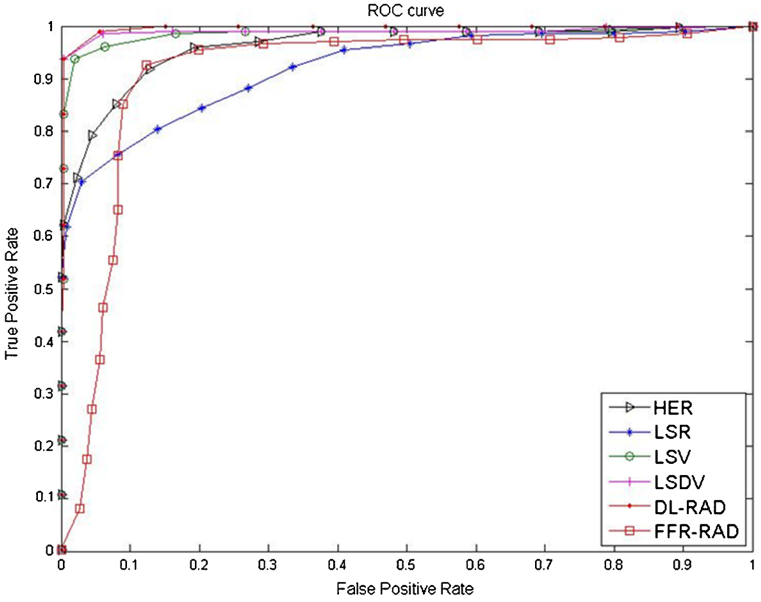


Fig. 8 The ROC curve of DL-RAD, sub-features and FFR-RAD

Table 1 TPR, TNR of each classifier from different brand of mobile phone under our proposed scheme DL-RAD (%)

D M	TPR	TNR				
		D_iPh	D_Sam	D_Mei	D_ZTE	D_Mix
M_iPh	97.72	98.15	96.31	90.77	96.68	94.25
M_Sam	96.68	94.10	96.68	88.93	90.77	91.25
M_Mei	95.94	88.56	94.10	93.73	92.99	92.25
M_ZTE	89.67	88.19	92.99	94.83	100	93.00
M_Mix	98.15	97.79	95.57	93.36	100	98.75

M_mix, and then detect the remaining original and replay audio samples from each mobilephones replay database. Tables 1 and 2 show these results. Where "M" represents classifier model, "D" represents tested audio database. D_(x) is the replay audio samples selected from (x) brand mobile phones, D_Mix is the tested replay audio samples selected from each brand mobile phones randomly. M_(x) is the classifier which is trained from the orginial audio samples and the replay audio samples of (x) brand mobile phones, M_Mix is the classifier which is trained from the orginial and replay audio samples of all brand mobile phones.

Table 1 shows the performance of DL-RAD. The maximum TPR of all these separated classifiers is 97.72%, and the minimum is 89.67%. Each separated classifier has the highest accuracy on the audio samples from their own brand. The TNR of D_Mix under M_mix is 98.75%. This shows that the classifier trained from the data with high diversity will perform better. The TNR of D_ZTE under M_ZTE and M_Mix is 100%, it is due to the distortion of ZTE loudspeaker is bigger than the other smartphones, and the number of audio samples is small. Table 2 shows the performance of FFR-RAD. The maximum TPR of all separated classifiers is 89.30%, and the minimum is 79.34%. These are lower than the TPR of DL-RAD. The TNR of D_Mix under M_mix is 82.25%. This is also lower than that of DL-RAD.

The above experiments show that TPR and TNR of the proposed method have achieved a good performance. The experiment to compare with the existing scheme FFR-RAD shows that the proposed method outperforms the existing method. The generalization ability of the proposed scheme for different loudspeakers is also excellent, and also noticeably better than the existing scheme.

Table 2 TPR, TNR of each classifier from different brand of mobile phone under FFR-RAD [15] (%)

D M	TPR	TNR				
		D_iPh	D_Sam	D_Mei	D_ZTE	D_Mix
M_iPh	82.29	84.87	75.65	81.18	83.39	78.25
M_Sam	89.30	78.23	92.25	83.76	90.77	83.50
M_Mei	79.34	85.24	83.03	87.82	77.49	63.25
M_ZTE	80.44	79.70	84.13	84.87	91.88	79.25
M_Mix	85.98	84.13	79.70	83.39	88.93	82.25

5 Conclusion

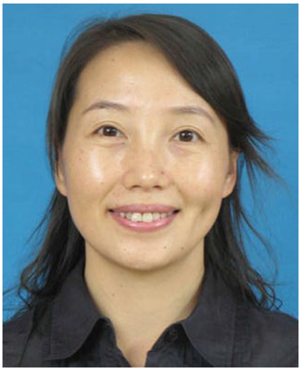
Replay attack is one of the biggest threats to ASV System. The paper proposes a novel replay attack detection scheme DL-RAD, extracts the features from the signal distortion caused by loudspeaker, which is the last link of replay attack chain and plays a decisive role on the replay signal. Four sub-features, HER, LSR, LSV, and LSDV, have been proposed. SVM is adopted as classifier to evaluate the performance of these features. Experimental results show that the performance of the proposed scheme in terms of TPR, TNR and ACC is significantly better than that of the existing channel characteristic based RAD method. The generalization ability of the proposed scheme for different loudspeakers has also been studied, and have a good performance. Since the proposed DL-RAD scheme is based on the tested speech signal itself, and has no relative to the content of the speech, so it can be used in both text-dependent and text-independent ASV system. Loudspeaker is used in most speech spoofing attack, e.g. speech synthesis, or voice conversion. We expect that the principles of the proposed scheme can be applied to analyze signal distortion of those replayed synthesized speech signal caused by loudspeaker, in order to discover the different characteristics between original speech and the replayed synthesized speech.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Brown S (2006) Linear and nonlinear loudspeaker characterization. Ph.D. thesis, Citeseer
2. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines (ACM)
3. findblometrics (2015) Voicevault biometrics to protect payments. https://findbiometrics.com/voicevault-biometrics-to-protect-payments-25131/
4. Gaka J, Grzywacz M, Samborski R (2015) Playback attack detection for text-dependent speaker verification over telephone channels. Speech Comm 67:143
5. Koga S, Makihara S, Yamanouchi Y (2010) In: IEEE international conference on acoustics speech and signal processing, pp 1678–1681
6. Kollewe J (2016) Hsbc rolls out voice and touch id security for bank customers–business. The Guardian
7. Lindberg J, Blomberg M (2012) Vulnerability in speaker verification - a study of technical impostor techniques
8. Ma Y, Luo X, Li X, Bao Z, Zhang Y (2018) Selection of rich model steganalysis features based on decision rough set α-positive region reduction. IEEE Trans Circ Chapman Hall/CRC Syst Video Technol PP(99):1
9. MPF (2015) DAILYMAIL.COM. Android can now unlock your phone when it hears your voice. http://www.dailymail.co.uk/sciencetech/article-3037733/OK-Google-Android-unlock-phone-hears-voice.html
10. Reynolds DA (2002) An overview of automatic speaker recognition technology 4, IV
11. (2015) Review: Jbl xtreme - how much bass can you handle? http://www.oluvsgadgets.net/2015/07/review-jbl-xtreme-how-much-bass-can-you-handle.html
12. Shen W, Khanna R (1997) Prolog to speaker recognition: a tutorial. Proc IEEE 85(9):1436
13. Shiota S, Villavicencio F, Yamagishi J, Ono N, Echizen I, Matsui T (2015) Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification
14. Villalba J, Lleida E (2010) In: Fala, pp 131–134
15. Villalba J, Lleida E (2011) In: Cost 2101 European conference on biometrics and Id management, pp 274–285
16. Villalba J, Lleida E (2011) Preventing replay attacks on speaker verification systems 47 (10), p 1
17. Wang ZF, Wei G, He QH, Wang ZF, Wei G (2011) Channel pattern noise based playback attack detection algorithm for speaker recognition 4, p 1708

18. Wang ZF (2011) Playback attack detection based on channel pattern noise. Huanan Ligong Daxue Xuebao/journal of South China University of Technology 39(10):7
19. Wang J, Li T, Shi YQ, Lian S, Ye J (2016) Forensics feature analysis in quaternion wavelet domain for distinguishing photographic images and computer graphics. Multimedia Tools Chapman Hall/CRC Appl 76(22):1
20. Wu Z, Evans N, Kinnunen T, Yamagishi J, Alegre F, Li H (2014) Spoofing and countermeasures for speaker verification: a survey. Speech Comm 66:130
21. Wu Z, Gao S, Cling ES, Li H (2015) In: Signal and information processing association summit and conference, pp 35–45
22. Wu Z, Li H (2016) On the study of replay and voice conversion attacks to text-dependent speaker verification. Multimedia Tools Appl 75(9):5311
23. Zhang L, Tan S, Yang J, Chen Y (2016) In: ACM Sigsac conference on computer and communications security, pp 1080–1091
24. Zhang L, Cao J, Xu M, Zheng F (2008) Prevention of impostors entering speaker recognition systems, Journal of Tsinghua University
25. Zhang Y, Qin C, Zhang W, Liu F, Luo X (2018) On the fault-tolerant performance for a class of robust image steganography, Signal Processing



Yanzhen Ren received the M.S. degree in Mechanical Design and Theory from Huazhong University of Science and Technology, Wuhan, China in 1999 and Ph.D. degree in Communication and Information System from Wuhan University, Wuhan, China in 2009. She is currently an associated professor in Computer School of Wuhan University. Her research interests are multimedia information security, multimedia forensics and multimedia recognition and retrieval.