International Journal of Security and Its Applications
Vol. 10, No. 12 (2016), pp.299-310
http://dx.doi.org/10.14257/ijsia.2016.10.12.24

Voice Playback Detection based on Long-window Scale-factors

Yanan Chen, Rangding Wang*, Diqun Yan, Chao Jin

College of Information and Engineering Ningbo University, Zhejiang Ningbo 315211, China
wangrangding@nbu.edu.cn

Abstract

With the popularity of high fidelity and portable recording device, it becomes very easy for the attacker to spoof speaker verification system by voice playback. In this paper, it found that there exists obvious difference in the scale-factors, a parameter of MP3 codec, when the original and playback voices are compressed by MP3. So, a detection algorithm for the playback voice is proposed. The experimental results show that the detection accuracy of the proposed algorithm can reach to 99.51% for 4 different types of eavesdropping devices. Meanwhile, the algorithm is integrated into the speaker verification system based on GMM-UBM. The equal error rate (EER) of the system has dropped 32% and the system's ability to resist playback attack is improved.

Keywords*: speaker verification system; playback attack; MP3 codec; scale-factors;*

1. Introduction

In the field of biometric verification technology [1], speaker recognition system with the advantages of the convenient access, the inexpensive sound pickup equipment, the off-site verification, has been gradually applied in finance, insurance, and so on. With the popularity of the high quality portable recording equipment, however, the pass phrases are easily obtained by impostor when the legitimate client accesses the speaker recognition system. If the impostor successfully attacks the system with the pre-recorded, the security of the speaker verification system and the interests of the client will be threatened. Therefore, it is necessary to integrate playback speech detection into the verification.

Currently, detection algorithms for voice playback attack have the following two categories: 1) Detection algorithm based on speech production randomness. In other words, though the same person says the same content at different times, the speech signals still have great randomness. In [2] and [3], a playback detector based on the random of speech sounds was designed to compare the verification recording with stored recording of past access attempts. Wu, in [4], adopted the same algorithm on the study of playback speech detection and extended the experimental samples. Besides, the algorithm was improved in [5] in which the spectral peaks are combined into pairs with a seed peak. This method, in [2-5], is only based on text-dependent speaker verification system and the efficiency of detection will significantly decrease When the number of request certification for client increase. 2) Detection algorithm based on the speech channel. In [6], according to the difference between the playback recording and the authentic recording channel and the fact that the rich channel information in silent segment, a method based on MFCC (Mel) of silent segment is proposed. However, the algorithm can't obtain the stable characteristics of the speech with a short period of time. Different recording and playback devices will result in various channel noise in replay speech signals, a playback attack detection method based on noise channel model was put forward in [7]. Hafiz, in [8], thought that the circuit components can cause nonlinear distortion in playback speech because the produce process of the playback speech experienced the eavesdropping devices and speaker device. They finally use the

ISSN: 1738-9976 IJSIA

bispectrum to construct distortion model, and construct the 12 Hu moment dimension as the characteristic. In [9-10], a countermeasure was implemented to prevent replay attacks using far-field recordings. The countermeasure was designed based on the fact that the noise and reverberation levels will increase in the far-field recorded signals and that playback devices have low frequency response in low frequency region. Although the second kind of detection algorithm has solved the problems related to text, but it is still only on a kind of eavesdropping devices.

As above-mentioned, the research work of playback voice detection still exists the following problems. On the one hand, in the algorithms analyzed above, only one type of the eavesdropping device was involved, and its encoding format was limited to wav format. In fact, recording devices such as smart phones and digital audio recorders are more likely to be adopted during spoofing. Meanwhile, mp3 and m4a are the most possible formats for the attacker. On the other hand, the existing algorithms analyzed above are all based on the acoustic properties and the channel characteristics of speech signals, which were not involved in the feature of the encoding parameters. Based on the above consideration, this paper puts forward playback voice detection algorithm, which is applicable to a variety of portable eavesdropping devices. The feature is extracted from the statistical data of the long-window scale-factors. The experimental results show that the algorithm for the recording from the various eavesdropping device has a high detection rate.

The remainder of this paper is organized as follows: Section II describes the mathematical model of the voice playback attack. The feature extraction and the verification integrated with playback detection are presented in Section III. The details of the dataset are given in Section IV. Section V shows the experimental results. Conclusions are made in Section VI.

2. Mathematical Model of Voice Playback Attack

Figure 1 shows the scene of the playback attack on speaker verification system. First, the attacker puts the recording device into his pocket or places it in some private location. Once the authenticated user speaks out his voice to verify his identity, the voice will be recorded by the attacker's device. Then the attacker will try to spoof the speaker verification system by the recorded voice.

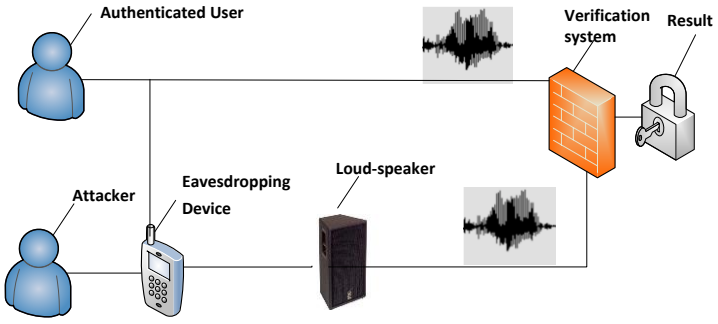

Figure 1. Diagram of Voice Playback Attack

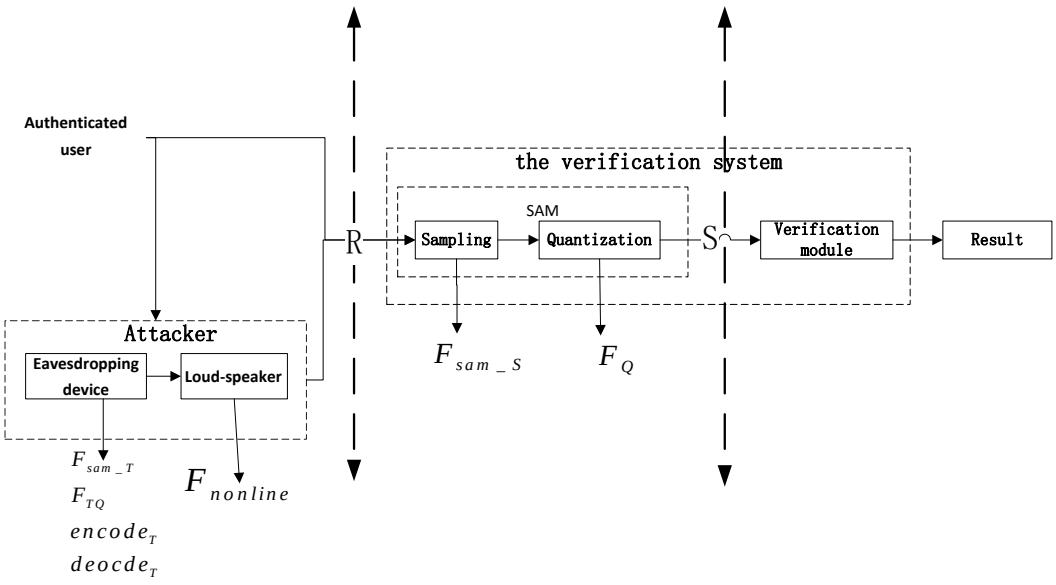

Figure 2. Mathematical Model of Voice Playback Attack

The verification system can be divided into the following three parts. The first part describes the authenticating voices from two sources, which is prior to R shown in Figure 2. Between R and S in Figure 2, the speech acquisition module (SAM) is the second part to convert the voice from analog to digital. The last part is to verify the validity of the voice and output the final result. The aim of this work is to improve the verification system's ability to resist the playback attack by adding a playback detection module before the last part.

Generally, several forms exist during the voice transmission. As for the legitimate case, the user produces the voice. Then the verification system captures the voice and converts it to digital version. However, it is more complex in the case of attacking. The voice of the legitimate user will be recorded by the attacker's device. And the recorded voice will be replayed by some device. Finally, the playback voice will be captured by the verification system. As shown in Figure 2.

According to the voice forms and the device, the function of SAM in verification system can be denoted as sampling response function F_{sam_S} and quantization response function F_Q. The function of the eavesdropping device can be denoted as sampling response function F_{sam_T}, quantization response function F_{TQ}, encoding function $encode_T$ and decoding function $deocde_T$. The primary function of the loud-speaker is power amplification which can be expressed as $F_{nonline}$. Suppose X as the detecting voice and Y is the original voice, then its model is,

$$Y = F_Q(F_{sam\_S}(X, sr), qs) \quad (1)$$

Where sr and qs denote the sampling rate and quantization step of SAM, respectively.

The playback voice Z can be expressed by Eq. (2), (3).

$$V_{re} = encode(F_{TQ}(F_{sam\_T}(X, sr_T), qs_T), cr_T) \quad (2)$$

$$Z = F_Q(F_{sam\_S}(F_{nonline}(deocde_T(V_{\text{wiretap}}, cr_T)), sr), qs) \quad (3$$

)

Where sr_T , qs_T and cr_T denote the sampling rate, quantization step and the code-rate of eavesdropping device, respectively. V_{re} is the voice recorded by the eavesdropping device. It can be seen that in the playback voice model, quantization distortion is introduced by F_{sam_T} and F_{TQ} . $encode_{RT}$, $encode_T$ and $deocde_T$ will cause codec distortion. Non-linear distortion will be caused by $F_{nonline}$.

From the above analysis, compared to the original voice, the playback voice will include various distortions, such as quantization and non-linear distortions. Therefore, the codec parameters related with quantization could be considered to extract the features for detecting the playback voice.

3. Feature Extraction and Verification

In this paper, the scale-factors are considered as the feature of playback speech detection module (PSDM) which is an added part of the verification system. Compared to the verification system without PSDM, the security of the aforementioned new system has been improved by the following way: if the request speech is classified as the playback speech by the PSDM, the verification system will be failed to access.

3.1. Feature Extraction

Through a poly phase filter bank analysis, the speech signal $x(n)$ in S (See as Figure 2) is filtered into 32 subbands with equal bandwidth,

$$x(n) = N * sub(n)$$

(4)

Where $sub(n)$ represents the signal of all subbands, and N is defined as the total subband number.

Each subband signal is further divided into 18 finer subbands by applying a modified discrete cosine transform ($MDCT$), and 576 spectral coefficients form a granule $\mathrm{gr} = \{xr_1, xr_2, \mathrm{L}\ , xr_{576}\}$, where xr denotes the spectral coefficients. The coefficients in a granule are divided into several scale-factor bands according to the window type, 21 bands for long window and 12 bands each for short windows. From the observation of experiments, long window function would be used for more than 90% frames during the MP3 coding [11]. Hence, just the scale-factors of long windows are utilized in our algorithm, and the term scale-factors is referred to as long window scale-factor in the remaining part, except where noted. According to standard ISO/IEC 13818, the scale-factors in a granule have different widths and the 576 frequency coefficients are grouped as follows:

$$sfb = \left\{ \begin{matrix} \left(xr_0 \quad xr_1 \quad \mathrm{L} \quad xr_5\right)_6 \\ \left(xr_6 \quad xr_7 \quad \mathrm{L} \quad xr_{17}\right)_6 \\ \left(xr_{17} \quad xr_{18} \quad \mathrm{L} \quad xr_{23}\right)_6 \\ \mathrm{M} \\ \left(xr_{464} \quad xr_{465} \quad \mathrm{L} \quad xr_{521}\right)_{58} \\ \left(xr_{522} \quad xr_{523} \quad \mathrm{L} \quad xr_{575}\right)_{54} \end{matrix} \right\} = \left\{ \begin{matrix} \mathrm{sfb}_0 \\ \mathrm{sfb}_1 \\ \mathrm{M} \\ \mathrm{sfb}_{20} \end{matrix} \right\} \quad (5$$

)

Where sfb denotes the scale-factor band.

Then spectral coefficients are quantized and coded within three nested iteration loops,

frame loop, outer loop, and inner loop. Frame loop, the top level loop, initializes the demands and calculates the remaining of the allocated bits. These remaining bits will be added in the bit reservoir and provide for next frame. The inner loop quantizes the input vector and increases the quantization step size until the output vector can be coded with the available amount of bit. After completion of the inner loop an outer loop checks the distortion of each scale-factor band and, if the allowed distortion is exceeded, amplifies the scale-factor band and calls the inner loop again. The scale-factors exist in the outer loop, and inner loop.

In each scale factor band, spectral coefficients in the outer loop and inner loop are handled as follows: in inner loop, $xr(i)$ are quantized by Eq. (6),

$$ix(i) = nint(\ (\frac{|xr(i)|}{\sqrt[4]{2^{stepsize}}})^{0.75} - 0.0946) \quad (6)$$

nint is the round function, which leads to an irreversible distortion $xfsf(sb)$ for input data:

$$xfsf(sb) = \sum_{i=b(sb)}^{b(sb)+bw(sb)-1} \frac{(|xr(i)| - ix(i)^{4/3} * \sqrt[4]{2^{stepsize}})^2}{bw(sb)} \quad (7)$$

In the outer loop, the quantization errors and the masking threshold SMR provided by psychoacoustic model are compared. If $xfsf(sb) < SMR$, the corresponding scale-factors will be increased by one, and the quantization step will become smaller. Then the inner loop is called again until the termination condition of outer loop can be matched. The adjusted scale-factors are obtained:

$$sf = \begin{bmatrix} sf_{1,1} & sf_{1,2} & \text{L} & sf_{1,21} \\ sf_{2,1} & sf_{2,2} & \text{L} & sf_{2,21} \\ \text{M} & \text{M} & \text{M} & \text{M} \\ sf_{N,1} & sf_{N,2} & \text{L} & sf_{N,21} \end{bmatrix} \quad (8)$$

The mean value of scale-factors, $mean\_sf$, is calculated as,

$$mean\_sf = \frac{1}{M}\left[\sum_{i=1}^{M} sf_{i,1} \quad \sum_{i=1}^{M} sf_{i,2} \quad \text{L} \quad \sum_{i=1}^{M} sf_{i,21}\right] = [f_1 \quad f_2 \quad \text{L} \quad f_{21}] \quad (9)$$

Where, $sf_{i,j}$ denotes the scale-factors of the j th scale-factor band in the i th granule; f_j represents the mean value of scale-factors in the j th scale-factor band.

In order to expose the alteration of scale-factors statistics of playback speech, we calculate the mean values of scale-factors extracted from the original speech and the replay speech recorded by 4 different eavesdropping devices. Figure 3 illustrates the histogram of the mean values. It can be seen that the statistics in the same scale-factor band have different abilities of identifying replay speech for the different eavesdropping devices. For instance, the scale-factor bands of indexes 1 and 9 hardly distinguish the replay speech of device R6620 from the original speech, while excellent classification performance can be achieved in the scale-factor band of index 14. Therefore, to guarantee the proposed algorithm is suitable for different eavesdropping devices, the statistics in all 21 scale-factor bands are taken as the detection features.

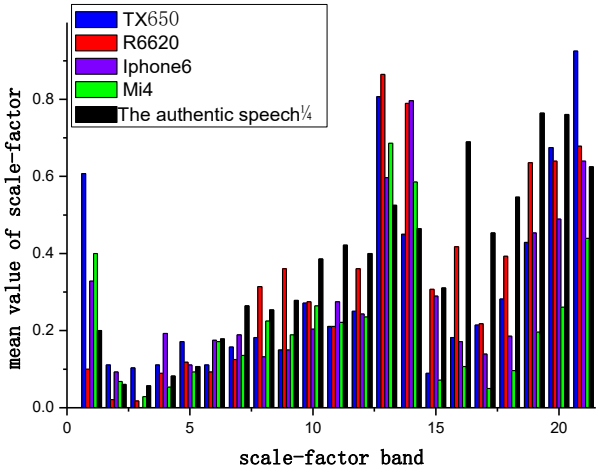

Figure 3. Histogram of Scale-factors

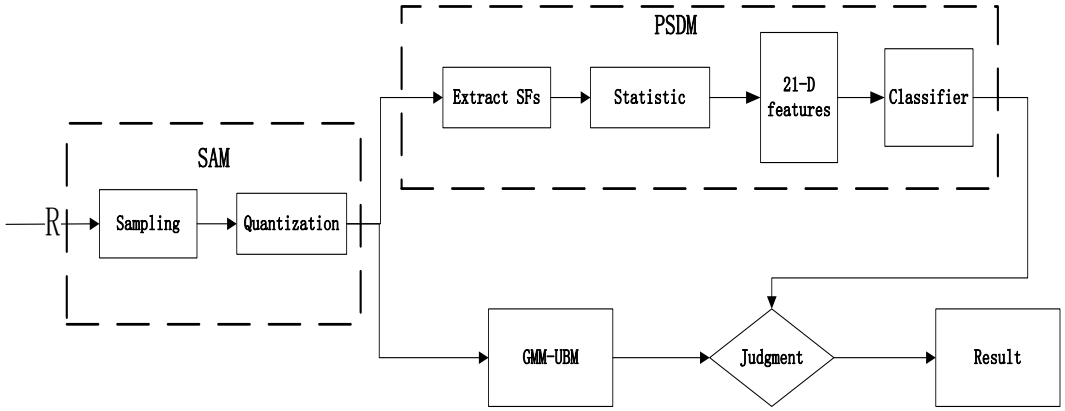

Figure 4. Block Diagram of Process Certification Process

3.2. Verification with PSDM

In the verification system, the verification module is implemented by GMM-UBM [12] algorithm and the specific acquisition equipment is used as a substitute for SAM. Figure 4 outlines the complete verification system integrated with the proposed PSDM. The verification task is executed by the following steps:

Step 1: The acquisition equipment input the acquired speech signal to the GMM-UBM verification module and PSDM simultaneously;

Step 2: In PSDM, the scale-factors of the original speech are extracted and the 21-dimensional features are constructed according to Eq. (8) and Eq. (9), respectively. These features are considered as the input of classifier, and then the classification result *result_1* will be output based on the model trained by the same classifier. If the given speech is discriminated as replay speech, *result_1=0*. Otherwise, *result_1* is set as 1.

Step 3: This step is operated along with Step (2). The given speech is matched with the model of the authenticated user, and verification module will output the verification result. Then, *result_2=1* for the authenticated speech and *result_2=0* for unauthenticated speech.

Step 4: Combining the classification result of Step (2) with the result output from step (3), the final decision will be made according to the rule described in Table 1. Only the result equals 1, can the given speech be determined as the authenticated one.

Table1. Final Judgment Rule

Result1	Result2	Result
0	0	0
0	1	0
1	0	0
1	**1**	**1**

4. Database Set

4.1 Corpus Set

The detail information of corpus used in this work is shown in Table 2. Meanwhile, considering the influence of the pronunciation habits to the recording, 20 females and 25 males from eight provinces of China have been participated in building the database. The age of participants between 20 ~ 30 accounts for 80% of the total.

Table 2. Corpus Set

	Sections	Source of the corpus	Description of detail
Speaking section	Common phrase	network language Internet password letter string digital string	Each type of sentence included 84 samples is divided into seven portions; Each participant needs to read 12 * 4 phrases.
	Long sentences	From 863 corpus database[13]	Each participants need to emotionally read 25 long sentences
Verbal section	Questions to answer	The occupation, work units, individual hobby, hometown impression, phone number, all together 20 questions	For each question, participants are free and timeless to answer the question.
	Spoken monologue	Economic issues, social issues, technology and culture development, military , all together 20 topics	Each of the participants need to select a topic, giving his opinion on this topic for more than 3 minutes

4.2 Playback Speech Database set

Since SAM in the verification system usually employs 16 KHz sampling rate when it deals with speech signal, R6620, a recording device, whose sampling rate is same with SAM, has been used in the experiment. In real scene, the attacker can easily obtain a variety of high fidelity smart phone or digital recorder. Therefore, during recording the playback speech, many kinds of the eavesdropping devices would be employed. The details of the devices adopted in this work are shown in Table 3.

Table 3. Device Information

	Acquisition equipment	Eavesdropping devices			
Type	Aigo R6620	Sony TX650	Aigo R6620	iPhone6	mi4
Recording format	wav	wav	wav	m4a	mp3
Parameter	16K,16bit	44.1K,16 bit	16K,16bit	64kbps	128kbps

In a quiet room, each participant speaks the corpus in Table 2 with standard Mandarin with their own reading habits. According to the actual situation, the acquisition and eavesdropping devices are placed at 10 and 70 cm from the participants, respectively. The recording voice sampled by acquisition equipment is defined as original voice. Then, the recording voice sampled by the eavesdropping devices is replayed by Philips DTM3155 which is a loud speaker. Then the replayed voice will be sampled by the acquisition device of the verification system. The distance between the loud speaker and the acquisition device is about 10cm. Here, the recording sampled by the acquisition equipment can be defined as playback voice, which contains four type sources of voice depended on the eavesdropping device.

The obtained recordings are cut into each of the 3 s short sound segment. Finally, 5900 original speech samples, and 5900 playback voice samples from each eavesdropping device are obtained as shown in Table 4. The detail of database is shown in Table 4.

Table 4. Details of the Database

Source	Original voice	Playback voice			
		TX650	R6220	iPhone6	Mi4
male	3600	3600	3600	3600	3600
female	2300	2300	2300	2300	2300

5. Experimental Results

In order to comprehensively evaluate the performance of the proposed algorithm, several factors, such as bit rates, device types and classifiers, are considered in the experiments. In bit rates, generally, the times of regulating the scale factor are inversely proportional to the bit rate. So for a comprehensive inspection rate influence on test results, this paper utilizes the seven bit rates, which include 64 Kbps, 80 Kbps, 96 Kbps, 128 Kbps, 160 Kbps, 192 Kbps and 256 Kbps. In terms of the playback voice source, since an attacker can acquire various eavesdropping devices, this paper uses the digital recorders (TX650, R6620), smart phones (iPhone6, Mi4) for testing. In the respect of classifier, four different types of classifiers and 10-fold cross validation, in the data mining tool weka 3.6 [14], are employed. Finally, the verification system is compared with the verification system with PSDM to demonstrate the performance improvement brought by PSDM.

5.1 Influences of the Different Classifiers

This section investigates the performance of different classifiers based on the detection. For a comprehensive analysis of the evaluation results, the paper employs four types of classifiers, Libsvm, Logistic, IBLC and ADtree in weka, in this section. All classifiers were utilized with default parameters and setting unless stated otherwise. Besides, in consideration of the aim of this paper to detect different voice source, the playback voice is from the four eavesdropping devices. The experimental samples are shown in Table 5.

Table 5. The Experimental Samples

	Original voice	Playback speech			
		TX650	R6220	iPhone6	Mi4
Train speech	1500	1000	1000	1000	1000
Test speech	1500	1000	1000	1000	1000

The experimental results are shown in Table 6-9, where, TPR is true positive rate, FPR is false positive rate, and ACC is identification accuracy percentage. The overall results demonstrate that all the four classifiers achieve good results, and that the proposed algorithm can detect playback voice from different sources. Though Libsvm, Logistic, IBk have the similar acceptable detection, the best result is produced by Libsvm classifier for all bit rates, especially it is up to 99.51% at 64 kbps. So in the PSDM, we chose the scale-factor of 64 Kbps as the final features, and use Libsvm classifier.

Table 6. Detection Accuracy in Logistic

Bit Rate	TPR (%)	FPR (%)	ACC (%)
64kbps	99.60	0.600	99.51
80kbps	96.89	96.44	96.77
96kbps	99.30	3.00	98.71
128kbps	98.30	4.200	97.59
160kbps	98.30	2.400	98.07
192kbps	98.70	1.800	98.55
256kbps	99.80	1.800	99.35

Table 7. Detection Accuracy in Libsvm

Bit Rate	TPR (%)	FPR (%)	ACC (%)
64kbps	99.60	0.600	99.51
80kbps	99.60	1.200	99.30
96kbps	99.80	1.800	99.36
128kbps	99.80	4.800	98.55
160kbps	99.80	1.800	99.35
192kbps	99.80	1.800	99.35
256kbps	98.00	3.600	97.59

Table 8. Detection Accuracy in IBk

Bit Rat	TPR (%)	FPR (%)	ACC (%)
64kbps	99.30	1.200	99.19
80kbps	99.60	0.600	99.51
96kbps	99.10	0.600	99.14
128kbps	99.10	1.200	99.03
160kbps	98.87	1.200	98.71
192kbps	98.90	1.200	98.76
256kbps	98.70	1.200	98.71

Table 9. Detection Accuracy in ADTree

Bit Rat	TPR (%)	FPR (%)	ACC (%)
64kbps	98.87	4.800	97.75
80kbps	99.10	4.800	98.07
96kbps	98.89	6.700	94.13
128kbps	98.5	4.800	97.59
160kbps	97.8	5.500	96.95
192kbps	99.30	4.200	98.39
256kbps	99.30	4.700	98.23

5.2 Influences of the Eavesdropping Devices' Coding Format

Because the proposed algorithm is based on the codec distortion in the process of producing the playback, it is very important to detect a coding format if we want to further explore the influence of the distortion. Using the chosen features and classifier shown in section 5.1, we respectively detect playback voice from a certain eavesdropping devices, and make a preliminary analysis for eavesdropping devices' coding format for playback voice. The experimental samples are shown in Table 5.

The results of the experiment are shown in Table 10. Under the limited conditions, the ACC for all the eavesdropping devices are relatively high, which confirms that the proposed algorithm works very well. Besides, all parameters used to describe the performance of the algorithm in Table 10, works well for R6620. Furthermore, the equipment, namely R6620, undertakes a comparative role because the device played an

eavesdropping device role and an acquisition equipment role. Contrasting with R6620, the detection of TX650 is lower with ACC, which shows that the sampling rate of eavesdropping device is high when the playback voice detection is low. Compared with R6620, the ACC of the smart phones are higher, which indicates the coding format of the two smart phones possess better fidelity than the acquisition equipment whose sampling rate is 16K. It is demonstrated by the experiment that the recording sampled by high fidelity smart phone or digital recorder is easily spoof verification system.

Table 10. The Detection of Different Eavesdropping Devices

Classifier	Libsvm					
	MAE	RMSE	RAE (%)	RRSE (%)	KS	ACC (%)
TX650	0.0109	0.104	2.284	21.37	0.9770	98.91
R6620	0.0069	0.0832	1.414	16.80	0.9859	99.38
iphone6	0.0073	0.0853	1.514	17.40	0.9848	99.27
mi4	0.0108	0.1041	2.261	21.26	0.9775	98.17

MEA=mean absolute error, RMSE=root mean square error, RAE=relative absolute error percentage, RRSE=root squared error percentage, KS=kappa Statistic, ACC=identification accuracy percentage

5.3 Verification System with PSDM

GMM-UBM, Gaussian Mixture Model-Universal Background Model, is the mainly method in the field of text independent speaker recognition system. Since the MFCC considers the human ear with different auditory sensitivity in different frequencies of sound waves, it has become the most widely used acoustic characteristics in speaker recognition system. In this paper, the features used in the construction process of UBM and speaker models are 24-D $MFCC$ and 24-D $\Delta MFCC$.

Based on GMM-UBM, we test whether the playback voice from four eavesdropping devices can successfully attack the system. Ten user models are used in this section, in which each user model has 150 original voice samples, while 320 playback voice samples are from four eavesdropping devices. As shown in Figure 5, the error rates of four eavesdropping devices are about 40%, which means the eavesdropping devices used in this paper successfully attack the recognition system.

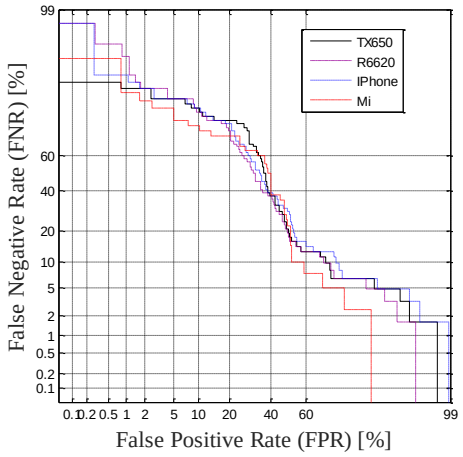

Figure 5. The Error Rates when Playback Attack the Speaker Verification

The anti-attack capacity is compared between the proposed system and the crude system. According to the experimental results of section 5.1, we finally choose the selected feature and Libsvm. The error rates of the proposed system and the crude system are 40% and 8%, respectively, as shown in Figure 6. This shows that the algorithm proposed in this paper for playback attack has excellent anti-attack capacity based on GMM-UBM speaker verification system.

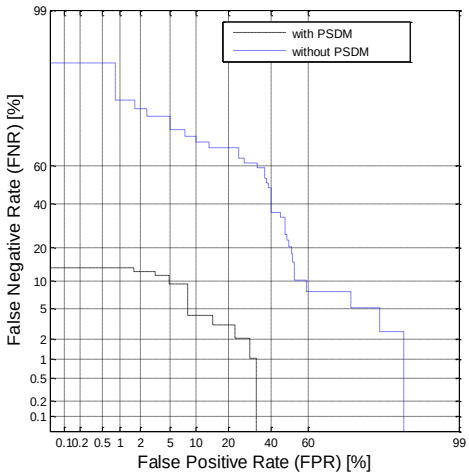

Figure 6. Comparison of Error Rates between Speaker Verification Systems with and without PSDM

6. Conclusions

A detection algorithm for playback attack has been proposed based on long-window scale-factors. The experimental results show that the original voices and playback voices which are from different playback devices can be effectively identified by the proposed algorithm. In case of Libsvm as the classifier and the scale-factors of 64Kbps as the features, the detection accuracy can reach 99.51%. Meanwhile, the EER of the speaker verification system combined with PSDM is 32% lower. In addition, various playback devices and compression formats are considered, which are more close to the real scene. The generation mechanism of the playback voice and the influences caused during playback process will be studied in the further work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 61672302, 61300055), Zhejiang Natural Science Foundation (Grant No. LZ15F020010), Ningbo University Fund (Grant No. XKXL1405, XKXL1420, XKXL1509, XKXL1503) and K.C. Wong Magna Fund in Ningbo University.

References

[1] D. Zhu, B. Ma and H. Li, Speaker verification with feature-space MAPLR parameters, IEEE Transactions on Audio Speech & Language Processing, Vol.19 No.3, **(2011)**, pp: 505-515.
[2] W. Shang, M. Stevenson, A playback attack detector for speaker verification systems, International Symposium on Communications, Control and Signal Processing (ISCCSP), St Julians, Bordeaux, **(2008)**, pp: 1144-1149, March 12-14.
[3] W. Shang, M. Stevenson. Score normalization in playback attack detection. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), dallas tx, America, **(2010)**, pp :1678-1681, March 14-19.
[4] J. Gałka, G. Marcin ,S Rafał, Playback attack detection for text-dependent speaker verification over

telephone channels, Journal of Speech Communication, Vol.67, **(2015)**, pp:143-153.
[5] Z. Wu, S. Gao, E. S. Cling, et al. A study on replay attack and anti-spoofing for text-dependent speaker verification, IEEE 2014 Summit and Conference, Asia-Pacific Signal and Information Processing Association, Siem Reap, Cambodia, **(2014)**, pp:35-45.
[6] L. P. Zhang, J. Cao, and M. X. Xu, Prevention of impostors entering speaker recognition systems, Journal of J T singhua univ (Sci&Tech), Vol.48 No.S1, **(2008)** , pp: 699 - 703.
[7] Z. F. Wang, Q. H. He, X. Y. Zhang, et al. Channel pattern noise based playback detection algorithm speaker recognition. Journal of South China University of Technology(Natural Science Edition), Vol.39 No.10, **(2011)**, pp: 7 - 12.
[8] M. Hafiz, Securing speaker verification system against replay attack, 46th International Conference: Audio Forensics , (**2012**), June 1.
[9] J. Villalba, E. Lleida, Detecting replay attacks from far-field recordings on speaker verification systems, Proceedings of the COST 2101 European conference on Biometrics and ID management, Brandenburg, Germany, **(2011)**, pp: 274-285, March 8-10.
[10] J. Villalba, E. Lleida, Preventing replay attacks on speaker verification systems, 2011 IEEE International Carnahan Conference on Security Technology (ICCST), Barcelona, Spain,Vol.47 No.10, **(2011)**, pp: 1-8, October 18-21.
[11] P. Ma, R.. D. Wang, D. Q. Yan, et al, Detecting double-compressed MP3 with the Same Bit-rate, Journal of Software, Vol.9 No.10, (**2014**), pp: 2522-2527.
[12] R. Togneri, D. Pullella, An overview of speaker identification: accuracy and robustness issues, IEEE Circuits & Systems Magazine, Vol.11 No.2, (**2011**), pp: 23-61.
[13] T. Q. Wang, A. J. Li, The design of the continuous Chinese speech recognition corpus, The sixth national conference on modern phonetics learning, TianJin China **(2003),** October 18.
[14] M. Hall, E. Frank, G. Holmes, et al. The WEKA data mining software: an update, Acm Sigkdd Explorations Newsletter, Vol.11 No.1, (**2010**), pp: 10-18.

