Channel pattern noise based playback attack detection algorithm for speaker recognition

Conference Paper · July 2011

DOI: 10.1109/ICMLC.2011.6016982 · Source: DBLP

3 authors, including:

Qianhua He
South China University of Technology
75 PUBLICATIONS **642** CITATIONS

Proceedings of the 2011 International Conference on Machine Learning and Cybernetics, Guilin, 10-13 July, 2011

CHANNEL PATTERN NOISE BASED PLAYBACK ATTACK DETECTION ALGORITHM FOR SPEAKER RECOGNITION

ZHI-FENG WANG, GANG WEI, QIAN-HUA HE

School of Electronic and Information Engineering, South China University of Technology, GuangZhou, China
E-MAIL: wang.zf01@mail.scut.edu.cn, ecgwei@scut.edu.cn, eeqhhe@scut.edu.cn

Abstract

This paper proposes a channel pattern noise based approach to guard speaker recognition system against playback attacks. For each recording under investigation, the channel pattern noise severs as a unique channel identification fingerprint. Denoising filter and statistical frames are applied to extract channel pattern noise, and 6 Legendre coefficients and 6 statistical features are extracted. SVM is used to train channel noise model to judge whether the input speech is an authentic or a playback recording. The experimental results indicate that, with the designed playback detector, the equal error rate of speaker recognition system is reduced by 30%.

Keywords:

Speaker recognition; playback attack; channel pattern noise; long-term feature; Legendre polynomial

1. Introduction

The concept of voiceprint was firstly presented by Kersta for speaker recognition in 1962 in *Nature* [1]. Since then a number of theories and methods have been proposed by scientists and researchers. Nowadays speaker recognition systems have been widely employed in commercial community. For instance, speaker recognition technique has been applied in police investigation, e-commerce, banking transactions, etc [2]. Meanwhile, the security problems keep pace with the advancement of the speaker recognition technique.

Two common security problems for speaker recognition systems are voice mimicking [3] and playback attack [4]. For voice mimicking attack, an impostor attacks the system by mimicking the voice of a registered speaker [3]. As shown in Figure 1, a playback attack is executed by an intruder who obtains a recording of the client using a high quality portable recorder as the client accesses the speaker recognition system. When attacking the system, the intruder claims the identity of the legitimate client and plays back the recordings of the client's pass phrases. Comparing with voice mimicking attack, the source recordings of playback attack are from the legitimate clients, which makes the playback attack more dangerous to speaker recognition systems. On the other hand, the ready availability of inexpensive high quality portable recorder reduces the challenge of executing such attacks and increases threats to speaker recognition systems [4].

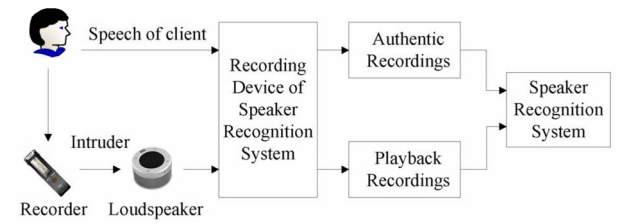

Figure 1. The playback attack

A playback attack detection algorithm has been proposed in [4] in the application of remote interaction via telephone. An utterance-specific feature called peakmap was proposed, and a similarity measurement has been assessed between the input and stored utterances. A playback attack is declared if the incoming utterance is deemed too similar to any of the stored utterances. In [5], mute voice are gathered from authentic recordings to train a channel model. If the channel characteristics for the test data do not match the training channel model, it is regarded as a playback attack.

This paper extracts channel pattern noise from authentic and playback recordings for playback attack detection. Authentic recordings are produced through the channel of recording device of the speaker recognition system, while the channel of playback recordings consists with three devices: the intruder recording device, the playback speaker, the recording device of the system. Different recording and playback devices will result in various channel noise

978-1-4577-0308-9/11/$26.00 ©2011 IEEE

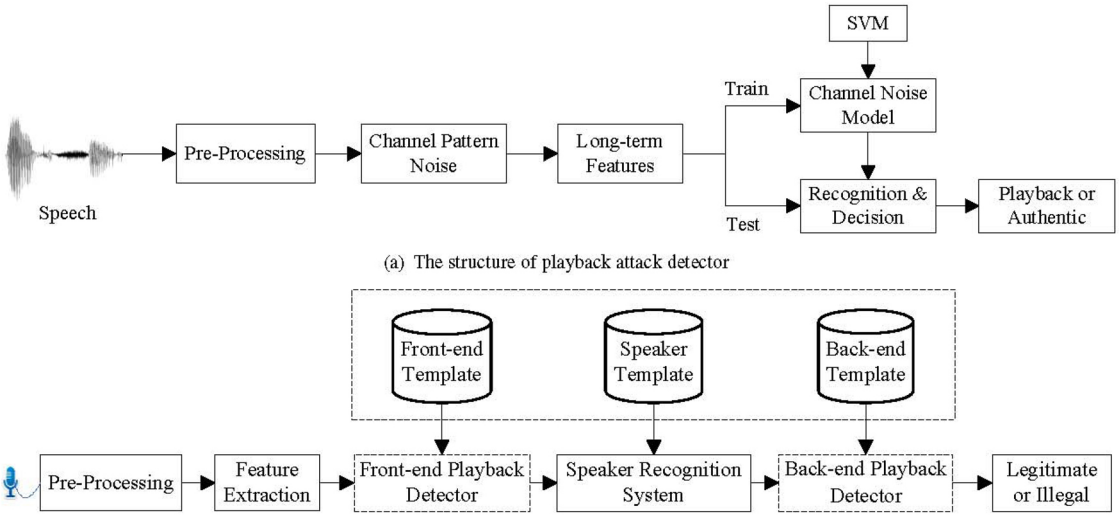

Figure 2. Playback attack detector based on channel pattern noise

in speech signals(microphone, loudspeaker, pre-amplifier, power amplifier, input and output filters, A/D, D/A, sample and hold circuit will cause channel noise [6]). We call these channel noise which are from transducers and different circuits as channel pattern noise. Authentic recordings contain the channel pattern noise from the recording device of speaker recognition system, while playback recordings include the channel pattern noise from the intruder recording device, the playback speaker and the recording device of the speaker recognition system. With extracted channel pattern noise, the playback attack can be detected.

2. Playback attack detection based on channel pattern noise

As shown in Figure 2(a) the designed playback attack detector contains four components: extraction of channel pattern noise, long-term features, channel noise model based on SVM, and decision module.

According to different positions and functions of playback attack detectors in speaker recognition systems, there are two types of playback attack detectors: front-end playback attack detector and back-end playback attack detector. As shown in Figure 2(b), for the front-end playback attack detector, before the input speech of the client enters the speaker recognition system, it should be firstly sent for playback attack detection to judge whether it is a playback speech. If the input is a playback attack speech, then the system will reject to serve the intruder directly. In this situation, the input for playback attack detector can be the speech of any client of the system, so the playback attack detector should have the capacity to detect the playback attack from any speakers of the system. For back-end playback attack detector, the client's speech firstly enters the speaker recognition system. If the system shows the identity of the illegal speaker, then the speech should be sent for playback attack detection. In this situation, the speaker is fixed by the speaker verification system. There should be one back-end playback attack detector for each client of the system.

2.1. Log-spectrum on statistical frame

Channel noise is all along with speech signals and varies slowly [7]. In order to get a stable channel noise distribution, long-term features can be used to describe channel pattern noise. Taking this into account, statistical frames are proposed to extract channel pattern noise. Statistical frames are mean of the same frequency components of the short-time frames. The extracting procedures of statistical frames are as in Figure 3.

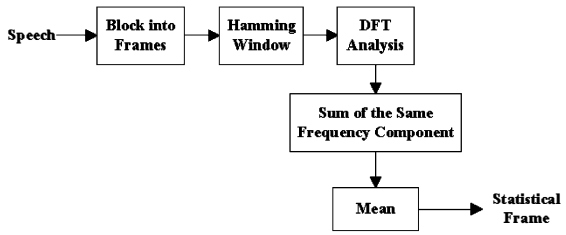

Figure 3. Block diagram of statistical frames.

The previous three steps for extracting statistical frames are block into frames, Hamming window and DFT analysis, which are similar to general analysis based on tradi-

tional short-time frames. However, the next step is to take the sum of the same frequency components, and then calculate the mean of each frequency components. The final output is the statistical frame. There are several advantages for extracting statistical frames: Firstly, the statistical frame analysis utilizes a large number of short-time frames to fit a stable channel noise distribution. The statistical frames have the statistical characteristics, and they are very helpful for extracting long-term features of channel pattern noise. Secondly, it is a normalization process for extraction of statistical frames, since it can map speech with different length into the same length frame in frequency domain. What's more, statistical frames can reduce the complexity of computation.

2.2. Channel pattern noise and features

The channel noise caused by transducers and different circuits of the recording and playback device is convolution signal in time domain [8], so channel pattern noise can be extracted in the log-spectrum domain, which transfer nonlinear model into linear model. The low frequency component is mainly the effect of convolution noise introduced by channel noise [9]. As shown in Figure 4, in the log-spectrum domain, authentic recording differs with playback recording in low frequency band, and this difference is caused by channel pattern noise of authentic an playback recordings. This fact confirms that channel pattern noise distributes in the low frequency band of speech signals.

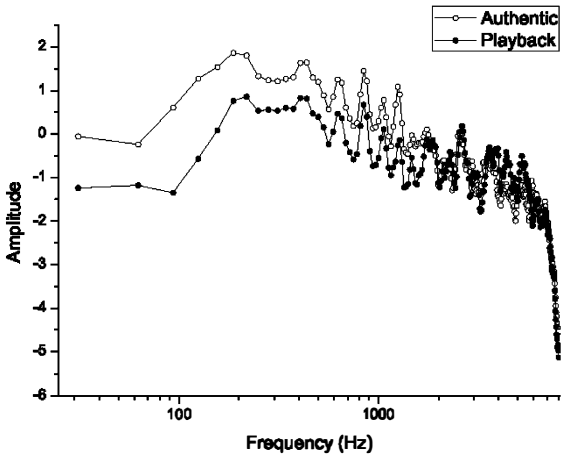

Figure 4. The authentic and playback recordings are from the same speaker with the same text "5940247874".

A denoising filter is used to extract channel pattern noise in the log-spectrum domain. As explained in Equation (1) and (2), the input recordings ($y_a(n)$ and $y_p(n)$ are authentic and playback recording) firstly pass through a denoising filter '$Defilter$', which is a high pass filter and can denoise the pattern noise from the speech. Then the logarithmic power spectrum based on statistical frame is calculated, and '$\mathscr{F}$' stands for DFT. Finally, the channel pattern noise is obtained by subtracting non-filtered speech with filtered speech, while N_a and N_p are channel pattern noise of authentic and playback recordings.

$$N_a = log[\mathscr{F}(y_a(n))] - log[\mathscr{F}(Defilter(y_a(n)))] \quad (1)$$

$$N_p = log[\mathscr{F}(y_p(n))] - log[\mathscr{F}(Defilter(y_p(n)))]. \quad (2)$$

There are two reasons for using high pass filter as denoising filter to extract channel pattern noise. Firstly, the channel noise is varying slowly compared to the variations naturally occurring in speech, and the channel distribution can be estimated to build a filter to denoise the channel noise. Secondly, the bandwidth of channel pattern noise is very narrow, and it would be very difficult to build a low pass filter with very narrow pass-band and extremely wide cutoff band. We have experimented with several denoising filters and eventually decided to use the denoising filter as follows.

$$H(z) = 1 - \frac{\sum_{n=1}^{N} \alpha^n z^{-n}}{\sum_{n=1}^{N} \alpha^n} \quad (3)$$

When extracting the channel pattern noise, $N = 32$ and $\alpha = 0.94$. Long-term features based on pattern noise features are extracted according to the following steps as in Figure 5.

2.3. Legendre Polynomials Coefficients

Legendre polynomials are a set of orthogonal basis, and this technique of approximation by Legendre polynomials has been successfully applied in language identification and other engineering applications [10]. On the other hand, the orders of Legendre polynomials for fitting curves can be in a small number in practical application. Actually, 5-order coefficient is enough for fitting the curves of channel pattern noise. A function can be expressed by Legendre series in the following form

$$f(x) = \sum_{n=0}^{\infty} L_n P_n(x) \quad (4)$$

And the form of Legendre polynomial is as follows

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n}(x^2 - 1)^n, x \in [-1, 1]. \quad (5)$$

Each coefficient of the Legendre polynomial models a particular aspect of the channel pattern noise. For example, L_0 is interpreted as the direct current component of the curve, L_1 gives information about the slope of the curve, L_2 shows information about the curvature of the curve, and L_3, L_4, L_5 model the fine details of the curve.

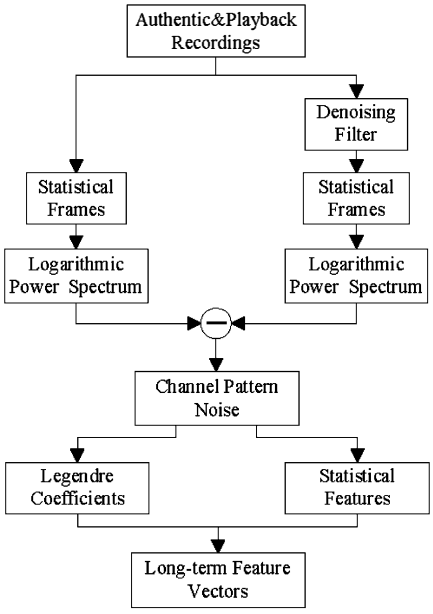

Figure 5. The extracting steps of channel pattern noise and two sets of features on statistical frames for playback attack detection.

2.4. Statistical features

Since channel pattern noise is all along with the speech signal and varying slowly, the statistical features are used to describe channel pattern noise. Taking this into account, we used a set of statistical features to represent channel pattern noise features. This set of statistical long-term features consists of six statistical features:

- PN_min: The minimal value of pattern noise.
- PN_max: The maximal value of pattern noise.
- PN_mean: The average value of pattern noise.
- PN_median: The value of the 50th percentile. It is less sensitive to outliers.
- PN_diff: The difference between PN_max and PN_min as a measure of local range of pattern noise.
- PN_stdev: The standard deviation as a measure of the variance of pattern noise.

2.5. Channel noise model based on SVM

SVM was used to build playback attack detector. The previous combined 12-demension long-term features on channel pattern noise were used. Radial basis function was selected as kernel function ($K(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2), \gamma > 0$). SOM algorithm and grid searching algorithm are used to find the best error penalty parameter $C(C > 0)$ and γ, and then the best parameters were used to train the optimal channel noise model.

3. Authentic and Playback Speech Database (APSD)

Over a span of six months, 4 females and 9 males contribute to the database by Mandarin Chinese. Among them, 6 participants are from north China, and other 7 participants are from south China. They are all graduate students, and their ages are from 20 to 35 and can speak Mandarin Chinese fluently. With considering the influence of different speaking rate, and different types of texts, the corpus is designed as follows:

- 20 phrases as commands for access control system.
- 10 digit strings consisting of ten digits each.
- 60 phonologically and syllabically balanced utterances of 8-16 word length. 30 utterances are from 863 Continuous Speech Database. 15 utterances are from the China Daily, and other 15 utterances are from Xinhua Wang which is an important official website in China.
- Two phonologically and syllabically balanced texts, of 223 words, read at an normal speaking rate.
- Asking the speaker to read the previous fixed text at a slow and fast speaking rate respectively.
- There are five topics, and the speakers can choose one and give a more than 2 minutes of spontaneous speech.

There are 1473 words in the corpus, which include 528 different words. There are 251 syllables in this corpus. What's more, it contains all the 60 phonemes in common spoken Chinese.

The authentic speech data was recorded at the system end using a sound card(Creative 5.1) with a unique microphone. The authentic recording was in 16 bits per sample and the sampling rate was 16 kHz. Meanwhile the speaker's utterance was simultaneously recorded using a high quality digital voice recorder(DVR), set at a sampling rate of 22.05 kHz in 16 bits. Then these set of speeches recorded by high quality DVR are regarded as intruder's recordings. When producing playback recordings, the intruder's recordings are played back by a loudspeaker(ALTEC iMT237) to simulate the whole physical process of playback attack.

4. Experiments and Discussion

In the following experiments, the data sets selected from APSD contain 12220 speech samples, which include 13 speakers' speech, and there are 940 samples for each speaker(470 authentic recordings and 470 playback recordings for each speaker). The distribution of the selected data sets are shown in TABLE 1. The whole duration range of

the selected data is nearly 10 hours and the storage space is about 2.8 GB. False rejection rate(FRR) and false acceptance rate(FAR) are used as evaluation indexes to estimate the performance of playback attack detection algorithm. Meanwhile, 10 fold cross-validation are used in the following experiments.

TABLE 1. DISTRIBUTION OF SELECTED DATA SETS

Data sets	Authentic	Playback
Phrases	520	520
Digit Strings	260	260
Sentences(normal speed)	2750	2750
Sentences(fast speed)	1240	1240
Sentences(slow speed)	1240	1240
Paragraphs	100	100

Three experiments are designed based on selected data sets to evaluate the proposed algorithm:

- An auditory perception experiment is designed to check whether playback recordings can be distinguished from authentic recordings by auditory perception. Speaker recognition experiments are performed on the selected data sets by GMM and HMM speaker identification systems respectively, which can reveal the threat of playback attack to speaker recognition systems.
- The front-end playback attack detectors are constructed. The performances of the speaker verification system which is in conjunction with the front-end playback detector are estimated. These experiments are used to illustrate the importance of playback detector in speaker recognition systems.
- The back-end playback attack detectors are built up, and their experimental performances are compared with the front-end playback detectors.

4.1. Auditory perception and speaker identification experiments

We picked ten participants and asked them to listen to the authentic and playback recordings. All of the ten participants can hardly distinguish these two types of recordings. It means that auditory perception cannot be used to draw a clear distinction between authentic and playback recordings. In the following, playback and authentic recordings are sent for speaker recognition by GMM and HMM speaker identification systems. The identification results are shown in TABLE 2.

For these two systems, playback recordings still keep high identification accuracies similar to the authentic recordings, which shows that these two common speaker identification systems have no ability to prevent playback attacks.

TABLE 2. SPEAKER IDENTIFICATION EXPERIMENTS(%)

Accuracy	Authentic	Playback
HMM	87.2781	81.0897
GMM	97.7318	93.6966

4.2. Front-end playback attack detector and conjunction with speaker verification system

All the speech samples of the 13 speakers are used to build a front-end playback detector, and the experimental results are shown in TABLE 3. The algorithm in [5] is also implemented on the selected data sets from APSD to build another front-end playback detector. Comparing with the algorithm in [5], FRR and FAR of the algorithm in this paper are reduced by almost 13%.

TABLE 3. COMPARING WITH OTHER ALGORITHM(%)

Error rate	Algorithm in this paper	Algorithm in [5]
FRR	2.8619	15.6732
FAR	2.4507	15.6732

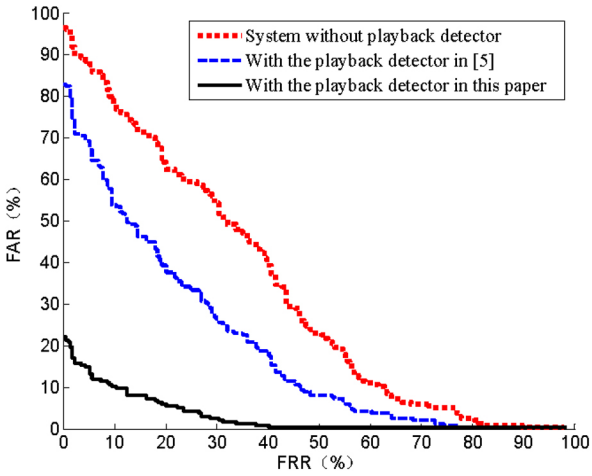

Figure 6. DET curves of speaker verification systems.

The front-end playback attack detectors both in this paper and in [5] are in conjunction with a GMM-UBM speaker verification system. In this situation, the playback recordings from speaker A are not regarded as the speech samples of speaker A, even though they would obtain high scores in this speaker verification system. As shown in Figure 6, when there are some playback recordings in the input recordings, the error rates of the system without playback attack detector are very high. The equal error rate(EER) of

this system is 40.1709%, and the speaker verification system is in poor security. With the playback attack detector in this paper, the EER of the system is 10.2564%, which is reduced by nearly 30%. With the playback attack detector in [5], the EER is 29.0598%.

4.3. Back-end playback attack detector

As show in Figure 7, the back-end playback detectors are constructed for each speaker. F01 F04 are 4 female speakers, and M01 M09 are 9 male speakers. FRR and FAR of F03, M03 and M04 are 0%. The experimental results indicate that the back-end playback detectors obtain better performance than the front-end detectors in this paper.

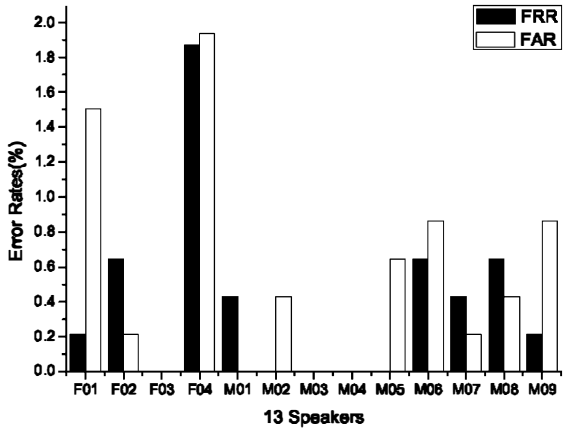

Figure 7. Back-end playback detectors.

5. Conclusion

In this paper, we have presented a channel pattern noise based approach to guard speaker recognition system against playback attacks. The experiment results show that the FRR is 2.8619% and the FAR is 2.4507%. With the playback detector, the equal error rate of speaker verification system is reduced by 30%.

The future work include doing more research on the properties of channel pattern noise and expanding other portable recording equipment just like PDA and mobile phone, which would be with high quality in recording.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Item No. 60972132) and the Natural Science Team Foundation of Guangdong Province, China (Item No. 9351064101000003).

References

[1] L. G. Kersta, "Voiceprint identification", Nature, Vol 196, No. 4861, pp. 1253-1257, 1962.

[2] E. E. Vale, and A. Alcaim, "Adaptive weighting of subband-classifier responses for robust text-independent speaker recognition", Electronics Letters, Vol 44, No. 21, pp. 1280-1282, 2008.

[3] L. Yee Wah, "Vulnerability of speaker verification to voice mimicking", Proceeding of ISIMP2004 Conference, Hong Kong, pp. 145-148, June 2004.

[4] Shang Wei, and S. Maryhelen, "Score normalization in playback attack detection", Proceeding of ICASSP2010 Conference, Dallas, pp. 1678-1681, 2010.

[5] Zhang Lipeng, Cao Jiang, and Xu Mingxing, "Prevention of impostors entering speaker recognition systems", Journal of Tsinghua University, Vol 48, No. S1, pp. 699-703, 2008.

[6] K. C. Pohlmann, Principles of Digital Audio, Sixth Edition, Mc Graw-Hill, New York, 2010.

[7] B. A. Hanson, and T. H. Applebaum, "Subband or cepstral domain filtering for recognition of Lombard and channel-distorted speech", Proceeding of ICASSP1993 Conference, Minneapolis, pp. 79-82, November 1993.

[8] Lawrence Rabiner, and Ronald Schafer,Theory and Applications of Digital Speech Processing, Prentice Hall, Prentice, 2010.

[9] H. Hermansky, and N. Morgan, "RASTA processing of speech", IEEE Transactions on Speech and Audio Processing, Vol 2, No. 4, pp. 578-589, 1994.

[10] Lin Chi-Yueh, and Wang Hsiao-Chuan, "Language Identification Using Pitch Contour Information", Proceeding of ICASSP2005 Conference, Philadelphia, pp. 601-604, 2005.