



A Replay Speech Detection Algorithm Based on Sub-band Analysis

Lang Lin, Rangding Wang^(✉), and Yan Diquan

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

Abstract. With the development of speech technology, various spoofed speech has brought a serious challenge to the automatic speaker verification system. The object of this paper is replay attack detection which is the most accessible and can be highly effective. This paper investigates discrimination between the replay speech and genuine speech in each sub-band. For sub-bands with discrimination information, we propose a new filter design approach. Finally, experiments are conducted on the ASV spoof 2017 data set using the algorithm proposed in this paper which demonstrates a 60% relative improvement in term of equal error rate compared with the baseline of ASV spoof 2017.

Keywords: Replay attack · Automatic speaker verification
Spoofing and anti-spoofing · GMM model

1 Introduction

Automatic speaker verification (ASV) is a biometric authentication technique that is intended to recognize people by analysing their speech. With the rapid development of this authentication technique, ASV technique has been extensively used in the fields of life, judicial, and the financial. Compared to other biometric authentication techniques, such as fingerprints, irises, and faces, Voiceprint authentication does not require users to perform face to face contact. Therefore speech is more susceptible to spoofing attacks than other biometric signals [1, 2]. Secondly, high-quality audio capture devices and powerful audio editing software are more conducive to spoof voice to attack ASV systems.

Spoofing attacks can be categorized as impersonation, replay, speech conversion and speech synthesis [3]. For impersonation attacks, existing ASV techniques have been able to effectively resist this spoofing attacks. Speech conversion and speech synthesis requires the counterfeiters has more specialized technical. In addition, this spoof attacks can be effectively defended by existing solutions [4, 5]. However, replay attacks are the most accessible and can be highly effective. More importantly, popularity and portability of high-fidelity audio equipment in recent years have greatly increased the threat of replaying speech to ASV systems.

In the past two years, replay attacks have received extensive attention from researchers. The ASV spoof 2017 Challenge uses the Constant-Q Cepstral Coefficients (CQCC) to detect spoofing attack and its equal error rate (EER) is 24.55% [6]. In this

database, the multi-feature fusion methods and the integrated classifier methods are used for replay attack detection [7] and its EER is 10.8%. The fusion of the two features of RFCC and LFCC reduced the EER to 10.52% [8]. In addition, the I-MFCC feature has also been shown to be effective in detecting replay speech [9]. At the same time, high-frequency information features obtained by CQT transformation has also proven to be effective [10]. Recently, Delgado et al. used the Cepstral Mean and Variance Normalization (CMVN) method on CQCC features [11]. The results show that this method is very effective for detecting replay attacks. Although the above work is significantly improved compared to the baseline, the computational complexity is relatively high due to the introduction of the CQT transformation.

Recent work focused on how to find effective features rather than analysing the differences between replay and genuine voice in each sub-band. Further, according to the differences reflected in different sub-bands, feature extraction approaches are discussed in this Work.

2 Database

The ASV spoof 2017 corpus is used in our investigations. The corpus is partitioned into three subsets: training, development, and evaluation. A summary of their composition is presented in Table 1. This paper uses Train and Development to train the model and Evaluation to test the performance of the model.

Table 1. Statistics of the ASV spoof 2017 corpus.

	#Speaker	#Replay session	#Replay configuration	#Replay speech	#Genuine speech
Train	10	6	3	1508	1508
Development	8	10	10	760	950
Evaluation	24	161	110	1298	12008

3 Sub-band Analysis

First, the speech signal is transformed from the time domain to the frequency domain by time-frequency transformation method. Then the entire frequency band is divided into 16 sub-bands and 8 sub-bands. During the experiment, one sub-band is removed at a time, and the remaining sub-bands are used to extract the sub-band features and used the GMM model for training; the equal error rate (EER) is used as the metrics of feature performance. Finally, a classification level measure of discriminative ability is estimated using EER ratio of a sub-band based spoofing detection system.

3.1 Sub-band Division and Analysis

The sub-bands feature extraction process is shown in Fig. 1. For each frame of speech, frequency bins are subdivided into sub-bands based on DFT bin groupings. The

number of the DFT bins is 256, and the window function is the Hanning. During the experiment, one sub-band is removed at a time. Within remaining sub-bands, DCT is applied to the corresponding log magnitude to obtain the remaining sub-band features. The features include 150 dimensions, comprising of 50 DCT coefficients along with the deltas and delta-deltas. Cepstral mean and variance normalization (CMVN) [12] is an efficient normalization technique used to remove nuisance channel effects. Therefore, the CMVN technique is applicable to sub-band feature.

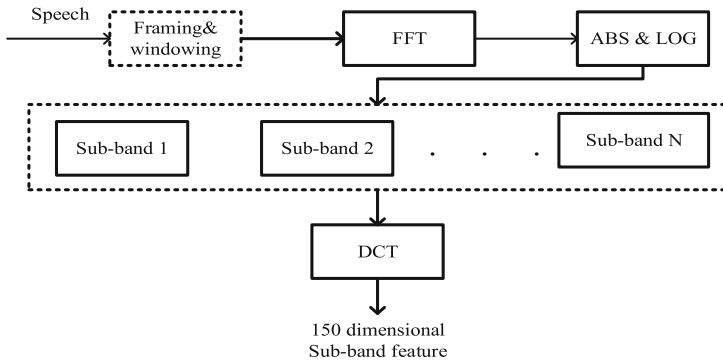


Fig. 1. Sub-band feature extraction

The EER represents the equal error rate of all sub-bands, EER_i represents the equal error rate of the remaining sub-bands after removing the i -th sub-band, and r_i represents the ratio of EER_i and EER which represents the contribution capacity of the i -th sub-band. The ratio is defined as follows:

$$r_i = EER_i / EER \tag{1}$$

The first approach involved dividing the speech bandwidth into uniform 1 kHz wide sub-bands. And the second approach involved dividing the speech bandwidth into uniform 0.5 kHz wide sub-bands. The two approaches are referred to as 8-band and 16-band divisions in the rest of the paper.

3.2 GMM Models and Performance Indicators

In Sect. 3.1, we removed each sub-band feature at a time. Within the remaining sub-bands, a 256-component GMM system is used to determine the discriminative ability within a removed sub-band. The process of GMM model training and identification is shown in Fig. 2. The primary metric is the EER [13].

3.3 Sub-band Division and Analysis

Table 2 shows the EER_i and r_i for the 8 sub-bands. The experimental results demonstrate that the r_i of the 1st and 8th sub-bands are obviously greater than 1.

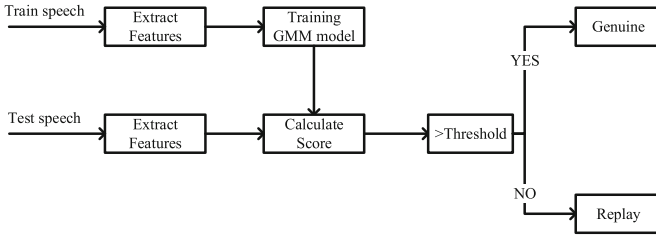


Fig. 2. GMM training process

Table 2. The experimental result of 8-bands

Sub-band	EER_i	r_i	Sub-band	EER_i	r_i
1 (0–1 kHz)	17.22	1.468	5 (4–5 kHz)	11.68	0.997
2 (1–2 kHz)	12.00	1.025	6 (5–6 kHz)	11.20	0.956
3 (2–3 kHz)	11.59	0.988	7 (6–7 kHz)	12.12	1.035
4 (3–4 kHz)	11.54	0.985	8 (7–8 kHz)	21.90	1.870
All sub-bands	11.71	–			

Table 3. The experimental result of 16-bands

Sub-band	EER_i	r_i	Sub-band	EER_i	r_i
1 (0–0.5 kHz)	15.81	1.350	9 (4–4.5 kHz)	11.45	0.978
2 (0.5–1 kHz)	11.86	1.011	10 (4.5–5 kHz)	11.85	1.012
3 (1–1.5 kHz)	12.21	1.041	11 (5–5.5 kHz)	11.68	0.997
4 (1.5–2 kHz)	11.75	1.003	12 (5.5–6 kHz)	11.51	0.983
5 (2–2.5 kHz)	12.06	1.030	13 (6–6.5 kHz)	12.19	1.041
6 (2.5–3 kHz)	11.78	1.006	14 (6.5–7 kHz)	11.73	1.002
7 (3–3.5 kHz)	11.41	0.977	15 (7–7.5 kHz)	12.98	1.108
8 (3.5–4 kHz)	11.96	1.020	16 (7.5–8 kHz)	18.07	1.543
All sub-bands	11.71				

Specifically, the 0–1 kHz and 7–8 kHz sub-bands are identified as the most discriminative frequency regions.

Table 3 shows the EER_i and r_i for the 16-bands. The experimental results show that at low-frequencies, 0–0.5 kHz contains more discriminatory information than 0.5 Hz–1 kHz. Also in the high-frequency region, 7.5 kHz–8 kHz contains more discriminative information.

As can be seen from Tables 2 and 3, the 0–0.5 kHz and 7–8 kHz sub-bands are identified as the most discriminative frequency regions. And compared to low-frequencies, high frequencies contain more discriminative information.

4 Filter Banks Design

For the better use of the discriminative information brought by the 0–1 kHz sub-band and the 7–8 kHz sub-band, we have proposed two filter design approaches. The basic idea behind the proposed approaches is the allocation of a greater number of filters within the discriminative sub-bands [3].

Two different filter banks design approaches are presented in this paper. All two approaches involve assigning the center frequencies of triangular filters across the speech bandwidth. The initial approach is allocating more linear filters in discriminative frequency bands based on the r_i in Sect. 3. The second approach is also based on r_i , which is allocating Mel filter banks at low-frequencies bands, linear filter banks at intermediate frequency bands, and I-Mel filter banks at high-frequencies bands. The output of the filter is defined as the cepstrum coefficient which includes 46 dimensions, comprising of 15 DCT coefficients along with the deltas, delta-deltas, and log-energy. The process of feature extraction is shown in Fig. 3.

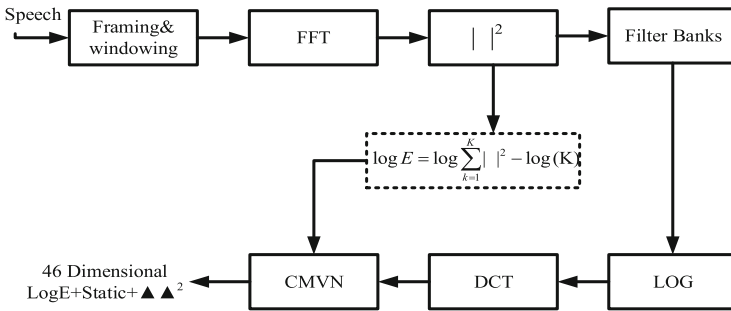


Fig. 3. Feature extraction

4.1 Linear Filter Design

This approach idea is the allocation of a greater number of filters within the discriminative sub-bands. The number of linear filters allocates in each band is related to the r_i . For example, in an 8-band experiment, the r_i at 0–1 kHz is 1.5, the r_i between 1–7 kHz is around 1.0, and the r_i between 7–8 kHz is around 1.8. Therefore the 8 -band filter design is to design 6 linear filters per 1 kHz in the 0–1 kHz frequency band. In the frequency band of 1–7 kHz, 4 linear filters are allocated per 1 kHz. In the 7–8 kHz frequency band, 7 linear filters are allocated per 1 kHz. The shape of the filter banks is shown in Fig. 4.

According to the 8 sub-band design idea, the 16-band filter bank is designed to allocate 3 linear filters in 0–0.5 kHz, 26 linear filters in 0.5–7 kHz, and 7 linear filters in 7–8 kHz. The shape of the filter banks is shown in Fig. 5.

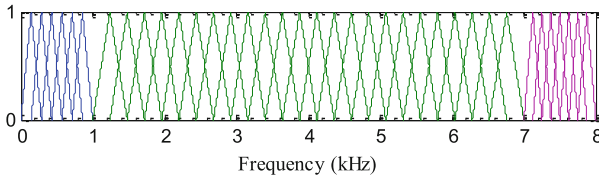


Fig. 4. 8 sub-band linear filter design

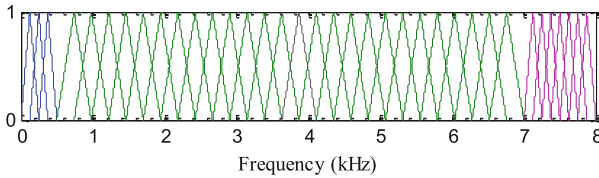


Fig. 5. 16 sub-band linear filter design

4.2 Mel, Linear, and I-Mel Filter Design

This approach idea is not only to allocate a greater number of filters within the discriminative sub-bands but also assign more appropriate filter types to the corresponding sub-bands. At low frequencies, we use the Mel filter design to enhance the details of the low frequencies. At high frequencies, we use I-Mel filters (inverting the Mel scale from high frequency to low frequency) to enhance the detail of the high frequencies, while the Intermediate frequency uses linear filters. According to the above theory, the 8-band filter is designed to allocate 6 Mel filters per 1 kHz in the frequency band of 0–1 kHz. In the frequency band of 1–7 kHz, 4 linear filters are allocated per 1 kHz. In the 7–8 Hz frequency band, 7 I-Mel filters are allocated per 1 Hz. The shape of the filter design is shown in Fig. 6.

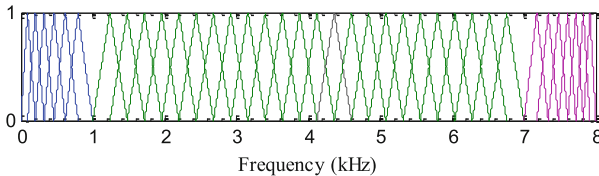


Fig. 6. 8 sub-band combination filter design

According to the 8-band design idea, the 16-band filter bank is designed to allocate 3 Mel filters in 0–0.5 kHz frequency band and 26 linear filters in 0.5–7 kHz frequency band. And in 7–8 kHz frequency band, 7 I-Mel filters are used in the sub-band. The design of the filter is shown in Fig. 7.

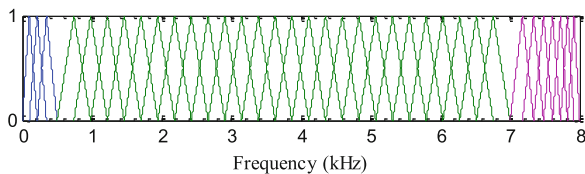


Fig. 7. 16 sub-band Mel, Linear, and I-Mel filter design

5 Results and Discussion

This paper proposes a new filter design method by calculating the EER ratio for each sub-band to determine the number and shape of filters for each sub-band. In order to verify the validity of the filter bank designed in this paper, we compare the cepstrum coefficient proposed by the filter bank proposed in this paper with the cepstrum coefficient proposed by the traditional filter. The cepstrum coefficient proposed by the traditional filter is defined as LFCC. MFCC, I-MFCC, extraction process as showed in Fig. 3. The cepstrum coefficient includes 46 dimensions, comprising of 15 DCT coefficients along with the deltas, delta-deltas, and log-energy. In addition, we compare the algorithm proposed in this paper with the algorithm proposed by other researchers. Experimental results show that our algorithm is superior to other literature to varying degrees (Table 4).

Table 4. Experimental results

	Detailed description	EER (%)
Basic features	36 Linear filter bank (LFCC)	13.07
	36 Mel filter bank (MFCC)	19.50
	36 I-Mel filter bank (I-MFCC)	13.92
This paper features	6 Linear filter in 0–1 kHz + 24 Linear filter in 1–7 kHz + 7 Linear filter in 7–8 kHz	10.66
	3 Linear filter in 0–0.5 kHz + 26 Linear filter in 0.5–7 kHz + 7 Linear filter in 7–8 kHz	10.16
	6 Mel filter in 0–1 kHz + 24 Linear filter in 1–7 kHz + 7 I-Mel filter in 7–8 kHz	10.81
	3 Mel filter in 0–0.5 kHz + 26 Linear filter in 0.5–7 kHz + 7 I-Mel filter in 7–8 kHz	9.88
Other literature	[6]	24.55
	[11]	12.24
	[7]	10.8
	[9]	18.37
	[10]	17.31
	[8]	10.25

6 Conclusions

In this paper, we have used EER ratio to identify sub-bands that contain discriminative information between genuine and replay speech. Two such discriminatory sub-bands were identified: 0–0.5 kHz and 7–8 kHz. We have then proposed two approaches to designing banks of triangular filters that allocate a greater number of filters to the more discriminative sub-bands. The two approaches were experimentally validated on the ASV spoof 2017 corpus and outperform other approaches proposed by other researchers. Considering that the number of filters in the filter bank is a key parameter that may have a significant effect on system performance. Therefore, future work will pay more attention to the choice of each sub-band filter.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (Grant No. U1736215, 61672302), Zhejiang Natural Science Foundation (Grant No. LZ15F020002, LY17F020010), Ningbo Natural Science Foundation (Grant No. 2017A610123), Ningbo University Fund (Grant No. XKXL1509, XKXL1503). Mobile Network Application Technology Key Laboratory of Zhejiang Province (Grant No. F2018001).

References

1. Wu, Z., Kinnunen, T., Chng, E.S., Li, H., et al.: A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case. In: Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific, pp. 1–5 (2012)
2. Kinnunen, T., Wu, Z., Lee, K.A., et al.: Vulnerability of speaker verification systems against speech conversion spoofing attacks: the case of telephone speech. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4401–4404 (2012)
3. Sriskandaraja, K., Sethu, V., Le, P.N., et al.: Investigation of sub-band discriminative information between spoofed and genuine speech. In: INTERSPEECH, San Francisco, pp. 1710–1714 (2016)
4. Hanilçi, C., Kinnunen, T., Sahidullah, M., et al.: Spoofing detection goes noisy: An analysis of synthetic speech detection in the presence of additive noise. *Speech Commun.* **85**, 83–97 (2016)
5. Pal, M., Paul, D., Saha, G.: Synthetic speech detection using fundamental frequency variation and spectral features. *Comput. Speech Lang.* **48**, 31–50 (2017)
6. Todisco, M., Delgado, H., Evans, N.: A new feature for automatic speaker verification anti-spoofing: constant Q cepstral coefficients. In: Odyssey 2016-The Speaker and Language Recognition Workshop, Piscataway, NJ, pp. 283–290. IEEE (2016)
7. Ji, Z., Li, Z.Y., Li, P., et al.: Ensemble learning for countermeasure of audio replay spoofing attack in ASV spoof 2017. In: INTERSPEECH 2017, Stockholm, pp. 87–91 (2017)
8. Font, R., Espin, J.M., Cano, M.J.: Experimental analysis of features for replay attack detection — results on the ASV spoof 2017 challenge. In: INTERSPEECH, Stockholm, pp. 7–11 (2017)
9. Lantian, L., Yixiang, C., Dong, W.: A study on replay attack and anti-spoofing for automatic speaker verification. In: INTERSPEECH 2017, Stockholm, pp. 92–96 (2017)
10. Witkowski, M., Kacprzak, S., Żelasko, P., et al.: Audio replay attack detection using high-frequency features. In: INTERSPEECH 2017, Stockholm, pp. 27–31 (2017)

11. Delgado, H., Todisco, M., Sahidullah, M.: ASV spoof 2017 Version 2.0: meta-data analysis and baseline enhancements. In: *Odyssey 2018 - The Speaker and Language Recognition Workshop*, Les Sables d'Olonne, France, pp. 1–9 (2018)
12. Auckenthaler, R., Carey, M., Lloyd-Thomas, H.: Score normalization for text-independent speaker verification systems. *Digit. Sig. Process.* **10**, 42–54 (2000)
13. Kinnunen, T., Sahidullah, M., Delgado, H., et al.: The ASV spoof 2017 challenge: assessing the limits of replay spoofing attack detection. In: *INTERSPEECH 2017*, Stockholm, pp. 1–6 (2017)
14. Wu, Z., Yamagishi, J., Kinnunen, T., et al.: ASV spoof: the automatic speaker verification spoofing and countermeasures challenge. *IEEE J. Sel. Top. Sign. Proces.* **11**, 588–604 (2017)
15. Wu, Z., Kinnunen, T., Evans, N., et al.: ASV spoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In: *16th Annual Conference of the International Speech Communication Association, INTERSPEECH 2015*, Dresden, vol. 11, pp. 588–604 (2015)