Regular Paper

Replay Attack Detection Based on Spatial and Spectral Features of Stereo Signal

Ryoya Yaguchi[1,a)] Sayaka Shiota[1,b)] Nobutaka Ono[1] Hitoshi Kiya[1]

Received: March 14, 2020, Accepted: December 1, 2020

Abstract: In this paper, we propose a replay attack detection (RAD) method that uses spatial and spectral features of a stereo signal. To distinguish genuine and replayed utterance, we focus on non-speech segments, in which a human does not emit sound, but a loudspeaker for replay attack might emit some recorded noise or its electromagnetic noise. The generalized cross-correlation (GCC) based spatial features capture this difference. To improve the robustness against the variety of recording environments, we combine the spatial features with spectral features. In particular, we fuse the output scores of GCC-based and spectral feature-based methods. In experiments, we confirm the effectiveness of the combination of spatial and spectral features.

Keywords: automatic speaker verification, replay attack, spoofing countermeasure, generalized cross-correlation

1. Introduction

Recently, biometric authentication systems have become popular for use in various areas such as banking protection and immigration control [1], [2], [3]. Automatic speaker verification (ASV), which uses voice as a biometric template, is one such technique. With voice templates, ASV systems can easily be linked with voice interface systems. However, it has been reported that spoofing attacks (e.g., replay and speech synthesis) have become a serious problem for ASV systems [4]. As a means of considering countermeasures for spoofing attacks, ASV Spoofing and Countermeasures (ASVspoof) challenges were held in 2015 [5], 2017 [6], and 2019 [7]. Through these challenges, many countermeasures using various acoustic features have been proposed [8], [9], [10].

The ASVspoof challenges assume two types of spoofing attacks. One is a physical access (PA) attack, and the other is a logical access (LA) attack. A block diagram of the PA attack is shown in **Fig. 1**. Since the ASVspoof database was recorded by using single-channel microphones, almost all proposed countermeasures assume a single-channel situation. Meanwhile, since recording with multi-channel microphones has become easy, replay attack detection (RAD) systems assuming multi-channel recording have also been proposed [11], [12], [13]. In Ref. [13], we use generalized cross-correlation (GCC) [14] of stereo signals for RAD. GCC-based systems focus on non-speech segments, in which no sound is emitted from humans, but loudspeakers tend to generate some noise and non-perceptual signals, and these signals can be easily captured in non-speech segments. GCC-based methods have achieved high performances in some primitive experiments. However, this performance needs to improve as the methods are situation-dependent. The GCC of stereo signals is regarded as a spatial feature, and it captures different characteristics compared with spectral features. To utilize the different aspects of these two features, we fuse the output scores of the GCC-based and spectral feature-based methods [15]. Additionally, a convolutional neural network (CNN)-based RAD system that was submitted to ASVspoof 2019 was compared and discussed with systems using the proposed method. In an experiment, one of the systems achieved a relative error reduction of 72.5% compared with a single-GCC-based method and a relative error reduction of 96.6% compared with the single-spectral-based system.

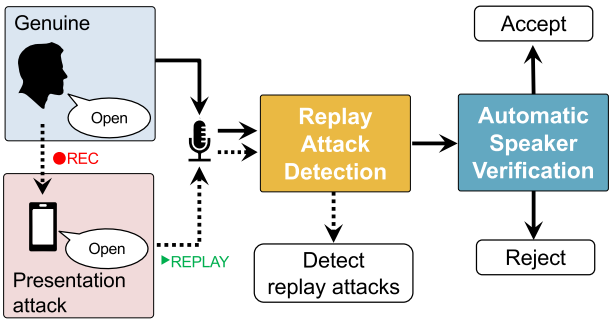

Fig. 1 Block diagram of replay attack detection and ASV systems.

The remainder of this paper is organized as follows. Related work on using cross-correlation methods is detailed in Section 2. Section 3 introduces spectral feature-based systems proposed in ASVspoof, and Section 4 provides the proposed score-fusion method that uses cross-correlation and spectral features. Section 5 describes the experimental setup and the results of detection tests. Finally, Section 6 concludes this paper.

2. GCC-based RAD Method

2.1 Characteristics of Loudspeakers in Non-speech Segments

Suppose that we record a speech by two microphones a and b.

[1] Tokyo Metropolitan University, Hino, Tokyo 191–0065, Japan
[a)] yaguyaguchiii@gmail.com
[b)] sayaka@tmu.ac.jp

For a genuine speaker, the recorded signals can be represented in the time-frequency domain:

$$M_a(t,f) = H_a(t,f)S(t,f) + N_a(t,f), \quad (1)$$

$$M_b(t,f) = H_b(t,f)S(t,f) + N_b(t,f), \quad (2)$$

where $M_a(t,f)$ and $M_b(t,f)$ are signals observed at each microphone, and $S(t,f)$ is the sound source. $H_a(t,f)$ and $H_b(t,f)$ are transfer functions from a speaker to each microphone. $N_a(t,f)$ and $N_b(t,f)$ are background noises. In non-speech segments, the source signal $S(t,f)$ is equal to 0. Thus, the signals observed in non-speech segments include only background noise:

$$M_a(t,f) = N_a(t,f), \quad (3)$$

$$M_b(t,f) = N_b(t,f). \quad (4)$$

In this case, they are not highly correlated because background noise is usually diffuse or the direction is not fixed. In comparison, the replay attack case is different. Let

$$M_p(t,f) = H_p(f)S(t,f) + N_p(t,f) \quad (5)$$

be a speech signal recorded by a microphone, p, for a replay attack. When this recorded signal is played by a loudspeaker, the signals observed by the two microphones are written as

$$M_a(t,f) = H'_a(f)(M_p(t,f) + N_s(t,f)) + N_a(t,f), \quad (6)$$

$$M_b(t,f) = H'_b(f)(M_p(t,f) + N_s(t,f)) + N_b(t,f), \quad (7)$$

where $H'_a(f)$ and $H'_b(f)$ are transfer functions, and $N_s(t,f)$ represents electromagnetic noise generated by the loudspeaker. In non-speech segments, $S(t,f) = 0$ yields $M_p(t,f) = N_p(t,f)$. Then, Eqs. (8) and (9) can be rewritten as

$$M_a(t,f) = H'_a(f)(N_p(t,f) + N_s(t,f)) + N_a(t,f), \quad (8)$$

$$M_b(t,f) = H'_b(f)(N_p(t,f) + N_s(t,f)) + N_b(t,f). \quad (9)$$

The equations mean that the recorded noise $N_p(t,f)$ and the electromagnetic noise $N_s(t,f)$ are still emitted even in non-speech segments. Then, the noise in non-speech segments can be localized, and GCC values become high. These characteristics help to distinguish spoofing attacks from genuine utterances. Let $r_1(t,f)$ and $r_2(t,f)$ be zero-mean signals captured by two microphones. Then, the GCC between them can be calculated as below.

$$\phi_g(\tau;t) = \frac{1}{L}\sum_f \frac{r_1^*(t,f)r_2(t,f)}{|r_1^*(t,f)r_2(t,f)|} e^{j2\pi f\tau/L}, \quad (10)$$

where $t = [1,\ldots,T]$ and f are the frame and the frequency index, respectively. τ is the time difference, and L is the frame length. In a genuine-speaker case, the maximum GCC is low in non-speech segments because no sound is emitted from a genuine speaker [16]. In the case of a loudspeaker, since recorded or electromagnetic noises from loudspeakers can be emitted, the maximum GCC becomes high even in non-speech segments. **Figure 2** illustrates an example of calculating GCC from a genuine utterance and spoofed one.

Figures 2 (a) and (b) show the waveforms of a genuine utterance and a replayed one and the trajectories of the maximum GCC for each frame, respectively. The red boxes in Fig. 2 (b) denote non-speech segments. According to these trajectories, the maximum GCCs were low for the genuine utterance, and those of the replayed utterance were high in the non-speech segments. Figure 2 (c) shows the GCC of one frame in both a speech segment and non-speech one for the genuine and the replayed utterances. The red dots denote the maximum GCC in each frame. In the speech segments, the peak of both utterances had a high value. In the non-speech segments, the peak of the genuine utterance was low, whereas the peak of the replayed utterance was high. From this investigation, recorded background and electromagnetic noises could be an effective factor in spoofing countermeasures.

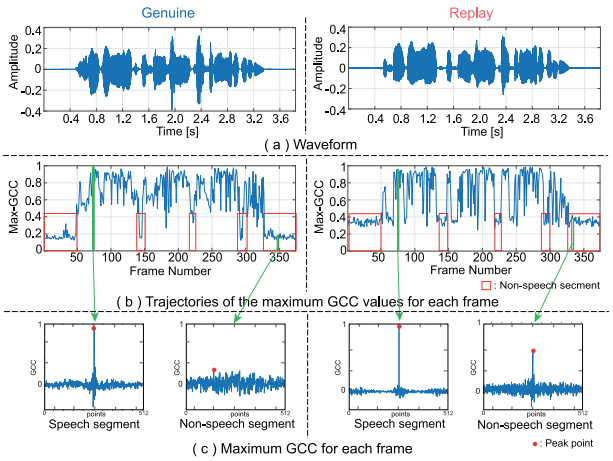

Fig. 2 GCC examples in genuine and replay attack cases.

2.2 Spoofing Detection Using Maximum GCC in Non-speech Segments

The GCC-based method [13] focuses on the trajectories of the maximum GCC (max-GCC) in non-speech segments for spoofing detection. The max-GCC for each frame is defined as

$$\phi_{\max}(t) = \max_{\tau} \phi_g(\tau;t). \quad (11)$$

As shown in Fig. 2 (b), there were two types of non-speech segments; "short pauses" appeared during a speaking period, and "silent segments" appeared both before the start of speaking and after the end. Therefore, two scores are defined for calculating the detection score with the maximum GCC. One focuses on the minimum value from among the maximum GCCs for short pauses, which is called "GCC(min)." The other focuses on the average value of the maximum GCCs for silent segments, called "GCC(avg)." These definitions are expressed as:

$$\text{GCC(min):}\ \Phi_{\min} = \min_{t_s \le t \le t_e} \phi_{max}(t), \quad (12)$$

$$\text{GCC(avg):}\ \Phi_{\text{ave}} = \frac{1}{K}\sum_{T_s \le t < t_s, t_e < t \le T_e} \phi_{\max}(t), \quad (13)$$

where t_s and t_e are the start and end points of an utterance, respectively, and K is the total number of frames in segment t. Parameters T_s and T_e represent the start and end points for calculating GCC(avg), respectively. The value of these parameters can be set arbitrarily under the constraints $1 < T_s < t_s$, $t_e < T_e < T$, where the parameter T represents the end point of an utterance. In this paper, these methods were treated as the GCC-based methods.

3. Spectral Feature-based RAD Methods

3.1 ASVspoof 2019 Results

For the ASVspoof 2019 PA scenario, 50 systems were submitted [19]. Many countermeasures used DNNs such as the CNN, light-CNN (LCNN), and residual network (ResNet) as backend systems [20], [21], [22], [23], [24], [25], [26]. For input features, spectrogram and phase information [22], [27], linear frequency cepstral coefficients (LFCC) [18], constant Q cepstral coefficients (CQCC) [17], Mel-frequency cepstral coefficients (MFCC), inverted MFCC (IMFCC) [28], and rectangular filter cepstral coefficients (RFCC) [29] were adopted. According to the results, the systems that obtained the lowest EERs used several kinds of DNNs for frontend or backend systems and adopted an ensemble of classifiers. As this paper mentioned, the ASVspoof 2019 database is composed only of single channel signals. And, almost all systems submitted to ASVspoof challenges used spectral features only.

3.2 Benchmark System for ASVspoof 2019

The ASVspoof 2019 challenge provided two benchmark systems that use a Gaussian mixture model (GMM)-based classifier. The GMM of each system is trained with spectral features, CQCC [17] and LFCC [18], respectively. In the past ASVspoof challenges, many countermeasures used CQCC and LFCC as effective spectral features. Thus, they were adopted as benchmark systems. The features are extracted from input speech signals, and a log-likelihood ratio (LLR) is calculated by using the GMMs as below.

$$\mathrm{LLR} = \log p(\boldsymbol{X}|H_0) - \log p(\boldsymbol{X}|H_1), \tag{14}$$

where $\boldsymbol{X} = x_1, x_2, \ldots, x_n$ is a speech utterance, and n is a number of frames. H_0 is a null hypothesis, and H_1 is an alternative hypothesis, and they correspond to whether $\boldsymbol{X}$ is genuine speech or spoof speech. p represents conditional probabilities whether $\boldsymbol{X}$ is H_0 or H_1. Since the LLR is calculated per frame, the average of the LLR for an utterance is referred to as "LLR." In this paper, the reverse sign of LLR is used as a detection score.

3.3 CNN-GRU for RAD Method

A lot of countermeasures using DNN have been proposed for ASVspoof 2019 [19]. One of these countermeasures used high-resolution spectrograms as input features, and CNN and gated recurrent unit (GRU) were used as a classifier, and this countermeasure was named CNN-GRU [20]. The DNN architecture of CNN-GRU is composed of convolutional layers, pooling layers, ResNet layers, and a GRU layer. For spectrograms that are used as input features, magnitude, phase spectrogram, power spectral density (PSD) are extracted. In the result of ASVspoof 2019, the EER of the CNN-GRU system obtained 2.45% that ranked 10th of all systems. The authors of CNN-GRU provided a GitHub URL about the single system that uses a high-resolution magnitude spectrogram as an input feature, and ResNet is used as a classifier. The EER of this ResNet system was 4.79% under the conditions of the ASVspoof 2019 PA scenario. In this paper, this single ResNet system was used as one of the spectral feature-based methods. In this system, the value calculated from the last node was directly used for a detection score which is referred to as "CS." The training manner is the same as what they proposed in Ref. [20].

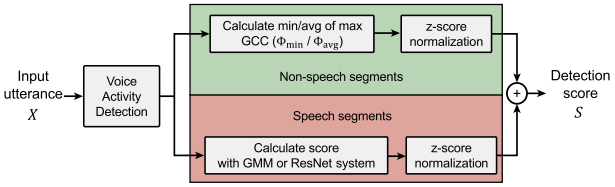

Fig. 3 Flow of the proposed method.

4. Score Fusion System

4.1 Motivation

It has been reported that GCC-based methods achieved high performances, especially in quiet situations [13]. However, this performance is situation-dependent. Thus, robustness must be improved for obtaining a stable performance. GCC-based methods focus on spatial characteristics in non-speech segments. Through the ASVspoof challenges, many approaches based on various kinds of acoustic features have been reported [8], [9], [10]. Since these spectral features are extracted from spectral characteristics, characteristics different from those of GCC-based methods can be utilized. Thus, it is expected that fusing the scores of spatial and spectral feature-based systems can enable the systems to compensate for each other, improving robustness.

4.2 Procedure

The procedure of the proposed score-fusion system is illustrated in **Fig. 3**. First, an input utterance is separated into speech and non-speech segments by voice activity detection (VAD). From all non-speech segments of the input utterance, the GCC scores Φ_{min} and Φ_{avg} are calculated by Eqs. (12), (13). By using all frames of speech segments in an input utterance, the LLR and CS were calculated under the manner of Sections 3.2 and 3.3, respectively. Each score from the non-speech or speech segment is normalized by using z-score normalization:

$$z = \frac{I - \mu}{\sigma}, \tag{15}$$

where z is the normalized score, I is the input score, and μ and σ are the mean and the standard deviation of scores in training data. Finally, the added scores are used as a detection score. We show an example of the detection score S with GCC(min) and the LLR of CQCC:

$$S = \frac{\Phi_{\mathrm{min}} - \mu_{\Phi_{\mathrm{min}}}}{\sigma_{\Phi_{\mathrm{min}}}} + \frac{\mathrm{LLR_{CQ}} - \mu_{\mathrm{LLR_{CQ}}}}{\sigma_{\mathrm{LLR_{CQ}}}}, \tag{16}$$

where $\mu_{\Phi_{\mathrm{min}}}$ and $\mu_{\mathrm{LLR_{CQ}}}$ are the averages of Φ_{min} and $\mathrm{LLR_{CQ}}$, respectively. Also, $\sigma_{\Phi_{\mathrm{min}}}$ and $\sigma_{\mathrm{LLR_{CQ}}}$ are the standard deviations.

5. Experiments

To evaluate the performance of the score-fusion system, experiments on replay attack detection were carried out.

5.1 Database

Figure 4 illustrates the testing flow of the experiments in both

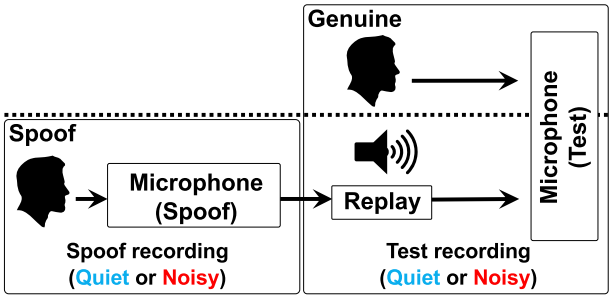

Fig. 4 Testing flow and recording process.

genuine and spoof cases. There were two types of recording: spoof and test. Although many recording situations could be considered, "Quiet" and "Noisy" were assumed for our experiments. "Quiet" means there was no extra background noise such as an air conditioner in a common space. "Noisy" represents the presence of a stationary sound, such as an air conditioner running on low, and a non-stationary sound, such as a TV program playing at a moderate volume. To construct databases of stereo signals for replay attack detection, all situations based on the above assumptions were performed. Additionally, several kinds of microphones and loudspeakers were prepared. To analyze each aspect of the situations, two databases were used.

The first database (DB1) was used for the comprehensive analysis of various situations in terms of the recording processes. For DB1, two types of microphones were used for spoof recording: AKG P170 (AKG) and TAMAGO-03 (TMG). The AKG is a condenser microphone and has strong directivity. The TMG has omnidirectional microphones with weak directivity to allow flexibility in terms of the speaker's position. For the TMG, two of the eight microphone channels were used, whereas two AKGs were installed in parallel and facing in the same direction. For replay attacks, four different types of loudspeaker were used, Elecom LBT-SPP300 (Elecom), Apple iPhone 6s (iPhone), Sony SRS-ZR7 (Sony-S), and Creative Inspire 2.0 1300 (CI). The Sony-S is 300-mm wide, 86-mm deep, and 93-mm high. It generates a non-perceptual electromagnetic noise in silent segments of replayed attacks. The CI is comprised of two separate stereo loudspeakers. Each speaker is 99-mm wide, 131-mm deep, and 221-mm high. The Elecom is a portable loudspeaker and tends to generate an electromagnetic noise when in use. The iPhone features no distinctive electromagnetic noise but produces a slightly more muffled sound than the original sound. For all the data in DB1, the TMG was also used for the testing part.

For the second database (DB2), we assume that spoof recording was carried out secretly. Therefore, only noisy recordings for spoofing were prepared. For DB2, two types of microphones were used for spoof recording, a Sony C-357 (Sony-C, a condenser microphone) and the TMG. Two Sony-Cs were installed in parallel and facing the same direction. For replay attacks, four different types of loudspeakers were used: the Elecom, Sanwa Supply MM-SPL8UBK (SNW), JBL Professional Control 2P (JBL), and Huawei P20 Lite (Huawei). The SNW is a small loudspeaker powered by USB. The JBL is a desktop loudspeaker. It is 159-mm wide, 143-mm deep, and 235-mm high. The Huawei is a smartphone and has the same features as the iPhone. The TMG or the Sony-C was used for the detection test for DB2.

To analyze the effects on the combination of the environments, four situations were carried out:

(N-Q) Noisy-Quiet: Spoof and test recordings carried out in noisy and quiet environments, respectively.

(N-N) Noisy-Noisy: Both recordings carried out in a noisy environment.

(Q-Q) Quiet-Quiet: Both recordings carried out in a quiet environment.

(Q-N) Quiet-Noisy: Spoof and test recordings carried out in quiet and noisy environments, respectively.

For DB1, all four situations were carried out. The average signal-to-noise ratio (SNR) of DB1 was set to about 18 dB. For DB2, only N-Q and N-N were carried out. The average SNR of DB2 was set to about 14 dB. Comparing these situations with the ASVspoof 2019 settings, the room size for DB1 and DB2 was 5–10 square meters, which corresponded to ASVspoof 2019 ENVIRONMENT_ID S = b. The Talker-to-ASV distance for DB1 was 10–50 cm, which corresponded to ENVIRONMENT_ID D_s = a, and that for DB2 was 50–100 cm, which corresponded to ENVIRONMENT_ID D_s = b. The Attacker-to-ASV distance was about 10 cm for DB1 and DB2, which corresponded to ATTACK_ID D_a = A.

DB1 consisted of 40 genuine speech samples uttered by two male and two female speakers and 640 spoofing attack samples obtained by replaying the genuine speech samples. DB2 consisted of 150 genuine speech samples uttered by three male and two female speakers and 2,400 spoofing attack samples obtained by replaying the genuine speech samples. For DB1, all speech samples were sampled at 16 kHz. For DB2, different recording conditions were used for each microphone for spoof recording. The samples recorded by TMG were sampled at 16 kHz, and those recorded by Sony-C were sampled at 48 kHz.

The ASVspoof 2019 database for the PA scenario contained three parts; training, development and test. From this database, only training data was used for training the ResNet system. The training set included 48,600 spoof utterances and 5,400 genuine utterances.

5.2 Comparison Methods

As a benchmark system, we used two GMM-based systems with CQCC and LFCC as spectral features, respectively. For the training of the benchmark systems, we used the same manner as defined in ASVspoof 2019 for 16-kHz sampled conditions and the parameter of the systems was simply tripled for 48-kHz sampled conditions. To train each GMM, we used 900 genuine utterances and 900 replayed utterances from the Voice Liveness Detection (VLD) database [11]. In Ref. [11], the proposed VLD method required stereo signals for a detecting genuine speech from a replayed one. All utterances in the VLD database were recorded through two AKGs, and the spoof utterances were replayed by a Bose 111AD loudspeaker. The mean and standard deviation scores for z-score normalization were calculated with the VLD database. In all experiments using the GCC-based methods, hand-labeled data was used for the start point t_s and the end point t_e of each utterance. For GCC(avg), the average time was 0.5

Table 1 System performance in terms of EER for DB1 and DB2 (TMG: 16 kHz, Sony-C: 48 kHz).

		DB1				DB2			
Testing microphone		TMG				TMG		Sony-C	
Situation		N-Q	N-N	Q-Q	Q-N	N-Q	N-N	N-Q	N-N
Single system	Score								
GCC(min) [13]	$\Phi_{\min}$	2.73	6.07	4.09	7.27	9.80	20.56	3.79	15.61
GCC(avg) [13]	Φ_{avg}	4.32	6.00	4.20	7.39	**4.88**	**7.86**	1.25	5.51
CQCC [17]	LLR_{CQ}	37.12	35.24	39.70	33.00	39.27	40.30	20.51	13.93
LFCC [18]	LLR_{LF}	39.68	38.74	39.75	37.45	43.50	43.39	10.67	7.87
ResNet [18]	Classification Score	48.75	47.40	44.62	42.70	46.40	48.32	17.12	20.76
Fusion system									
GC(min)-CQ	$\Phi_{\min} + \text{LLR}_{\text{CQ}}$	5.00	4.74	7.61	6.67	11.83	23.56	3.29	10.04
GC(min)-LF	$\Phi_{\min} + \text{LLR}_{\text{LF}}$	4.09	3.89	5.91	5.50	15.81	24.64	2.33	5.98
GC(min)-RN	$\Phi_{\min}$ + CS	10.50	10.48	10.65	14.40	25.03	35.96	3.33	10.33
GC(avg)-CQ	$\Phi_{\text{avg}} + \text{LLR}_{\text{CQ}}$	11.55	8.40	5.42	7.88	7.70	13.54	0.98	4.22
GC(avg)-LF	$\Phi_{\text{avg}} + \text{LLR}_{\text{CQ}}$	12.09	8.20	2.73	7.00	7.42	15.18	0.30	2.56
GC(avg)-RN	Φ_{avg} + CS	11.22	9.55	9.00	12.20	17.56	28.05	0.89	4.13
GC(min)-GC(avg)	$\Phi_{\min} + \Phi_{\text{avg}}$	2.29	2.86	2.86	4.33	5.00	10.06	1.41	7.00
CQ-LF	$\text{LLR}_{\text{CQ}} + \text{LLR}_{\text{LF}}$	37.66	35.55	39.24	35.85	41.08	42.68	13.33	9.12
CQ-RN	LLR_{CQ} + CS	41.52	36.17	38.46	33.63	46.81	46.90	8.76	8.73
LF-RN	LLR_{LF} + CS	47.11	43.17	39.50	38.80	46.22	46.83	6.35	5.33
GC(min)-CQ-LF	$\Phi_{\min} + \text{LLR}_{\text{CQ}} + \text{LLR}_{\text{LF}}$	10.00	8.51	11.18	9.79	15.78	29.42	3.62	5.56
GC(avg)-CQ-LF	$\Phi_{\text{avg}} + \text{LLR}_{\text{CQ}} + \text{LLR}_{\text{LF}}$	13.77	12.67	8.57	10.52	10.56	19.43	1.26	3.24
GC(min)-GC(avg)-CQ	$\Phi_{\min} + \Phi_{\text{avg}} + \text{LLR}_{\text{CQ}}$	3.64	**1.67**	3.24	3.33	6.29	12.58	0.15	5.79
GC(min)-GC(avg)-LF	$\Phi_{\min} + \Phi_{\text{avg}} + \text{LLR}_{\text{LF}}$	**2.22**	**1.67**	**1.82**	**2.22**	7.37	13.99	**0.00**	**2.53**
GC(min)-GC(avg)-RN	$\Phi_{\min} + \Phi_{\text{avg}}$ + CS	4.69	4.41	4.50	5.36	14.00	25.11	0.89	5.25
GC(min)-GC(avg)-CQ-LF	$\Phi_{\min} + \Phi_{\text{avg}} + \text{LLR}_{\text{CQ}} + \text{LLR}_{\text{LF}}$	4.09	2.78	4.71	3.75	9.40	15.91	0.15	3.35
GC(min)-GC(avg)-CQ-LF-RN	$\Phi_{\min} + \Phi_{\text{avg}} + \text{LLR}_{\text{CQ}} + \text{LLR}_{\text{LF}}$ + CS	3.82	3.41	5.83	4.48	13.31	23.31	0.15	2.15

GC: GCC, CQ: CQCC, LF: LFCC, RN: ResNet

seconds from T_s to t_s and t_e to T_e. For the GCC-based methods, the frame length was set to 256 points for 16-kHz sampled signals and 1,024 points for 48-kHz sampled signals. For the score-fusion systems, all combinations of the GCC-based methods and the spectral feature-based methods were compared as shown in **Table 1**. The equal error rate (EER) was used for an evaluation measurement.

Since the GCC-based methods require stereo signals, the score-fusion systems cannot be evaluated with the ASVspoof database. Instead of adopting the ASVspoof 2019 database for the systems, the ResNet system was used with the evaluation data of this experiment. The authors of Ref. [20] provided the software for a single ResNet system on GitHub. For training the ResNet system, the ASVspoof 2019 database was used. Since the ASVspoof 2019 database was sampled at 16 kHz, the data recorded by SONY-C was downsampled from 48 kHz to 16 kHz only for the ResNet system.

5.3 Results

Table 1 shows the EERs of each spoofing detection system for DB1 and DB2. First, the results of DB1 are discussed. Comparing situation N-Q with N-N or Q-Q with Q-N, it can be seen that the EERs of the GCC-based single systems were higher in the noisy recording for testing than those in the quiet recording. While most of the single GCC-based systems obtained low EERs, the EERs of CQCC, LFCC and ResNet were comprehensively high. One reason was the mismatches between the training data and the test one. The domain of ASVspoof 2019 fairly differs from our databases (DB1 and DB2). Although the VLD database was recorded with stereo signals, the recording conditions and the other details were not same from DB1 and DB2. As considering tendency between situations, opposite to the GCC-based systems, the performances of CQCC, LFCC, and ResNet in the noisy recording for testing were better than those in the quiet testing recording. For example, comparing situation N-Q with N-N or Q-Q with Q-N, it can be seen that the EERs of the machine learning-based systems in the situations N-Q and Q-Q were higher than those in the situation N-N and Q-N. In the case of the score fusion with two systems, the combination of two GCC-based methods GC(min)-GC(avg) achieved lower EERs than those of two-system combinations for all situations. The system stability increased when using both scores of the GCC-based systems. In the case of the score fusion with three systems, GC(min)-GC(avg)-LF achieved the lowest EERs for all situations. In comparison, the score fusion with four systems could not improve the performance more than GC(min)-GC(avg)-LF. This indicates that the characteristics extracted by CQCC were not suitable for combination with the spatial features, but LFCC was suitable for this.

Next, the results with DB2 in Table 1 are discussed. In the case of using TAMAGO for test recording, all score-fusion systems had lower performances than the single GCC(avg). For the TAMAGO recording, the SNRs of almost all test utterances were lower than the average SNR. In Ref. [13], it was also discussed that a test recording requires a sufficient enough SNR in order for GCC-based methods to perform well. This means that when SNRs are low, it is difficult to detect spoofing attacks as well as CQCC and LFCC-based methods and ResNet system. In contrast, in the case of using Sony-C for test recording, fusion systems GC(min)-GC(avg)-LF yielded the lowest EERs compared with the single GC(avg) the same as in the results with DB1. In this case, the SNRs were almost the same as those of DB1. From these results, if the quality of the testing microphone is high and

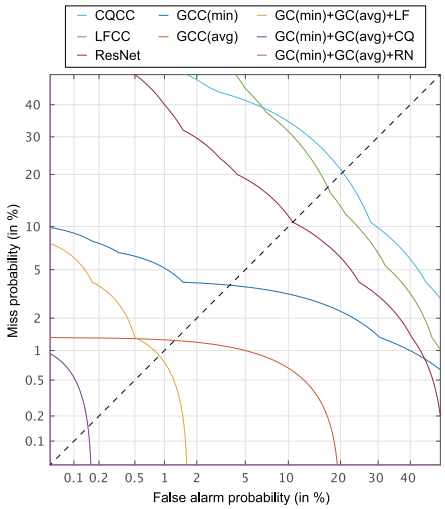

Fig. 5 DET curves of each single system and combination system in the N-Q situation with Sony-C in DB2.

Table 2 System performance in terms of EER with Sony-C for DB2 downsampled to 16 kHz.

Testing microphone: SONY-C	16 kHz only		16 kHz and 48 kHz	
Situation	N-Q	N-N	N-Q	N-N
Single system				
GC(min) [13]	1.28	10.76	1.28	10.76
GC(avg) [13]	1.21	4.91	1.21	10.76
CQ [17]	44.67	45.84	20.51	13.93
LF [18]	41.74	43.43	10.67	7.87
RN [20]	17.12	20.76	17.12	20.76
Fusion system				
GC(min)-CQ	3.11	18.02	0.72	7.72
GC(min)-LF	3.25	18.04	0.48	4.74
GC(min)-RN	1.33	8.89	1.33	8.89
GC(avg)-CQ	2.44	6.56	0.80	3.14
GC(avg)-LF	3.03	5.20	0.36	1.33
GC(avg)-RN	0.92	4.29	0.92	4.29
GC(min)-GC(avg)	0.15	6.03	0.15	6.03
CQ-LF	41.98	45.14	13.33	9.12
CQ-RN	23.42	23.81	8.76	8.73
LF-RN	22.95	20.94	6.35	5.33
GC(min)-CQ-LF	8.72	23.18	1.20	5.07
GC(avg)-CQ-LF	4.49	9.23	**0.00**	2.13
GC(min)-GC(avg)-CQ	0.71	7.00	**0.00**	2.48
GC(min)-GC(avg)-LF	0.91	6.18	**0.00**	**1.33**
GC(min)-GC(avg)-RN	**0.00**	**4.22**	**0.00**	4.22
GC(min)-GC(avg)-CQ-LF	1.14	7.75	**0.00**	1.89
GC(min)-GC(avg)-CQ-LF-RN	0.44	5.44	**0.00**	**1.13**

a situation in which a suitable SNR can be arranged, the score-fusion system could perform well without situation dependence. Since conditions are prepared by developers who want to protect systems from replay attacks, the systems using the proposed method can be regarded as a realistic technique. **Figure 5** shows DET curves of each single system and a combination system in the N-Q situation with Sony-C in DB2. In the case of the combination methods, the weakness of the single systems was relaxed with the score fusion. These results show the fusion systems were effectively performed.

From our primitive experiments, the performance of the GCC-based methods depended on sampling frequencies. To analyze the effects of the sampling frequency on the score fusion systems, DB2 using SONY-C as test microphone was downsampled from 48 kHz to 16 kHz. **Table 2** shows the EERs of each spoofing detection system for downsampled DB2. From the results of the rows "16 kHz only" in Table 2, the EERs of single GCC-based methods were lower in downsampled DB2 than in original DB2. However, the EERs of CQCC and LFCC were drastically increased in downsampled DB2. Due to the influence of single performances, the EERs of fusion systems which included CQCC or LFCC, such as GC(min)-CQ-LF and GC(avg)-CQ-LF, gained as well. On the other hand, the EERs of the fusion systems which included the GCC-based methods mainly, such as GC(min)-GC(avg), GC(min)-GC(avg)-CQ, GC(min)-GC(avg)-LF and GC(min)-GC(avg)-RN, reduced compared with those in original DB2. From these results, the spectral feature-based systems were adequate to use a higher sampling rate. Therefore, we performed some experiments under the adequate conditions for each system. The columns "16 kHz and 48 kHz" in Table 2 mean that the GCC-based methods and ResNet performed with downsampled DB2 at 16 kHz, and CQCC and LFCC were carried out with original DB2 sampled at 48 kHz. The results of "16 kHz and 48 kHz" show the fusion systems in the adequate conditions can improve the performances compared with those in rows "16 kHz only." Especially, the fusion system with all methods GC(min)-GC(avg)-CQ-LF-RN achieved the lowest EERs in both situations. And, "GC(min)-GC(avg)-LF" was the second best in both situations.

Considering the results of the Tables 1 and 2, the fusion system GC(min)-GC(avg)-LF performed the best in all situations. Thus, the proposed method using spatial and spectral feature, especially LFCC, outperformed the conventional systems and obtained a stable performance under several real situations.

6. Conclusion

We proposed a spatial and spectral feature-based RAD method. In previous work, as spatial-based methods, we proposed GCC-based RAD methods. While GCC-based methods have been reported to perform well under primitive experiments, the methods still suffer from situation dependency. Since spectral features extract different characteristics compared with GCC-based methods, it is expected that fusing the output scores of spatial and spectral feature-based methods can enable the methods to compensate for each other and improve robustness. From the experimental results, it was confirmed that the systems using the proposed method achieved the lowest EERs in almost all situations.

In future work, the proposed methods will also be combined with other spoofing countermeasures. Additionally, we will consider to use more complicated models such as a DNN-based modeling approach for the GCC-based method, and evaluation tests will be performed under a large amount of data.

Acknowledgments This work was supported, in part, by JSPS KAKENHI Early-Career Scientists Grant number JP19K20271, Grant-in-Aid for Scientific Research (A) JP16H01375, and ROIS-DS-JOINT (021RP2019) to S. Shiota.

References

[1] Verint VoiceVault: HSBC Embraces Mobile Voice Biometric Security Technology, available from ⟨http://voicevault.com/hsbc-embraces-mobile-voice-biometric-security-technology/⟩.
[2] SpeechPro: VoiceKey.ONEPASS: Bimodal biometric authentication for mobile platforms, available from ⟨https://speechpro-usa.com/

product/voice_authentication/voicekey-onepass⟩.
[3] Aley-Raz, A., Krause, N.M., Salmon, M.I. and Gazit, R.Y.: Device, system, and method of liveness detection utilizing voice biometrics, U.S. Patent No.8,442,824 (2013).
[4] Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F. and Li, H.: Spoofing and countermeasures for speaker verification: A survey, *Speech Communication*, Vol.66, pp.130–153 (2015).
[5] Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J., Hanilçi, C., Sahidullah, M. and Sizov, A.: ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge, *Proc. Interspeech* (2015).
[6] Wu, Z., Yamagishi, J., Kinnunen, T., Hanilci, C., Sahidullah, M., Sizov, A., Evans, N., Todisco, M. and Delgado, H.: ASVspoof: The automatic speaker verification spoofing and countermeasures challenge, *IEEE Journal of Selected Topics in Signal Processing*, Vol.11, No.4, pp.588–604 (2017).
[7] ASVspoof consortium: ASVspoof 2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan, available from ⟨http://www.asvspoof.org/asvspoof2019/asvspoof2019_evaluation_plan.pdf⟩.
[8] Chen, L., Guo, W. and Dai, L.: Speaker verification against synthetic speech, *Proc. 7th International Symposium on Chinese Spoken Language Processing*, pp.309–312 (2010).
[9] Lavrentyeva, G., Novoselov, S., Malykh, E., Kozlov, A., Kudashev, O. and Shchemelinin, V.: Audio Replay Attack Detection with Deep Learning Frameworks, *Proc. Interspeech*, pp.82–86 (2017).
[10] Chen, Z., Xie, Z., Zhang, W. and Xu, X.: Resnet and model fusion for automatic spoofing detection, *Proc. Interspeech*, pp.102–106 (2017).
[11] Shiota, S., Villavicencio, F., Yamagishi, J., Ono, N., Echizen, I. and Matsui, T.: Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification, *Proc. Interspeech*, pp.239–243 (2015).
[12] Zhang, L., Tan, S., Yang, J. and Chen, Y.: Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones, *Proc. ACM CCS*, pp.1080–1091 (2016).
[13] Yaguchi, R., Shiota, S., Ono, N. and Kiya, H.: Replay Attack Detection Using Generalized Cross-Correlation of Stereo Signal, *Proc. EURASIP EUSIPCO* (2019).
[14] Knapp, C. and Carter, G.: The generalized correlation method for estimation of time delay, *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol.24, No.4, pp.320–327 (1976).
[15] Yaguchi, R., Shiota, S., Ono, N. and Kiya, H.: Improving replay attack detection by combination of spatial and spectral features, *Proc. APSIPA ASC*, pp.833–837 (2019).
[16] Yaguchi, R., Shiota, S., Ono, N. and Kiya, H.: Spoofing detection method using generalized cross-correlation between multiple channels for speaker verification, *Proc. ASJ*, pp.1335–1338 (2018).
[17] Todisco, M., Delgado, H. and Evans, N.: A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients, *Proc. Odyssey*, Vol.25, pp.249–252 (2016).
[18] Sahidullah, M., Kinnunen, T. and Hanilci, C.: A comparison of features for synthetic speech detection, *Proc. Interspeech*, pp.2087–2091 (2015).
[19] Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., Yamagishi, J., Evans, N., Kinnunen, T. and Lee, K.: ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection, arXiv preprint arXiv:1904.05441 (2019).
[20] Jung, J., Shim, H., Heo, H. and Yu, H.: Replay attack detection with complementary high-resolution information using end-to-end DNN for the ASVspoof 2019 Challenge, *Proc. Interspeech*, pp.1083–1087 (2019).
[21] Lai, C., Chen, N., Villalba, J. and Dehak, N.: ASSERT: Anti-Spoofing with Squeeze-Excitation and Residual neTworks, *Proc. Interspeech*, pp.1013–1017 (2019).
[22] Cai, W., Wu, H., Cai, D. and Li, M.: The DKU Replay Detection System for the ASVspoof 2019 Challenge: On Data Augmentation, Feature Representation, Classification, and Fusion, *Proc. Interspeech*, pp.1023–1027 (2019).
[23] Williams, J. and Rownicka, J.: Speech Replay Detection with x-Vector Attack Embeddings and Spectral Features, *Proc. Interspeech*, pp.1053–1057 (2019).
[24] Chettri, B., Stoller, D., Morfi, V., Ramírez, M., Benetos, E. and Sturm, B,: Ensemble Models for Spoofing Detection in Automatic Speaker Verification, *Proc. Interspeech*, pp.1018–1022 (2019).
[25] Zeinali, H., Stafylakis, T., Athanasopoulou, G., Rohdin, J., Gkinis, I., Burget, L., Černockỳ, J., et al.: Detecting spoofing attacks using VGG and sincnet: But-omilia submission to ASVspoof 2019 challenge, *Proc. Interspeech*, pp.1073–1077 (2019).
[26] Alzantot, M., Wang, Z. and Srivastava, M.: Deep Residual Neural Networks for Audio Spoofing Detection, *Proc. Interspeech*, pp.1078–1082 (2019).
[27] Tom, F., Jain, M. and Dey, P.: End-To-End Audio Replay Attack Detection Using Deep Convolutional Networks with Attention, *Proc. Interspeech*, pp.681–685 (2018).
[28] Chakroborty, S., Roy, A. and Saha, G.: Improved closed set text-independent speaker identification by combining MFCC with evidence from flipped filter banks, *International Journal of Signal Processing*, Vol.4, No.2, pp.114–122 (2007).
[29] Hasan, T., Sadjadi, S., Liu, G., Shokouhi, N., Bořil, H. and Hansen, J.: CRSS systems for 2012 NIST speaker recognition evaluation, *Proc. ICASSP*, pp.6783–6787 (2013).
[30] He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition, *Proc. CVPR*, pp.770–778 (2016).

Ryoya Yaguchi was born in 1995. He received his B.E and M.S. degrees from Tokyo Metropolitan University, Tokyo, Japan in 2018 and 2020, respectively.

Sayaka Shiota received her B.E., M.E. and Ph.D. degrees in Intelligence and Computer Science, Engineering and Engineering Simulation from Nagoya Institute of Technology, Nagoya, Japan in 2007, 2009 and 2012, respectively. From February 2013 to March 2014, she had worked at the Institute of statistical mathematics as a project assistant professor. In April of 2014, she joined Tokyo Metropolitan University as an Assistant Professor. Her research interests include statistical speech recognition and speaker verification. She is a member of Acoustical Society of Japan (ASJ), IPSJ, IEICE, APSIPA, and IEEE.

Nobutaka Ono received his B.E., M.S., and Ph.D. degrees in Mathematical Engineering and Information Physics from the University of Tokyo, Japan, in 1996, 1998, 2001, respectively. He has worked with Tokyo Metropolitan University as a professor since Oct. 2017. His research interests include microphone array processing, blind source separation, and optimization algorithms for them. He is the author or co-author of more than 250 articles in international journal papers and peer-reviewed conference proceedings. He is a senior member of the IEEE signal processing society, and has been a member of IEEE Audio and Acoustic Signal Processing (AASP) technical committee since 2014. He received the best paper awards from IEEE ISIE in 2008 and from IEEE IS3C in 2014, the unsupervised learning ICA pioneer award from SPIE.DSS in 2015, the Sato paper award from the acoustic society of Japan and two telecom system technology awards from the telecommunications advancement foundation in 2018.

Hitoshi Kiya received his B.E and M.E. degrees from Nagaoka University of Technology, in 1980 and 1982 respectively, and his Dr.Eng. degree from Tokyo Metropolitan University in 1987. In 1982, he joined Tokyo Metropolitan University, where he became a Full Professor in 2000. From 1995 to 1996, he attended the University of Sydney, Australia as a Visiting Fellow. He is a Fellow of IEEE, IEICE and ITE. He served as President of APSIPA, and as Regional Director-at-Large for Region 10 of the IEEE Signal Processing Society from 2016 to 2017. He was also President of the IEICE Engineering Sciences Society from 2011 to 2012, and he served there as a Vice President and Editor-in-Chief for IEICE Society Magazine and Society Publications. He was Editorial Board Member of eight journals, including IEEE Trans. on Signal Processing, Image Processing, and Information Forensics and Security, Chair of two technical committees and Member of nine technical committees including APSIPA Image, Video, and Multimedia Technical Committee (TC), and IEEE Information Forensics and Security TC. He has organized a lot of international conferences, in such roles as TPC Chair of IEEE ICASSP 2012 and as General Co-Chair of IEEE ISCAS 2019. He has received numerous awards, including ten best paper awards.