# Improving replay attack detection by combination of spatial and spectral features

Yaguchi Ryoya* and Sayaka Shiota* and Nobutaka Ono* and Hitoshi Kiya*
* Tokyo Metropolitan University, Faculty of System Design, Department of Computer Science, Japan
E-mail: yaguchi-ryoya@ed.tmu.ac.jp

***Abstract*—In this paper, we propose a replay attack detection based on score fusion of spatial and spectral features-based systems. Recently, a replay attack detection (RAD) system using generalized cross-correlation (GCC) of a stereo signal has been proposed. The GCC is calculated from non-speech sections of input signals. It reported that the GCC-based method achieved high performance under several situations. However, since the performance of the GCC-based method depends on the situations, it is required to improve the performance without situation dependence. The GCC-based method uses spatial features, which utilize the different feature from spectral features. In this paper, we perform score fusion of the GCC-based and the spectral feature-based methods to improve the robustness of RAD systems. In the experiments, the proposed method achieved a relative error reduction of 69.5%, compared with a GCC-based single method under one of the hard tasks. And, the performance of score fusion systems improved without situation dependence.**

## I. Introduction

Recently, biometric authentication systems have become popular in various situations such as banking protection and immigration control. Automatic speaker verification (ASV), which uses voices as a biometric template, is one of the biometric authentication techniques. Due to using voice template, ASV systems can easily be linked with voice interface systems. On the other hand, it has been reported that spoofing attacks (i.e., replay and speech synthesis) have become a serious problem for ASV systems [1]. To consider countermeasures of spoofing attacks, the ASV Spoofing and Countermeasures (ASVspoof) challenges were held in 2015 [2], 2017 [3] and 2019 [4]. Through the ASVspoof challenges, many methods based on various spectral features have been proposed [5]–[7]. Such the spectral features are treated as "acoustic features" in this paper.

The ASVspoof challenges assume two types of spoofing attacks. One is physical access (PA) attack and the other is logical access (LA) attack. The flow of the PA attack is shown in Fig. 1. Since ASVspoof database is recorded by single-channel microphones, almost all proposed countermeasures assume the single-channel situation. Meanwhile, since recording with multi-channel microphone has become easy, the RAD system assuming the multi-channel recording has also been proposed [8]–[10]. In [10], generalized cross-correlation (GCC) of a stereo signal is used for replay attack detection. It focuses on non-speech sections, the reason why no sound is emitted from human but loudspeakers tend to generate some noise and non-perceptual sound during non-speech section. However, the GCC-based method is required to improve the performance due to restrictions of assumed situations. In this paper, we perform score fusion of the GCC-based and the acoustic feature-based methods. For the acoustic feature-based methods, the benchmark system of ASVspoof 2019 was used. In the experiments, the proposed method achieved lower equal error rate than those of single GCC-based methods. And, the performance of score fusion systems improved without situation dependence.

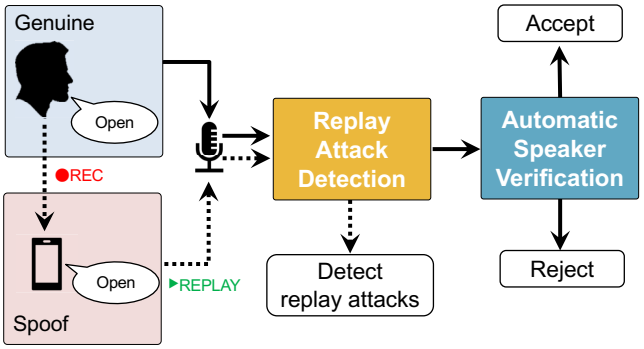


Fig. 1. System flow

The remainder of this paper is organized as follows. Related work using cross-correlation method is detailed in section 2. Section 3 introduces the proposed score fusion system using cross-correlation and spectral features. Section 4 describes the experimental setup and the results of detection tests. Finally, section 5 concludes this paper.

## II. Related Work

### A. Characteristics for loudspeakers in non-speech sections

The signals recorded by two microphones $a$ and $b$ for a genuine speaker can be represented as follows in the time-frequency domain:

$$M_a(t,f) = H_a(f)S(t,f) + N_a(t,f), \quad (1)$$
$$M_b(t,f) = H_b(f)S(t,f) + N_b(t,f), \quad (2)$$

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

$$M_a(t,f) = N_a(t,f), \quad (3)$$
$$M_b(t,f) = N_b(t,f). \quad (4)$$

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

On the other hand, the replay attack case is different. Let

$$M_p(t,f) = H_p(f)S(t,f) + N_p(t,f), \tag{5}$$

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

$$M_a(t,f) = H'_a(f)(M_p(t,f) + N_s(t,f)) + N_a(t,f), \tag{6}$$
$$M_b(t,f) = H'_b(f)(M_p(t,f) + N_s(t,f)) + N_b(t,f), \tag{7}$$

where $H'_a(f)$ and $H'_b(f)$ are transfer functions and $N_s(t;f)$ represents the electromagnetic noise generated by the loudspeaker. In non-speech sections, $S(t,f) = 0$ yields $M_p(t,f) = N_p(t,f)$. Then, Eqs. (9) and (10) can be rewritten as

$$M_a(t,f) = H'_a(f)(N_p(t,f) + N_s(t,f)) + N_a(t,f), \tag{8}$$
$$M_b(t,f) = H'_b(f)(N_p(t,f) + N_s(t,f)) + N_b(t,f). \tag{9}$$

The equations mean that even in non-speech sections, the recorded noise $N_p(t,f)$ and the electromagnetic noise $N_s(t,f)$ are still omitted. The electromagnetic noise may not be heard by human. Then, the noise can be localized and GCC can still take a high value. These characteristics make it possible to distinguish spoofing attacks from genuine utterances.

In a genuine-speaker case, the maximum generalized cross correlation (GCC) [11] is considered to be low in the non-speech sections because no sound is emitted from a genuine speaker. On the other hand, in the case of a loudspeaker, since the recorded noise or the electromagnetic noise of the loudspeaker can be emitted even in the non-speech sections, the maximum GCC can be high. As an example, Fig. 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 denote non-speech sections. According to these trajectories, the maximum GCC takes the lower values for the genuine utterance, and the maximum GCC of the replayed utterance has higher values. Figure 2(c) shows the GCC of one frame in both speech section and non-speech section for the genuine and the replayed utterances. The red dots denote the maximum point in each frame. In the speech sections, the peak of both utterances had a high value. In the non-speech sections, 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 can be regarded as an effective factor for localizing loudspeakers.

### *B. Spoofing detection using the maximum GCC in non-speech sections*

The GCC-based method [10] focused on the trajectories of the maximum GCC (max-GCC) values in non-speech sections for spoofing detection. The max-GCC for each frame is defined as

$$\phi_{max}(t) = \max_d \phi_g(\tau, d; t). \tag{10}$$

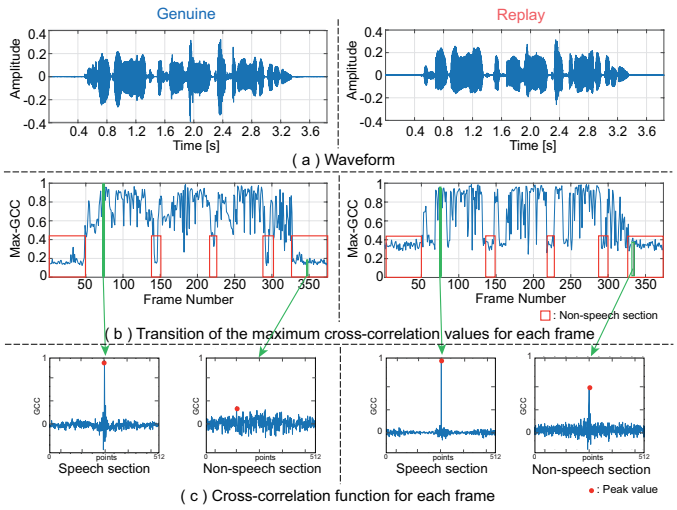


Fig. 2. Trajectories of the maximum GCC for each frame

As shown in Fig. 2(b), there are two types of non-speech sections: "short pauses" are in from a start point of speaking to an end point of one and "silent sections" are in both before the start speaking and after the end of one. Therefore, two scores are defined. One is focusing the minimum value of the max-GCCs in short pauses, which is named GCC(min). The other is focusing the average value of the max-GCCs in silent section, which is named GCC(avg). These definitions are expressed as:

$$\text{GCC(min):}\quad \Phi_{min} = \min_{t_s \le t \le t_e} \phi_{max}(t), \tag{11}$$

$$\text{GCC(avg):}\quad \Phi_{ave} = \frac{1}{K} \sum_{T_s \le t < t_s, t_e < t \le T_e} \phi_{max}(t), \tag{12}$$

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 section $t$. 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 a parameter $T$ represents the end point of an utterance. In this paper, these methods were treated as the GCC-based methods.

### *C. ASVspoof 2019 challenge benchmark system*

ASVspoof 2019 challenge focuses on spoofing countermeasures. The challenge provided two benchmark systems using Gaussian Mixture Model (GMM)-based backend classifier. The GMM of each system is trained with acoustic features, CQCC [12] and LFCC [13], respectively. In the past ASVspoof challenges, many countermeasures used CQCC and LFCC as one of effective acoustic features. Thus, they are adopted as benchmark systems. The features extracted from input speech signals, and a log likelihood ratio (LLR) is calculated by using the GMMs as below.

$$\text{LLR} = \log p(\chi|H_0) - \log p(\chi|H_1), \tag{13}$$

where $\chi$ is a speech utterance, $H_0$ is null hypothesis and $H_1$ is alternative hypothesis which are correspond to $\chi_t$ is a genuine

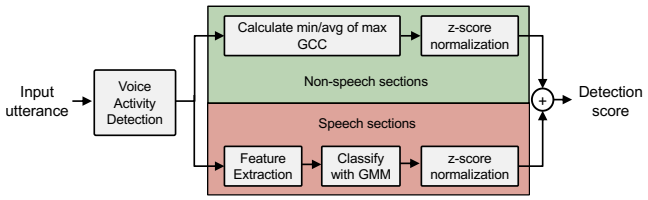


Fig. 3. Flow of the proposed method

speech or a spoof speech. $p$ represents conditional probabilities whether $\chi$ is $H_0$ or $H_1$. Since the LLR is calculated per frame, the average of LLR for an utterance is referd as "LLR". In this paper, reverse sign of LLR is used as a detection score.

## III. Score Fusion System

### A. Motivation

In [10], it has reported that the GCC-based method has high performance especially under quiet situations. On the other hand, the performance of the GCC-based method depended on the situations. Thus, it is required to improve the robustness without situation dependence.

The GCC-based method focuses on the spatial characteristics in non-speech sections. Through the ASVspoof challenges, many approaches based on various kinds of acoustic features have been reported [5]–[7]. Since these acoustic features are extracted from spectral characteristics, the different characteristics from the GCC-based methods can be utilized. Thus, it expects that score fusion of the spatial and spectral feature-based systems can compensate each other and improve the robustness.

### B. Procedure

The procedure of a score fusion system is illustrated in Fig. 3. First, an input utterance is separated into speech and non-speech sections by voice activity detection (VAD). From all non-speech sections of an input utterance, the GCC scores $\Phi_{min}$ and $\Phi_{avg}$ are calculated by Eqs. (11), (12). From all frames of speech sections, average LLR is calculated by the trained GMMs with acoustic features. Each detection score is normalized by using the z-score normalization as follows:

$$z = \frac{x - \mu}{\sigma}, \tag{14}$$

where $z$ is the normalized score, $x$ is the input score, $\mu$ and $\sigma$ are the mean and the standard deviation of a training set. Finally, the fused score is used as the detection score. We show an example of fused score $S$ with GCC(min) and LLR of CQCC as follows:

$$S = \frac{\Phi_{\min} - \mu_{\Phi_{\min}}}{\sigma_{\Phi_{\min}}} + \frac{\mathrm{LLR}_{\mathrm{CQ}} - \mu_{\mathrm{LLR}_{\mathrm{CQ}}}}{\sigma_{\mathrm{LLR}_{\mathrm{CQ}}}}, \tag{15}$$

where $\mu_{\Phi_{\min}}$ and $\mu_{\mathrm{LLR}_{\mathrm{CQ}}}$ are the average of $\Phi_{\min}$ and $\mathrm{LLR}_{\mathrm{CQ}}$, respectively. Also, $\sigma_{\Phi_{\min}}$ and $\sigma_{\mathrm{LLR}_{\mathrm{CQ}}}$ are the standard deviation of each score. In this paper, $\mu$ and $\sigma$ of each score are calculated from the other database for replay attack detection.

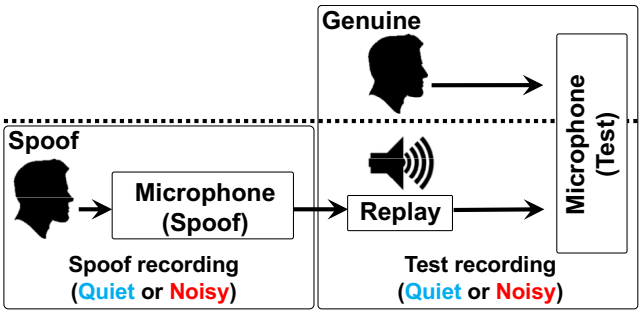


Fig. 4. Testing flow and recording process

## IV. Experiments

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

### A. Database

Figure 4 illustrates the testing flow of the experiments in both cases of genuine and spoof. There were two types of recording processes: spoof and test. Although several recording situations can be considered for both recording processes, two environments "Quiet" and "Noisy" were assumed for the experiments. "Quiet" indicates there was no extra background noise such as an air conditioner in a common space. "Noisy" represents there were stationary sound such as an air conditioner running on low and non-stationary sound such as a TV program playing at a moderate volume in the same room of "Quiet". To construct databases of stereo signals for replay attack detection, all situations based on 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 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 the speaker's position. For the TMG, two of the eight microphone channels were used whereas two AKGs were installed in parallel facing 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 INSPiRE2.0.1300 (CI). The SONY-S is 300 mm wide, 86 mm deep and 93 mm high. It generates non-perceptual electromagnetic noise in silent sections of replayed attacks. The CI is a separate stereo loudspeaker. It is 99 mm wide, 131 mm deep and 221 mm high for each one. The ELECOM is a portable loudspeaker and tends to generate 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 the spoof recording carried out secretly. Therefore, only noisy recording for

TABLE I
COMPARISON SYSTEMS

| System | Score | Score fusion |
|---|---|---|
| GCC(min) | $\Phi_{min}$ | - |
| GCC(avg) | $\Phi_{avg}$ | - |
| CQCC | $\mathrm{LLR_{CQ}}$ | - |
| LFCC | $\mathrm{LLR_{LF}}$ | - |
| GC(min)-CQ | $\Phi_{min}$ + $\mathrm{LLR_{CQ}}$ | ✓ |
| GC(min)-LF | $\Phi_{min}$ + $\mathrm{LLR_{LF}}$ | ✓ |
| GC(avg)-CQ | $\Phi_{avg}$ + $\mathrm{LLR_{CQ}}$ | ✓ |
| GC(avg)-LF | $\Phi_{avg}$ + $\mathrm{LLR_{LF}}$ | ✓ |
| GC(min)-GC(avg) | $\Phi_{min}$ + $\Phi_{avg}$ | ✓ |
| CQ-LF | $\mathrm{LLR_{CQ}}$ + $\mathrm{LLR_{LF}}$ | ✓ |
| GC(min)-CQ-LF | $\Phi_{min}$ + $\mathrm{LLR_{CQ}}$ + $\mathrm{LLR_{LF}}$ | ✓ |
| GC(avg)-CQ-LF | $\Phi_{avg}$ + $\mathrm{LLR_{CQ}}$ + $\mathrm{LLR_{LF}}$ | ✓ |
| GC(min)-GC(avg)-CQ | $\Phi_{min}$ + $\Phi_{avg}$ + $\mathrm{LLR_{CQ}}$ | ✓ |
| GC(min)-GC(avg)-LF | $\Phi_{min}$ + $\Phi_{avg}$ + $\mathrm{LLR_{LF}}$ | ✓ |
| GC(min)-GC(avg)-CQ-LF | $\Phi_{min}$ + $\Phi_{avg}$ + $\mathrm{LLR_{CQ}}$ + $\mathrm{LLR_{LF}}$ | ✓ |

spoof were prepared. For DB2, two types of microphones were used for spoof recording, SONY C-357 (SONY-C, a condenser microphone) and the TMG. Two SONY-Cs were installed in parallel 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 as follow under:

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

For DB1, all of 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 14dB. Comparing these situations with ASVspoof 2019 settings, the room size for DB1 and DB2 is categorized into 5-10 square meters. The Talker-to-ASV distance for DB1 is categorized into 10-50 cm and that for DB2 is categorized into 50-100 cm. The Attacker-to-ASV distance is about 10 cm for DB1 and DB2.

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 2400 spoofing attack samples obtained by replaying the genuine speech samples. For DB1, all speech samples were sampled at 16 kHz. For DB2 were adopted different recording conditions for each microphone in the spoof recording. The TMG was sampled at 16 kHz and the SONY-C was sampled at 48 kHz.

### *B. Comparison method*

As the benchmark system, we used GMM-based methods with CQCC and LFCC as acoustic features, respectively. The benchmark systems were trained with the same manner defined in ASVspoof 2019. To train each GMM, we used 900 utterances for genuine and 900 replayed utterances for spoof from VLD database [8]. All of the VLD database was recorded through the AKGs and the spoof utterances were replayed by a BOSE 111AD loudspeaker. The mean and standard deviation scores for z-score normalization were calculated on the VLD database [8].

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 s 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 1024 points for 48 kHz sampled signals.

For the score fusion systems, all combinations of the GCC-based methods and the acoustic feature-based methods were compared as shown in TABLE I. Equal error rate (EER) was used for an evaluation measurement.

### *C. Results*

Table II shows the EERs of each spoofing detection method for DB1. Comparing the 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 worse in the noisy recording for test than those in the quiet recording one. On the contrary, although the EERs of CQCC and LFCC were comprehensively high, the performances of CQCC and LFCC tended to be better in the noisy recording for test than those in the quiet testing one. These results indicated that the spatial and spectral features focused on different characteristics. In the case of the score fusion with two systems, the combination of two GCC-based methods (GC(min)-GC(avg)) achieved the lower EERs than those of two system combination in all situations. It was indicated that the system stability has been increased by 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 in all situation. On the other hand, the score fusion with four systems could not improve the performance than GC(min)-GC(avg)-LF. It indicated that the characteristics extracted by CQCC was not suitable to combine with the spatial features, but LFCC was suitable to combine with the spatial features.

Table III shows the EERs of each spoofing detection method for DB2. In the case using TAMAGO for test recording, all score fusion systems got lower performances than the single GCC(avg). On the TAMAGO recording, the SNRs of almost all test utterances were lower than the average SNR. In [10],

TABLE II
SYSTEM PERFORMANCE AS EER IN DB1

| | Situation | | | |
|---|---|---|---|---|
| | N-Q | N-N | Q-Q | Q-N |
| Single system | | | | |
| GCC(min) [10] | 2.73 | 6.07 | 4.09 | 7.27 |
| GCC(avg) [10] | 4.32 | 6.00 | 4.20 | 7.39 |
| CQCC [12] | 37.12 | 35.24 | 39.70 | 33.00 |
| LFCC [13] | 39.68 | 38.74 | 39.75 | 37.45 |
| Fused system | | | | |
| GC(min)-CQ | 5.00 | 4.74 | 7.61 | 6.67 |
| GC(min)-LF | 4.09 | 3.89 | 5.91 | 5.50 |
| GC(avg)-CQ | 11.55 | 8.40 | 5.42 | 7.88 |
| GC(avg)-LF | 12.09 | 8.20 | 2.73 | 7.00 |
| GC(min)-GC(avg) | 2.29 | 2.86 | 2.86 | 4.33 |
| CQ-LF | 37.66 | 35.55 | 39.24 | 35.85 |
| GC(min)-CQ-LF | 10.00 | 8.51 | 11.18 | 9.79 |
| GC(avg)-CQ-LF | 13.77 | 12.67 | 8.57 | 10.52 |
| GC(min)-GC(avg)-CQ | 3.64 | **1.67** | 3.24 | 3.33 |
| GC(min)-GC(avg)-LF | **2.22** | **1.67** | **1.82** | **2.22** |
| GC(min)-GC(avg)-CQ-LF | 4.09 | 2.78 | 4.71 | 3.75 |

TABLE III
SYSTEM PERFORMANCE AS EER IN DB2

| | Testing microphone | | | |
|---|---|---|---|---|
| | TAMAGO | | SONY-C | |
| | N-Q | N-N | N-Q | N-N |
| Single system | | | | |
| GCC(min) [10] | 9.80 | 20.56 | 3.79 | 15.61 |
| GCC(avg) [10] | **4.88** | **7.86** | 1.25 | 5.51 |
| CQCC [12] | 39.27 | 40.30 | 20.51 | 13.93 |
| LFCC [13] | 43.50 | 43.39 | 10.67 | 7.87 |
| Fused system | | | | |
| GC(min)-CQ | 11.83 | 23.56 | 3.29 | 10.04 |
| GC(min)-LF | 15.81 | 24.64 | 2.33 | 5.98 |
| GC(avg)-CQ | 7.70 | 13.54 | 0.98 | 4.22 |
| GC(avg)-LF | 7.42 | 15.18 | 0.30 | 2.56 |
| GC(min)-GC(avg) | 5.00 | 10.06 | 1.41 | 7.00 |
| CQ-LF | 41.08 | 42.68 | 13.33 | 9.12 |
| GC(min)-CQ-LF | 15.78 | 29.42 | 3.62 | 5.56 |
| GC(avg)-CQ-LF | 10.56 | 19.43 | 1.26 | 3.24 |
| GC(min)-GC(avg)-CQ | 6.29 | 12.58 | 0.15 | 5.79 |
| GC(min)-GC(avg)-LF | 7.37 | 13.99 | **0.00** | **2.53** |
| GC(min)-GC(avg)-CQ-LF | 9.40 | 15.91 | 0.15 | 3.35 |

it also discussed that the test recording required the enough SNR in order to obtain the high performance of GCC-based methods. It means that the situation of low SNRs is difficult to detect spoofing attacks as well for CQCC and LFCC-based methods. In contrast, in the case using SONY-C for test recording, the fused systems GC(min)-GC(avg)-LF yielded the lowest EERs than the single GCC(avg) as same as the results of DB1. In this case, the SNRs were almost the same as those of DB1.

From these results, when the quality of testing microphone is high and the situation obtained a suitable SNR can be prepared, the score fusion system can achieve high performance without situation dependence. Since the conditions are prepared by the developers who want to protect systems, the proposed systems can be regarded as a realistic technique.

## V. CONCLUSION

We proposed the score fusion of the spatial and the spectral features-based methods. Recently, as the spatial-based methods, the GCC-based replay attack detection method has been proposed. While the GCC-based methods have been reported to achieve high performance under several situations, the methods still have the situation dependency. Since the acoustic features extracted the different characteristics from the GCC-based methods, it expects that score fusion of the spatial and spectral feature-based systems can compensate each other and improve the robustness. From the experimental results, it was confirmed that the proposed methods achieved lowest EERs in all situations when the enough SNRs are available.

In future work, the proposed method will be combined with the other systems which obtained good results in ASVspoof 2019. Additionally, investigations for other situations and evaluation test with large amount of data will be performed.

## ACKNOWLEDGMENT

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

## REFERENCES

[1] T. Kinnunen, Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," *In Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4401–4404, 2012.

[2] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," 2015.

[3] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilci, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, "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.

[4] "ASVspoof 2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," http://www.asvspoof.org/asvspoof2019/asvspoof2019_evaluation_plan.pdf.

[5] L.W. Chen, W. Guo, and L.R. Dai, "Speaker verification against synthetic speech," *In Proc. 7th International Symposium on Chinese Spoken Language Processing*, pp. 309–312, 2010.

[6] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks.," *In Proc. INTERSPEECH*, pp. 82–86, 2017.

[7] Z. Chen, Z. Xie, W. Zhang, and X. Xu, "Resnet and model fusion for automatic spoofing detection," *In Proc. Interspeech*, pp. 102–106, 2017.

[8] S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui, "Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification," *In Proc. INTERSPEECH*, pp. 239–243, 2015.

[9] L. Zhang, S. Tan, J. Yang, and Y. Chen, "Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones," *In Proc. The 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1080–1091, 2016.

[10] R. Yaguchi, S. Shiota, N. Ono, and H. Kiya, "Replay attack detection using generalized cross-correlation of stereo signal," *In EURASIP European Signal Processing Confernce*, 2019. (accepted).

[11] R. Yaguchi, S. Shiota, N. Ono, and H. Kiya, "Spoofing detection method using generalized cross-correlation between multiple channels for speaker verification," *In The Acoustical Society of Japan*, 2018.

[12] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients," *In Proc. Odyssey, Bilbao, Spain*, vol. 25, pp. 249–252, 2016.

[13] M. Sahidullah, T. Kinnunen, and C. Hanilci, "A comparison of features for synthetic speech detection," *In Proc. Imterspeech, Annual Conf. of the Int. Speech Comm. Assoc.*, pp. 2087–2091, 2015.