# Replay Attack Detection Using Generalized Cross-Correlation of Stereo Signal

Ryoya Yaguchi, Sayaka Shiota, Nobutaka Ono and Hitoshi Kiya
Tokyo Metropolitan University, 6-6, Asahigaoka, Hino, Tokyo, JAPAN

***Abstract*—In this paper, we propose a replay attack detection method using the generalized cross-correlation (GCC) of a stereo signal for automatic speaker verification. In particular, this method focuses on a specific replay attack characteristics when speech is not active. In a genuine speaker case, when speech is not active, the maximum value of GCC is low since surrounding noise arrives from any direction. In contrast, in a replay attack case, even when the played speech is not active, the maximum value of GCC is high since recorded noise or electromagnetic noise is played by a loudspeaker for replay attack. Based on this assumption, two approaches of replay attack detection are introduced. One is to use the minimum value of GCC in short pauses. The other one is to use the average value of GCC in silent periods before the start point and after the end point of a target utterance. In experiments, it is confirmed that the proposed methods achieve low error rates without environmental restrictions.**

***Index Terms*—automatic speaker verification, spoofing countermeasure, generalized cross correlation, replay attack detection**

## 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. Since ASV systems have high affinity with voice interface systems, it is easy to combine ASV and voice interface systems [1], [2]. On the other hand, it has been reported that spoofing attacks (i.e., speech synthesis and replay) have become a serious problem for ASV systems [3]–[5]. In response, the ASV Spoofing and Countermeasures (ASVspoof) challenge was held in 2015 [6] and 2017 [7]. Since all samples in the ASVspoof2017 database were recorded by a singlechannel microphone, many methods assuming a singlechannel have been proposed [8]–[10]. To improve robustness against replay attacks, voice liveness detection (VLD) has also been established as a fundamental means of detecting replay attacks (Fig. 1). The concept of the VLD framework is to identify whether an input sample originates from a genuine speaker or a loudspeaker. Several methods using two channels of signals have been proposed for realizing the VLD framework [11], [12]. These VLD approaches focused on the characteristics appearing only for genuine utterances. The technique in [12] focuses on the fact that different phonemes are emitted from different locations within the human vocal tract system. In particular, the method uses the difference in the sound source position for each phoneme by calculating the time difference of arrival (TDOA) between two microphones [13]. While TDOA-based VLD obtained high performance, it can only be used in constrained situations with particularly high sampling rates such as 192 kHz.

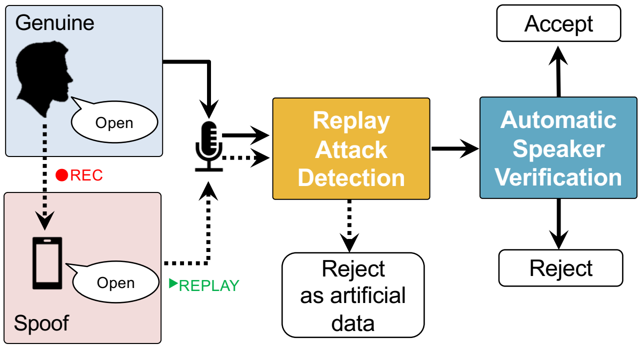

Fig. 1. System flow

The TDOA-based approach motivated us to consider other methods based on loudspeaker-specific characteristics. The proposed methods focuse on detecting phenomena peculiar to replay attacks rather than detecting those peculiar to genuine speech. In the case of genuine speech, no sound is emitted from the speaker in no-voice activity (no-VA) periods such as before and after the speech and during short pauses. On the other hand, in the case of prerecorded speech, there are some recorded background and electromagnetic noises even the during no-VA periods. In experiments, the proposed methods achieved higher performances than that of constant Q cepstral coefficients (CQCC) used for the baseline system of ASVspoof2017.

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

## II. Related Work

As a method for VLD, a sound-source-localization-based technique using TDOA has already been proposed [12]. In this technique, it is supposed that different phonemes are emitted from different locations within the human vocal tract system [14]. Figure 2 shows an example of TDOA for phonemes [s] and [u], where $\tau$ represents the arrival time of each phoneme from the speaker to each microphone. The

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

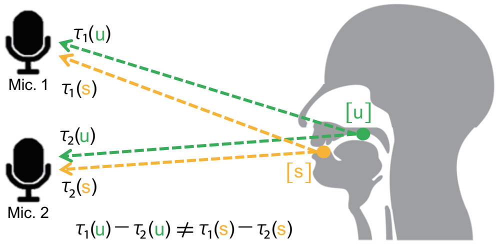

Fig. 2. Example of TDOA for phonemes [s] and [u]

TDOA is defined by the difference between $\tau_1$ and $\tau_2$ as follows.

$$\Delta t' = \tau_1 - \tau_2. \tag{1}$$

In this example, the TDOA of [s] is larger than that of [u]. This means that location where [u] is articulated is closer to the midpoint between the two microphones than that for [s].

Suppose that the detection system has multichannel microphones such as recent smartphones. Let $r_1(t, f)$ and $r_2(t, f)$ be zero-mean signals captured by two microphones. Then, the generalized cross-correlation (GCC) between them can be calculated as Fig. 2 [15].

$$\phi_g(\tau, d; t) = \frac{1}{L} \sum_f \frac{r_1^*(t, f) r_2(t + d, f)}{|r_1^*(t, f) r_2(t + d, f)|} e^{j2\pi f \tau / L}, \tag{2}$$

where $t = [1, ..., T]$ and $f$ are the frame and the frequency index, respectively, and $d$ is a time delay. $\tau$ is the time difference and $L$ is the frame length. The TDOA $\Delta t$ of a stereo signal can be estimated as follows:

$$\Delta t = \arg \max_d \phi_g(\tau, d; t). \tag{3}$$

If the phoneme-dependent TDOA is well trained in advance, the system can classify whether an input signal is from a genuine speaker or a loudspeaker by the TDOA and recognized phonemes. Note that this method is based on very accurate TDOA estimation, which is not easy in a real environment. Also, it is not robust to speaker movement.

## III. Proposed Methods

### A. Characteristics for loudspeakers during no VA periods

The TDOA-based approach motivated us to consider other methods based on loudspeaker-specific characteristics. In a genuine-speaker case, the maximum GCC is considered to be low in the no-VA periods 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 no-VA periods, the maximum GCC can be high.

These characteristics can be explained by the following observation models. The signals recorded by two microphones $a$ and $b$ for a genuine speaker can be represented as follows in the time-frequency domain:

$$M_a(t, f) = H_a(f) S(t, f) + N_a(t, f), \tag{4}$$
$$M_b(t, f) = H_b(f) S(t, f) + N_b(t, f), \tag{5}$$

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 no-VA periods, the source signal $S(t; f)$ is equal to 0. Thus, the observed signals in no-VA periods include only the background noise as follows:

$$M_a(t, f) = N_a(t, f), \tag{6}$$
$$M_b(t, f) = N_b(t, f). \tag{7}$$

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

$$M_p(t, f) = H_p(f) S(t, f) + N_p(t, f), \tag{8}$$

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

$$M_a(t, f) = H_a'(f)(M_p(t, f) + N_s(t, f)) + N_a(t, f), \tag{9}$$
$$M_b(t, f) = H_b'(f)(M_p(t, f) + N_s(t, f)) + N_b(t, f), \tag{10}$$

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 no-VA periods, $S(t, f) = 0$ yields $M_p(t, f) = N_p(t, f)$. Then, Eqs. (9) and (10) can be rewritten as

$$M_a(t, f) = H_a'(f)(N_p(t, f) + N_s(t, f)) + N_a(t, f), \tag{11}$$
$$M_b(t, f) = H_b'(f)(N_p(t, f) + N_s(t, f)) + N_b(t, f). \tag{12}$$

The equations mean that even in no-VA periods, the recorded noise $N_p(t, f)$ and the electromagnetic noise $N_s(t, f)$ are still omitted. 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.

To confirm the trends in “genuine and replayed utterances”, two utterances are investigated and the results are plotted in Fig. 3. Figures 3(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 no-VA periods. 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 3(c) shows the GCC of one frame in both a VA period and a no-VA period for the genuine and the replayed utterances. The red dots denote the maximum point in each frame. In the VA period, the peak of both utterances had a high value. In the no-VA periods, 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 no-VA periods

This paper focuses on the trajectories of the maximum GCC values in no-VA periods for spoofing detection. The maximum GCC for each frame is defined as

$$\phi_{max}(t) = \max_d \phi_g(\tau, d; t). \tag{13}$$

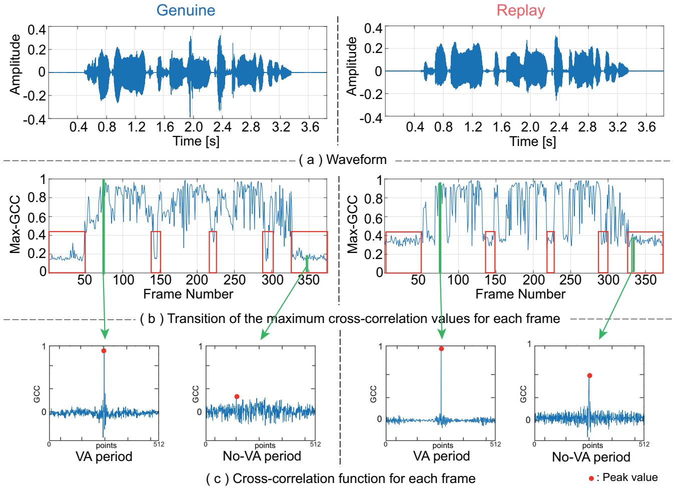

Fig. 3. Trajectories of the maximum GCC for each frame

As shown in Fig. 3(b), there are two types of no-VA periods: "short pauses" appearing in an utterance and "silent periods" before the start point and after the end point of speech signals. When short pauses are used for detection, the minimum value of GCC in its trajectory is selected as a detection score. For the silent periods, the average GCC value is calculated. These definitions are expressed as:

$$\Phi_{min} = \mathop{\arg\min}\limits_{t_s \le t \le t_e} \phi_{max}(t), \qquad (14)$$

$$\Phi_{avg} = \frac{1}{K} \sum_{T_s \le t < t_s, t_e < t \le T_e} \phi_{max}(t). \qquad (15)$$

where $t_s$ and $t_e$ are the start and end points of an utterance, respectively, and $K$ is the total number of frames $t$. Parameters $T_s$ and $T_e$ represent the start and end points used to calculate the average GCC value in silent periods, respectively. The value of these parameters can be selected arbitrarily under the constraints $1 < T_s < t_s, t_e < T_e < T$, where a parameter $T$ represents the end point of recording signal. The aim of using the minimum GCC value during short pauses is to capture step declines which are assumed to only occur for genuine speech. On the other hand, the aim of using the average GCC value in silent periods is to capture more stable characteristics that only occur for replayed signals.

## IV. Experiments

To evaluate the performance of the proposed methods, some experiments on replay attack detection were carried out.

### A. Database

In the experiments, there were two types of testing flow, as shown in Fig. 4. A different one was used in each testing flow.

DB1 was used for preliminary experiments in various conditions. For database 1 (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. The distance from the microphones to both the speaker and the loudspeaker was about 10 cm. 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 no perceptible electromagnetic noise in silent periods of replayed attacks. The CI is a separate stereo loudspeaker. It is 99 mm wide, 131 mm deep and 221 mm high for each side. 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 to record the detection test.

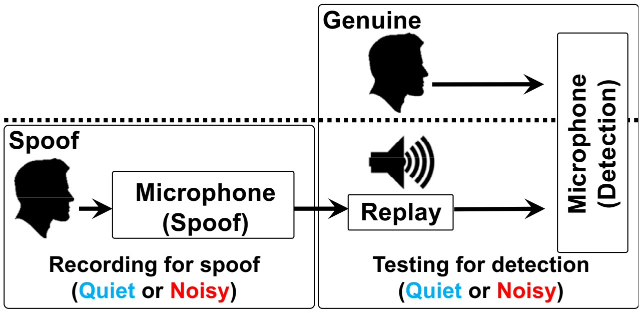

Fig. 4. Spoofing process and testing flow

DB2 was used as real spoofing situations. For database 2 (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. The distance from the microphones to the speaker was about 1 m and the distance from the microphones to the loudspeakers was about 10 cm. 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.

For spoof recording and detection tests, a quiet environment (Quiet) and an environment with background noise (Noisy) were assumed. "Quiet" was a common space with no extra background noise such as an air conditioner. "Noisy" was a room with stationary sound such as an air conditioner running on low and non-stationary sound such as a TV program playing at a moderate volume. Considering the environments for spoof recording and test recording, the following four environmental combinations were used.

(A) Noisy-Quiet: Spoof recording in a noisy environment and test recording in a quiet environment.

(B) Noisy-Noisy: Both spoof recording and test recording in a noisy environment.

(C) Quiet-Quiet: Both spoof recording and test recording in a quiet environment.

(D) Quiet-Noisy: Spoof recording in a quiet environment and test recording in a noisy environment.

All of the data in DB1 were recorded in these four environments. For DB2, spoof recording was performed in a quiet or noisy environment, and the detection test was recorded in a noisy environment. For all environmental combinations of DB1, the average signal-to-noise ratio (SNR) was set to about 18 dB. In the case of using the TMG for the spoof recording and test recording in DB2, the SNR was set to about 8 dB. In the case of using the SONY-C for the spoof recording and using the TMG for test recording, the SNR was set to about 21 dB. In the case of using the TMG for spoof recording and using the SONY-C for test recording, the SNR was set to about 10 dB. In the case of using the SONY-C for spoof recording and for test recording, the SNR was set to about 17 dB. 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 baseline system in these experiments, we used the constant Q cepstral coefficient (CQCC) [16], [17] for the acoustic features and Gaussian mixture models (GMMs) for the classifier [18]. The manner of use of the baseline system was the same as that defined in ASVspoof2017 [19]. Full details of the CQCC extraction were reported in [16]. To train the GMMs, we used 900 sentences for genuine utterances and 900 sentences for replayed speeches from a VLD database [11]. All of the VLD database was recorded through the AKGs and the spoof speeches were replayed by a BOSE 111AD loudspeaksers. The training conditions of the GMMs were the same as those of the baseline system of ASVspoof2017.

For the proposed approach, two methods were used. One was named GCC(min), which was defined in Eq. (14). The minimum value of the GCC during short pauses was used for the detection score. The other one was GCC(avg), which was defined in Eq. (15). The average value of the GCC during the no-VA periods was used for detection score. In all experiments on the proposed methods, hand-labeled data for the start point $t_s$ and end point $t_e$ of each utterance was used. Since GCC(min) required the detection of the periods of short pauses in an utterance, energy-based VA detection was used. For GCC(avg), two no-VA periods in an utterance were used to calculate the average of GCC; one was before the start point and the other was after the endpoint, and the averaging time was 0.5 s in both periods in the experiments. For the proposed methods, the frame length was set to 256 points for 16 kHz sampled speech and 1024 points for 48 kHz sampled speech. The evaluation was based on the equal error rate (EER).

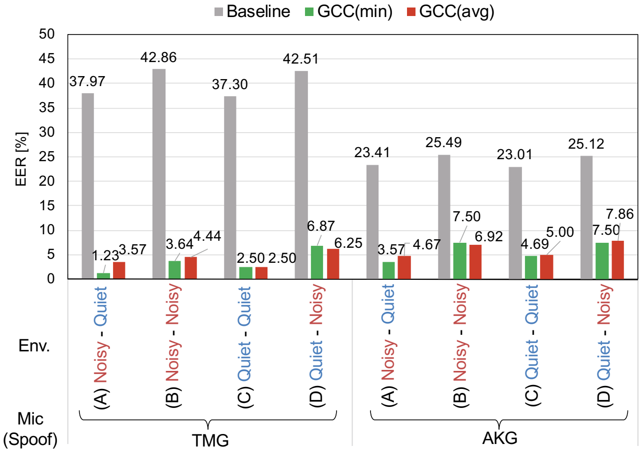

Fig. 5. EERs for DB1

### C. Results

Figure 5 shows the EERs of each spoofing detection method for DB1. Comparing the baseline system with the proposed methods, the EERs of the baseline were worse than those of the proposed methods, GCC(min) and GCC(avg), under every environmental combination. The mismatch of the frequency characteristics between the microphones used for spoofing and testing caused the deterioration of the baseline performance. The baseline was vulnerable against unknown recording conditions. For the environmental combinations (A) and (C), the proposed methods achieved high performance when using the TMG and AKG microphones for spoof recording. For the environmental combinations (B) and (D), the proposed methods produced a slightly higher error rate, meaning that the testing in a quiet environment yields a more accurate performance. Since the proposed methods focused on capturing some noise in no-VA periods, it is reasonable that the performance of the proposed methods depended on the testing conditions. This indicated that in the case of spoof recording in a noisy environment, the performance of the proposed methods were better than that in a quiet environment for spoof recording, which was for the same reason as above.

Figures 6 and 7 show the EERs of each spoofing detection method for DB2. The difference from Fig. 5 was that two types of microphones were used for testing, the TMG and the SONY-C. The results had a similar tendency to those for DB1. The EERs of the baseline were higher than almost all the EERs of the proposed methods. However, the EERs of the proposed methods had different trends. In particular, the EERs of GCC(min) using the TMG for spoof recording produced a higher error rate. From the database descriptions, there is some relationship between the EERs and the SNRs. When the SNR was under 10 dB, the performance of the proposed methods were unstable i.e., the proposed methods achieved

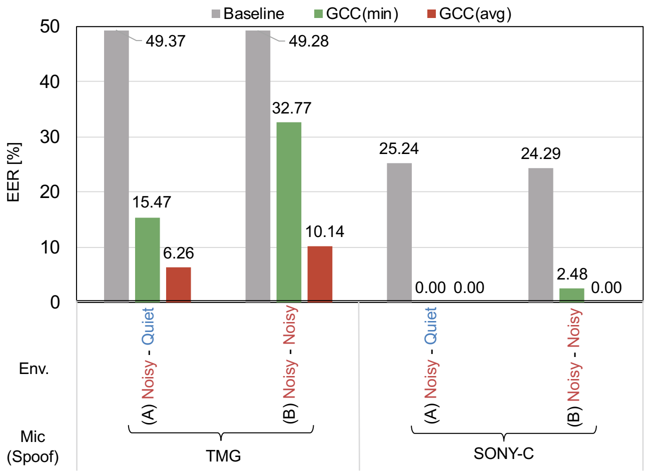

Fig. 6. EERs for DB2 (test recording by the TMG)

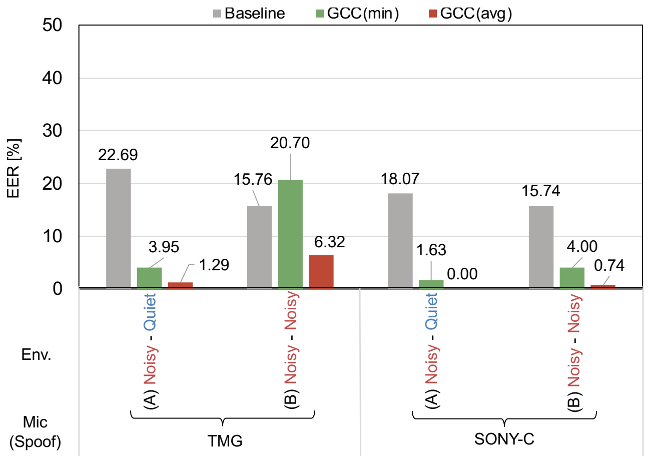

Fig. 7. EERs for DB2 (test recording by the SONY-C)

high performance in all tested conditions with a sufficient SNR.

The difference between DB1 and DB2 was the distance from the microphones to the speaker. The results for DB2 indicated that in real situations, such as when using a smart speaker, the proposed methods will work well.

## V. Conclusion

We proposed replay attack detection using the GCC function of a stereo signal. Replay attacks are regarded as a serious problem for ASV systems, and it was become important to consider countermeasures against spoofing. The proposed methods focused on using a stereo signal and detected specific characteristics of a replayed signal in no-VA periods. From the experimental results, it was confirmed that the proposed methods achieved low error rates without environmental restrictions.

As future work, the performance of the proposed methods will be investigated in other environments. The proposed methods will also be combined with other spoofing countermeasures.

## Acknowledgment

This work was supported in part by JSPS KAKENHI Early-Career Scientists Grant number JP19K20271 and Grant-in-Aid for Scientific Research(A) 16H01735 from the Japan Society for the Promotion of Science.

## References

[1] Saypay technologies. http://saypaytechnologies.com/.

[2] Voicekey mobile applications. http://speechpro-usa.com/product/voiceauthentication/voicekey#tab2.

[3] 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 *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4401–4404. IEEE, 2012.

[4] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, T. Leino, and A. M. Laukkanen. I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry. In *Interspeech*, pages 930–934, 2013.

[5] J. Villalba and E. Lleida. Detecting replay attacks from far-field recordings on speaker verification systems. In *European Workshop on Biometrics and Identity Management*, pages 274–285. Springer, 2011.

[6] 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. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[7] 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*, 11(4):588–604, 2017.

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

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

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

[11] 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 *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[12] L. Zhang, S. Tan, J. Yang, and Y. Chen. Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1080–1091, 2016.

[13] J. Liu, Y. Wang, G. Kar, Y. Chen, J. Yang, and M. Gruteser. Snooping keystrokes with mm-level audio ranging on a single phone. In *In ACM MobiCom*, pages 142–154, 2015.

[14] Joseph P Olive, Alice Greenwood, and John Coleman. *Acoustics of American English speech: a dynamic approach.* Springer Science & Business Media, 1993.

[15] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320–327, 1976.

[16] M. Todisco, H. Delgado, and N. Evans. A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients. In *Speaker Odyssey Workshop, Bilbao, Spain*, volume 25, pages 249–252, 2016.

[17] J. Brown. Computer identification of musical instruments using pattern recognition with cepstral coefficients as features. *The Journal of the Acoustical Society of America*, 105(3):1933–1941, 1999.

[18] T. B. Patel and H. A. Patil. Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[19] X. Wang, Y. Xiao, and X. Zhu. Feature selection based on CQCCs for automatic speaker verification spoofing. In *Proc. INTERSPEECH*, pages 32–36, 2017.