Replay Attack Detection in Automatic Speaker Verification Using Gammatone Cepstral Coefficients and ResNet-Based Model

Auwat Chaiwongyen$^{1,2}$, Suradej Duangpummet$^3$, Jessada Karnjana$^3$, Waree Kongprawechnon$^2$ and Masashi Unoki$^1$

$^1$Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292 Japan
$^2$Sirindhorn International Institute of Technology, Thammasat University, Thailand
$^3$National Science and Technology Development Agency, Thailand
E-mail: anuwat, unoki}@jaist.ac.jp, {jessada.karnjana, suradej.duangpummet}@nectec.or.th, waree@siit.tu.ac.th

Abstract This paper presents a method for detecting replay attacks in an automatic speaker verification system. The replay attack is of interest because it is the most straightforward, effective attack, and difficult to detect. Even though many speech features and classifiers have been proposed, the detection performance, such as an equal error rate (EER), accuracy, and balanced accuracy, need to be improved. Therefore, we propose a method for replay attack detection that applies the Gammatone cepstral coefficients with a ResNet-based model. The proposed method was evaluated and compared with existing methods and baselines in the ASVspoof 2019 challenge. The results indicated that the proposed method outperforms our previous method and the baselines in which the EER was 8.4%. In addition, the accuracy and balanced accuracy of the spoofing detection were improved.

Keywords: replay attack, automatic speaker verification, spoofing countermeasure

1. Introduction

Automatic speaker verification (ASV) systems use voice to validate the authenticity of a claimed voice for security [1]. Nowadays, the vulnerability of such systems to spoofing attacks, including impersonation, speech synthesis, voice conversion, and replay, has become a severe problem. A replay attack, which uses recording and playback devices, is the most accessible because it requires the least skill, experience, and equipment. It is, however, difficult to detect. Hence, this paper focuses on replay attacks for ASV.

Countermeasures for preventing replay attacks have been developed to identify a spoofed signal before ASV. The standard techniques are based on artifact identification to distinguish between genuine and spoofed signals using features extracted from a speech signal with classifications.

Many algorithms and methods based on using different features and classifiers have been proposed [2, 3, 4, 5, 6, 7]. For methods based on a single feature, in the ASVspoof 2019 challenge, the baselines are methods using either constant-Q cepstral coefficients (CQCCs) or linear-frequency cepstral coefficients (LFCCs) as features with a Gaussian mixture model (GMM) as the classifier [2]. Todisco et al. showed CQCCs and Mel-frequency cepstral coefficients (MFCCs) as features with a deep neural network (DNN) as the classifier [3]. This work reported that MFCCs with a DNN outperformed the baseline (CQCC-GMM) and CQCCs with a DNN. The method proposed by et al. revealed that using LFCCs with a Siamese convolutional neural network (CNN) outperformed the same classifier (Siamese CNN) with CQCCs [4].

For methods based on multiple features and classifiers, Adiban et al. proposed a technique based on CQCCs with the combination of an autoencoder and Siamese CNN [5]. A subsequent method based on the constant-Q transform (CQT) and stacking of classifiers, called SE-Res2Net50, surpassed Adiban et al.’s method [6]. The feature fusion method with a ResNet-based model, called ResNeWt18, reported the best performance in the ASVspoof 2019 challenge [7]. This work proposed an ensemble feature, including Mel-filter banks, CQT, CQTgrame, and the modified
group delay function (MGD), the so-called CQTMGD. From the aforementioned works, although no single feature performs well with all classifiers, the classifiers based on the ResNet model showed promising performance with many features. We previously investigated the ResNet-based model with LFCC and MFCC features, and the MFCC ones showed impressive results [8]. While Gammatone cepstral coefficients (GTCCs) performed better than MFCCs in speech-audio classifications and in ASV [9, 10], applying GTCCs for replay attack detection is still unclear. Therefore, this paper proposes a scheme based on GTCCs with a ResNet-based model for replay attack detection in ASV.

2. Proposed Method

In this section, we briefly introduce GTCCs and present our proposed method that utilizes GTCCs with a ResNet-based model.

2.1 Gammatone cepstral coefficients

GTCC features are a biologically inspired modification of the MFCCs that uses Gammatone filters with equivalent rectangular bandwidth (ERB) bands. The log-magnitude response of the Gammatone filter can be used to represent the human auditory response. In comparison with MFCCs, the Gammatone filter provides more frequency components in the low-frequency range with a narrow bandwidth and fewer frequency components in the high-frequency range with a wider bandwidth, as shown in Fig. 1. Thus, it can reveal the spectral information effectively [11].

The framework for GTCC extraction is divided into five steps, as shown in Fig. 2. First, the audio signal is windowed into short frames between 10 and 50 ms, and each frame is weighted by a Hamming window. Then, a short-time Fourier transform (STFT) is applied to convert a signal in the time domain into the time-frequency domain. Next, the Gammatone filter bank comprising various Gammatone filters is used for the STFT of the audio signal, and the energy of each sub-band is computed. After that, the logarithm (log) of each sub-band is computed. Finally, the discrete cosine transform (DCT) is used to obtain the GTCC features.

2.2 Proposed framework

The proposed framework takes the GTCC feature from the input signal without applying voice activity detection. The signal is divided into 128 overlapping frames of 1600 samples. Each frame is weighted by a Hamming window. The 60 Gammatone filters in ERB, as proposed in [9], are used to cover the frequency from 10 Hz to 11 kHz. Hence, the number of dimensions of the feature is 128 × 60.

For the classifier, which is ResNet-based models, we firstly apply ResNet34, which is a general ResNet model. The second model is the modified multi-
branch ResNet [6], namely Deep ResNet. It is used in splitting and aggregating strategies to gain accuracy effectively while maintaining the complexity. A basic block of Deep ResNet is shown Fig. 3 (b).

We use the Adam optimizer with a learning rate of $10^{-3}$ and a batch size of 16. The training is finished after 50 epochs. The loss function is the sparse categorical cross-entropy with two classes, i.e., bona fide and spoof. The dropout rate is 50%. The square bracket specifies the shape of each residual block. There are two stacked blocks. C=32 denotes the division of the ResNet into 32 groups. The details of the Deep ResNet are shown in Table 1.

### 3. Evaluations

To evaluate the proposed method, we used a dataset from the ASVspoof 2019 challenge (physical access sub-challenge). The dataset was divided into the training, development, and evaluation sets in which they contain both genuine utterances and replay attack access attempts recorded and playback under different conditions [2].

In this work, we mainly and directly compared our methods (GTCCs using ResNet34 and Deep ResNet) with the baselines in the ASVspoof 2019 challenge, i.e., LFCC-GMM and CQCC-GMM [2]. The existing features, i.e., LFCC, CQCC, and MFCC with the ResNet-based classifier, were also evaluated. However, several state-of-the-art methods including multiple features or advanced classifications are also illustrated for further improvement. For example, Siamese CNN is a twin neural network classifier whereas LFCC-GMM and CQCC-GMM as well as the proposed method are single-feature and single-classifier methods.

Table 2 illustrates the equal error rate (EER) comparison of each method. The EER is the rate at which the false rejection rate (FRR) and false acceptance rate (FAR) are equal. As illustrated in Table 2, the EER of the proposed methods is lower than that of the method using CQCC with the Siamese CNN [4]. It is also lower than that of the two baselines, i.e., LFCC-GMM and CQCC-GMM [2]. Note that the best performance of attack replay detection of each criterion is illustrated by the bold text. In addition to the EER, we also evaluated the performance of the proposed method with other classification metrics, including accuracy, balanced accuracy, and F1 score. Table 3 shows the performance of the proposed model and baselines in terms of the true positives rate (TPR), true negatives rate (TNR), false positives rate (FPR), false negatives rate (FNR), accuracy, balanced accuracy, precision, recall, and F1 score.

The results in terms of accuracy, balanced accuracy, and F1 score indicate that the proposed method, GTCCs with Deep ResNet, is the best. However, the TNR, FPR, and recall of the CQCC-GMM [2] are still slightly better than those of the proposed method.

### 4. Discussion

The comparison of replay attack detection between GTCCs as other features with the same classifiers shows a better slightly performance in terms of accuracy, recall, and F1 score. In contrast, it moderately outperforms in terms of balanced accuracy.

In addition, the performance of GTCCs with ResNet34 shows that the TNR, FPR, accuracy, balanced accuracy, recall, and F1 score outperform the other features. However, the TPR, FNR, and precision are slightly lower than those of MFCCs. We also re-implemented the two baselines in the ASVspoof 2019 challenge. However, we obtained a slight difference in EERs, as shown in Table 2. the ResNeWt18 [7]. However, our model might differ from the ResNeWt18. It also has a massive number of parameters, i.e., 146M parameters.

### 5. Conclusion

We proposed a method for detecting replay attacks as a countermeasure in an ASV system. The proposed method utilized GTCCs, a single front-end feature, and Deep ResNet as the classifier. We evaluated the performance of the proposed method using the ASVspoof 2019, a physical access dataset. The result indicated that the proposed method outperformed the two baselines at which the EER was 8.48%. In addition, the classification metrics, i.e., TNR, FPR, recall,
### Table 3  Classification performance on the evaluation set in the ASVspoof 2019

|               | GMM (%) | Deep ResNet (%) | ResNet34 (%) |
|---------------|---------|-----------------|--------------|
|               | LFCC    | CQCC            | LFCC         | CQCC | MFCC | GTCC | LFCC | CQCC | MFCC | GTCC | LFCC | CQCC | MFCC | GTCC |
| **TPR**       | 99.83   | 68.55           | 93.14        | 94.14 | 98.31 | 96.05 | 96.92 | 88.79 | 98.46 | 98.40 | 96.05 | 88.79 | 98.46 | 98.40 |
| **TNR**       | 4.48    | **98.16**       | 49.35        | 28.10 | 51.58 | 81.06 | 24.34 | 36.92 | 45.43 | 62.71 | 24.34 | 36.92 | 45.43 | 62.71 |
| **FPR**       | 95.52   | **1.84**        | 50.65        | 71.90 | 48.42 | 18.94 | 75.66 | 63.08 | 54.57 | 37.29 | 75.66 | 63.08 | 54.57 | 37.29 |
| **FNR**       | 0.17    | 31.45           | 6.86         | 5.86  | 1.69  | 3.95  | 3.08  | 11.21 | 1.54  | 1.60  | 3.08  | 11.21 | 1.54  | 1.60  |
| **Accuracy**  | 87.02   | 72.53           | 87.26        | 85.27 | 92.03 | **94.03** | 87.18 | 81.83 | 91.34 | 93.61 | 87.18 | 81.83 | 91.34 | 93.61 |
| **Balanced**  | 52.15   | 83.35           | 71.24        | 61.12 | 74.94 | **88.55** | 60.63 | 62.86 | 71.98 | 80.55 | 60.63 | 62.86 | 71.98 | 80.55 |
| **Accuracy**  | **99.83** | 68.55           | 93.14        | 94.14 | 98.31 | 96.05 | 96.92 | 88.79 | 98.46 | 98.40 | 96.05 | 88.79 | 98.46 | 98.40 |
| **Recall**    | 87.08   | **99.59**       | 92.22        | 89.41 | 92.90 | 97.03 | 89.20 | 90.08 | 92.09 | 94.45 | 89.20 | 90.08 | 92.09 | 94.45 |
| **F1-score**  | 93.02   | 81.20           | 92.68        | 91.71 | 95.53 | **96.54** | 92.90 | 89.43 | 95.17 | 96.38 | 92.90 | 89.43 | 95.17 | 96.38 |

accuracy, balanced accuracy, and F1 score, showed that the proposed method outperformed the baselines. These results suggested that the GTCC feature could capture replay attack characteristics to enable ResNet-based classifiers to distinguish the spoofed signal effectively. However, the performance of the proposed method is still lower than that of the methods based on feature fusion with advanced classifiers. Thus, incorporating GTCCs into other speech features and utilizing more advanced and generalized classifiers will be investigated in the future.

### Acknowledgment

This work was supported by a grant from the SIIT-JAIST-NSTDA Dual Doctoral Degree Program. This work was also supported by a Grant-in-Aid for Scientific Fund for the Promotion of Joint International Research (Fostering Joint International Research (B)) (20KK0233) and KDDI foundation (Research Grant Program).

### References

[1] D. A. Reynolds: An overview of automatic speaker recognition technology, Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Vol. 4. pp. 4072–4075, 2002.

[2] R. K. Das, J. Yang and H. Li: Long range acoustic and deep features perspective on ASVspoof 2019, 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 1018-1025, 2019, doi: 10.1109/ASRU.2019.9003845.

[3] M. Todisco, W. Alhamdani and F. T. Sheldon: Replay spoof attack detection using deep neural networks for classification, 2020 International Conference on Computational Science and Computational Intelligence (CSCI), 2020.

[4] Z. Lei, Y. Yang, C. Liu and J. Ye: Siamese convolutional neural network using gaussian probability feature for spoofing speech detection, Proc. Interspeech, pp. 1116–1120, 2020.

[5] M. Adiban, H. Samiti and S. Shehnoor: Replay spoofing countermeasure using autoencoder and siamese networks on ASVspoof 2019 challenge, Audio and Speech Processing, Vol. 64, pp 101-105, 2020.

[6] X. Li, N. Li, C. Weng, X. Liu, D. Su, D. Yu, and H. Meng: Replay and synthetic speech detection with res2net architecture, CASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.

[7] X. Cheng, M. Xu and T. F. Zheng: A multi-branch ResNet with discriminative features for detection of replay speech signals, APSIPA Transactions on Signal and Information Processing, Vol. 9, 2020.

[8] A. Chaiwongyen, W. Kongprawechnon, S. Duangpummet, J. Karnjana and M. Unoki: Replay attack detection using MFCC and ResNet18 for automatic speaker verification. RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP), 2022.

[9] V. Xavier and F. Alias: Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification. IEEE Transactions on Multimedia, pp. 1684-1689, 2012.

[10] C. Himanshu, D. Sadhya and V. Patel: Automatic Speaker Verification using Gammatone Frequency Cepstral Coefficients. 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN), IEEE, 2021.

[11] K. M. Malik, A. Javed, H. Malik and A. Irtaza: A light-weight replay detection framework for voice controlled IoT devices, in IEEE Journal of Selected Topics in Signal Processing, Vol. 14, pp. 982-996, 2020, doi: 10.1109/JSTSP.2020.2999828.
D. A. Reynolds: An overview of automatic speaker recognition. In Proceedings of the 1999 International Conference on Acoustics, Speech, and Signal Processing. 2002.

Waree Kongprawechmon received his B.S. degree in Electrical Engineering (First Class Honors) from Chulalongkorn University, Thailand. She received her M.Eng. degree in Control Engineering from Osaka University and her Ph.D. degree in Engineering (Mathematical Engineering and Information Physics) from the University of Tokyo, Japan in 1995 and 1999, respectively. Since 1999, she has joined as a Lecturer at the School of Information, Computer, and Communication Technology (ICT), SIT, Thammasat University, Thailand. Currently, she is an Associate Professor. Her research interests and areas of expertise are theory in H∞ control, control theory, robust control, system identification, adaptive control, learning control, neural network, and fuzzy control. She was also on an advisory committee in Engineering Science between Thailand and Japan (NRCT-JSPS) for cooperating projects from 2005 to 2020.

Anuwat Chaiwongyen received his B.S. degree in Technical Education (Computer Engineering) from the Rajamangala University of Technology Lanna Chiangmai, Thailand in 2008. He received his M.Eng degree in Computer Engineering from Thammasat University, Thailand in 2017. He was a research assistant at National Electronics and Computer Technology (NECTEC) from 2008 to 2020. He is currently a Ph.D. candidate in Sirindhorn International Institute of Technology (SIIT), Thammasat University, Thailand and the Japan Advanced Institute of Science and Technology (JAIST), Japan. His research interests include audio spoofing detection, machine learning, and deep neural networks.

Suradej Duangpummet received his B.Eng. and M.Eng degrees in Mechatronics Engineering from Pathumwan Institute of Technology and Asian Institute of Technology, Thailand in 2001 and 2011, respectively. He received his Ph.D. degree in Information Science (outstanding performance award) from JAIST under the supervision of Prof. Masashi Unoki in 2021. He has worked as a research assistant at NECTEC since 2003. He is currently a researcher in the Biometrics Technology Research Team at NECTEC. His research interests and areas of expertise are room acoustic signal processing, system identification, blind parameter estimation, control theory, machine learning, and deep neural networks.

Jessada Karnjana received his B.Eng. degree in Electronics Engineering from King Mongkut’s Institute of Technology LadaKrabang, Thailand in 1999, M.Eng. degree Microelectronics from Asian Institute of Technology, Thailand in 2006, his Ph.D. degree in Information Science from JAIST, Japan in 2016, and his Ph.D. degree in Engineering and Technology from Sirindhorn International Institute of Technology, Thammasat University in 2017. His research interests include data analysis and machine learning, reasoning with uncertainty, signal processing, and wireless sensor networks. He has worked at NECTEC, Thailand since 1999.