Replay Attack Detection Based on Voice and Non-voice Sections for Speaker Verification

Ananda Garin Mills*†, Patthranit Kaewcharuay*†, Pannathorn Sathirasattayanon*†,
Suradej Duangpummet†, Kasorn Galajit†, Jessada Karnjana†, and Pakinee Aimmanee*
* Sirindhorn International Institute of Technology, Thammasat University, Pathumthani, Thailand
E-mail:{6422771756, 6422782241, 6422782316}@g.siit.tu.ac.th, pakinee@siit.tu.ac.th
† NECTEC, National Science and Technology Development Agency, Pathumthani, Thailand
E-mail:{suradej.duangpummet, kasorn.galajit, jessada.karnjana}@nectec.or.th

***Abstract*—Voice can represent a person's identity. Thus, it can be used in automatic speaker verification (ASV) systems for authenticating secure applications. Unfortunately, existing ASV systems are vulnerable to spoofing attacks. A replay attack is a widely used spoofing technique because it is simple but difficult to detect. Hence, many methods are proposed for countermeasures against replay attacks. Most work inseparably considers voice and non-voice sections in the detection's performance. In this work, we investigate the spoof detection performances when the voice, non-voice, and both with different percentages of voice are used to obtain the optimal section. We also propose a method for detecting replay attacks using the optimal section of a signal. Mel-frequency cepstral coefficients are calculated from the optimal section as a feature, and the ResNet-34 model is used for classification. We evaluated the proposed method using a dataset from the ASVspoof 2019 challenge. The results depict that the optimal section for replay attack detection is when 10% and 20% of voice are included in the non-voice sections. It also showed that the proposed method outperforms the baselines with a 7.52% relatively improvement or an equal error rate of 1.72%.**

I. Introduction

Speech and voice can represent our identity as biometrics for authenticating secure systems using automatic speaker verification (ASV) systems. At the same time, an ASV is currently vulnerable to spoofing attacks in which someone disguises as another and illegitimately accesses a secure system. Hence, countermeasures against spoofing attacks are necessary to verify whether the claimed voice is a genuine or fake representation. Attackers might merely replay someone's voice to ASV, called replay attacks [1]. Other spoofing techniques, e.g., speech synthesis and voice conversion, might need advanced algorithms and expertise, whereas a replay attack is simple, using only a recorder and playback device. However, detection of replay attacks is challenging since current methods still have accuracy issues. Therefore, this paper focuses on detecting replay attacks.

In the past, many methods have been proposed to counter replay attacks by using different speech features or applying a variety of classifiers, such as the Gaussian mixture model (GMM), deep neural networks (DNN), and convolutional neural networks (CNN) [2–18]. Although most of the methods unitized the whole utterance for feature extractions, it was reported that non-voiced segments contain vital information from recorders and playback devices [10–15]. Eliminating non-voice segments also decrease the classification performance or causes the over-fitting problem [11, 17, 18].

A few methods thus exploited these clues for replay attack detection. For example, a method proposed by Saranya *et al.*, utilized only non-voiced sections [11]. Three GMM classifiers were trained from three feature extractions. The final decision was the voting from scores of these three models. Chettri *et al.* showed that weights of the CNN model are heavily present in the first and last 400 ms of the utterances [13]. Recently, Wang *et al.* proposed a method utilizing both voice and non-voice segments and fusing scores of the two models [16].

However, there is no research investigating the contributions of the proportional voice and non-voiced sections for replay attack detection. Therefore, this paper investigates the effects of using non-voice sections added with percentages of voice sections to obtain the so-called *optimal section*. Mel-frequency cepstral coefficients (MFCC) are calculated from these optimal sections as a feature. Then, ResNet-34 [19] is used as a classifier to detect whether the extracted feature of the optimal section is genuine or a replay attack.

The rest of this paper is organized as follows. Section II presents the investigation of voice and non-voice sections on replay attacks. In Section III, the proposed scheme based on the optimal section from the signal is described. Sections IV is experiments and evaluations. We discuss the finding in Section V. Lastly, Section VI is the conclusion of this paper.

II. Voice and Non-voice Analysis

We investigate the clues in non-voice sections of spoofed signals, i.e., the difference between voice and non-voice sections. Thus, the utterance is segmented to be voice and non-voice sections using voice activity detection (VAD).

Since there are many VAD algorithms, in this study, we use an algorithm based on energy and spectral spread analysis [20]. The detection of either voice or non-voice is done at a frame level. The energy of each frame is used for calculating a histogram. The decision threshold is then calculated from

978-616-590-477-3 ©2022 APSIPA

the histogram of all frames. The threshold T is determined as follows [20]:

$$T = \frac{W \times M_1 + M_2}{W + 1}, \quad (1)$$

where M_1 and M_2 are the first and second local-maxima positions, respectively, and W is a ratio of the weight of M_1 to that of M_2. W is set to 5 by default in a MATLAB function used in our experiment, affecting the duration of voice and non-voice sections. In our proposed method, the voice section duration is intentionally varied to investigate how it affects classification performance. Therefore, the choice of W alters only the percentage of voice duration used in our simulation without significantly changing the conclusion.

From preliminary experiments, we found that the quality of replay equipment makes it difficult for spoof detection since they seem to be identical, as shown in Fig. 1. The voice sections from almost the same conditions, including the same speaker, sentence, and environment, but one is genuine, and another is a replay attack.

On the other hand, a non-voice section shows meaningful clues that should be easier for distinguishing spoof signals from genuine ones. Examples of non-voice sections from the same utterance above are shown in Fig. 2.

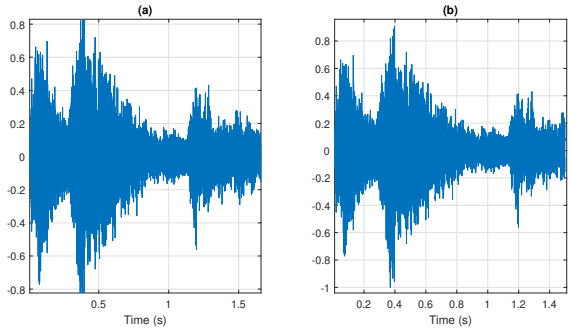

Fig. 1: Voice sections: (a) genuine and (b) spoof.

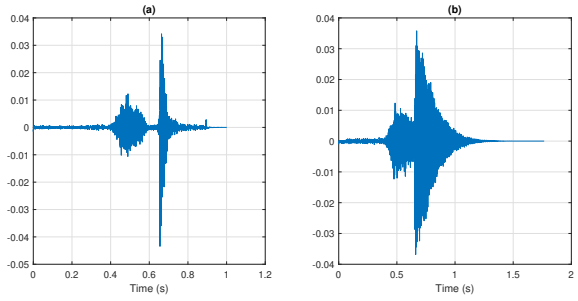

Fig. 2: Non-voice sections at the beginning of utterances: (a) genuine and (b) spoof.

Therefore, the clues of the replay process, including an attacker's microphone and loudspeaker, are presented in the early section of the voice. We then further investigate the classification performance of each section and find the optimal section to be a feature used in a replay attack scheme.

TABLE I: Dimensions of the MFCC calculated from the voice and non-voiced sections.

	whole utterance	voice	non-voice
Dimensions	60×723	60×448	60×610

III. Proposed Method

The proposed method concludes with four steps, as shown in Fig. 3. The sub-processes include VAD, optimal section selection, feature extraction, and classification model. Note that the optimal section is obtained from the experiments.

A. Optimal voice and non-voice ratio

From the input signal, four types of sound sections are identified to obtain the optimal section, including the whole utterance, voice only, non-voice only, and non-voice with percentages of voice. For voice only and non-voice only, VAD is done on the audio to get boundaries of silence. For non-voice with percentages of voice, we add a non-voice section with a percentage of the following the voice part. The percentage of the voice part is calculated from its boundary.

Fig. 4 illustrates the start and end boundaries of each section using the VAD. The red area is the silence or non-voice section, and the green areas are the additive of the voice part. The optimal section is, therefore, the concatenation between red and green areas. The percentage of the voice section (green area) will be determined in the next section.

B. Feature

In this work, we chose mel-frequency cepstral coefficients (MFCC) to be a front-end feature since it takes less computational power compared to others, such as constant-Q cepstral coefficients (CQCC) [21]. MFCC is a speech feature based on the human auditory system. It has been used for many speech signal processing and speech recognition, as well as replay attack detection [11, 22]. The MFCC is calculated by applying a cosine transform of the real logarithm of the short-term energy spectrum [23]. This short-term energy spectrum is expressed on a mel-frequency scale using a mel-scale filterbank.

For obtaining the optimal section, the MFCC of the section is calculated that are 19 coefficients of MFCC, delta, delta-delta, as well as the energy. Since there are four types of the section, the dimensions of the features are shown in Table I. An example of the MFCC feature extracted from the different sections of utterance is shown in Fig. 6. These features are then used as an input for the classifier.

C. Classifier

Deep residual neural networks (*ResNets*) are used to classify the feature extracted from the optimal voice and non-voice sections and to discriminate between genuine and replay

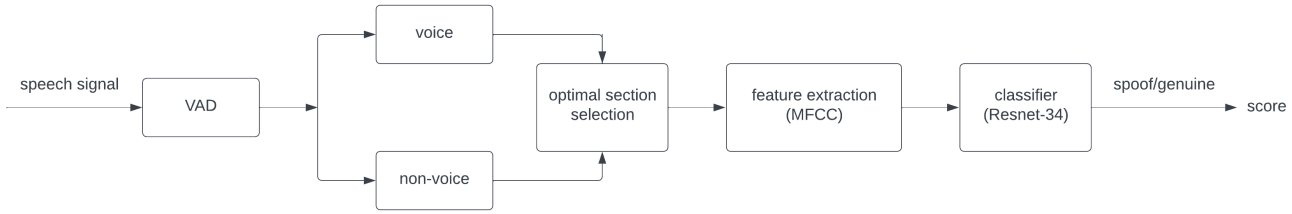

Fig. 3: Block diagram of the proposed method.

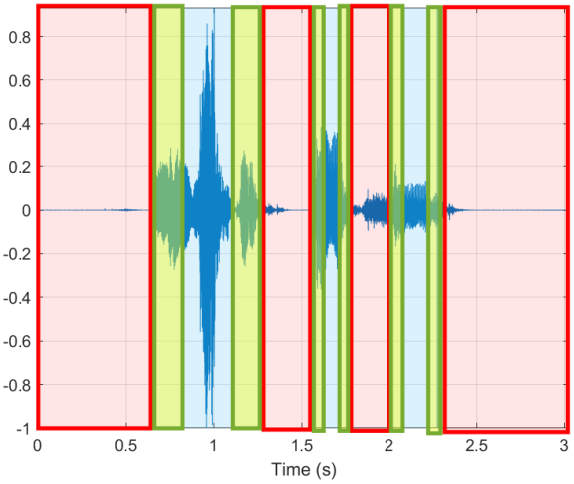

Fig. 4: Optimal sections for detecting replay attacks. Red boundaries indicate non-voice and silence. Green boundaries indicate extended regions from a percentage of the width of the voice region.

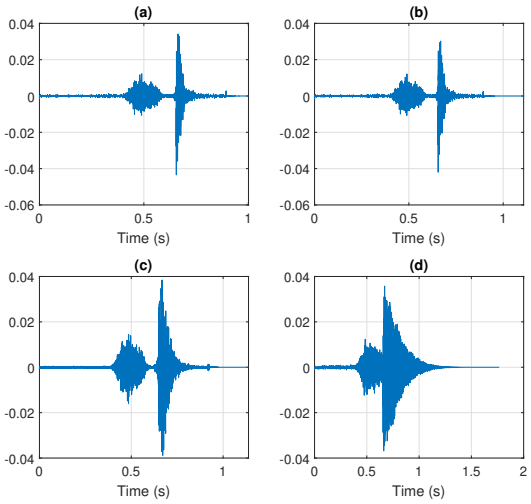

Fig. 5: Non-voice sections after applying VAD: genuine (a) and spoofs (b, c, d). The spoofs are from the same utterance but different quality recorders (from low to high) where the other configurations are the same, i.e., in a small room with a size of $2-5$ m^2, reverberation T60 of $200-500$ ms, and a talker-to-ASV distance of $10-50$ cm [16].

signals. ResNet, a deep convolutional neural network (CNN), was proposed for image classification [19]. Its effectiveness has been demonstrated in many researches not only in the area of image processing but also audio and speech signal processing, as well as spoofing-replay-attack detection [10, 12].

Training the ResNet is also easier compared to other models with a similar number of layers because of learning residual functions with reference to the input technique [19]. The leaning residual function is defined as:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x}, \tag{2}$$

where $\mathbf{x}$ and $\mathbf{y}$ are input and output vectors of a considering layer. $\mathcal{F}(\mathbf{x}, \{W_i\})$ stands for the residual mapping function. The square weight matrix, W_i, is optimized so that the difference between input and output (residual) is close to zero. This function can be represented by a block diagram as shown in Fig. 7. There are many modifications of ResNets, and the number of layers is also different (from 18 layers up to more than 150 layers). However, we mainly consider the effect of using voice and non-voice sections as a feature. Hence, we utilize *ResNet-34*, consisting of 34 convolutional layers, to be our classifier in this study.

IV. Experiments and Evaluations

TABLE II: Performance comparison between parts of voice used to create a model. The optimal model is the smallest EER. The voice and non-voice sections are split by applying voice activity detection. Note that MFCC and ResNet-34 are the features and classifiers of all conditions.

	EER (%)		Accuracy (%)
	Dev	Eval	Eval
Whole utterance	**1.72**	**1.86**	**98.90**
Voice only	32.04	31.45	68.55
Non-voice only	2.05	2.90	97.10

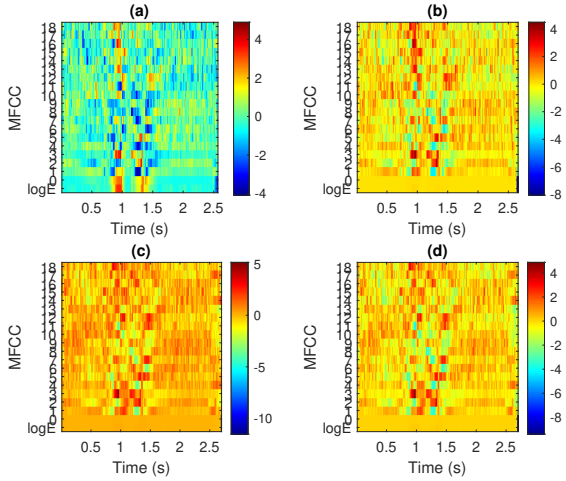

Fig. 6: MFCC of non-voice sections: (a) bonafide signal and (b,c,d) spoofed signal where the spoofed signals are from different quality of the recorders (from low to high) [16]. Note that these signals are under the same configurations, including room size of $2-5$ m^2, T60 of $500-200$ ms, and a talker-to-ASV distance of $10-50$ cm.

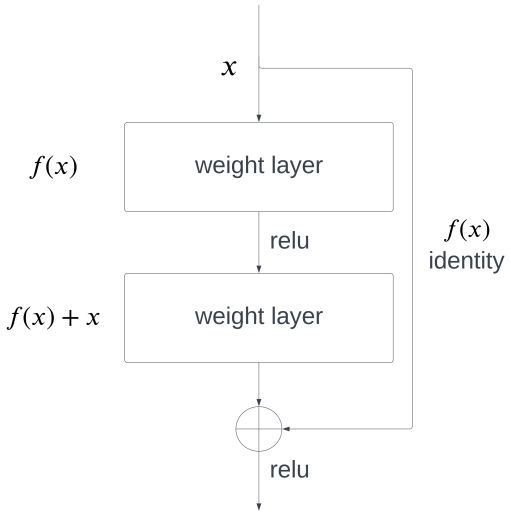

Fig. 7: Residual block in the ResNet.

A. Dataset

The dataset is from the ASVSpoof 2019: replay spoofing attacks in a physical access (PA) scenario [16]. This dataset includes $54,000$ audio files for training, $29,700$ files for development, and $135,000$ files for evaluation. The utterances were produced by 20 speakers, including 12 women and eight men for training, four males and six females for development. In contrast, a larger evaluation set consists of 21 for males and 27 for females. The range of the signals is between 1.45 and 10.31 seconds. We cut the longest signal to 8.23 seconds to make consistent dimensions between the training and evaluation sets. All files have a sampling rate of 16 kHz. The details of configurations in the dataset include room size, reverberation, replay device quality, and talker—attacking distance variables, including attacker-to-ASV distance. Note that this dataset is imbalanced.

B. Implementations

The segments of a signal were justified by the VAD algorithms [20] in MATLAB 2020a[1]. The MFCC was then calculated as a feature from each signal (whole utterance, voice, non-voice, and percentage of voice in the non-voice section). The ResNet-34 models[2] and their training were implemented on GoogleColab Pro+ in Python 3.7 with associated libraries, e.g., TensorFlow 2.1, Scikit-learn, and Librosa. We also trained the models from the training set and evaluated their performance using all trials from the development and evaluation sets.

C. Evaluation metrics

The evaluation metrics are an equal error rate (EER), which is a value where false accept rate is equal to false reject rate. The EER is commonly used as a prime metric of anti-spoofing [1]. Apart from the EER, the other classification metrics are also used, including accuracy, recall, and F-score.

D. Results

Table II shows the comparison results of three configurations: using the whole signal, using only the voice section, and using the only non-voice section. The results from a model using the only voice section are worse in all sets. It seems that the voice-only model suffers from over-fitting. In contrast, results show extraordinary performance from using the non-voice section and full signal. However, the model using the full signal still outperforms the non-voice model.

We carried out further experiments to observe the contribution of voice and non-voice combinations to detecting replay attacks. We used the model trained from the whole utterance as a baseline model. Then, the different combinations of voice and non-voice sections as input are fed to the model trained from the whole signal.

Table III shows the analysis results of using the non-voice section with different percentages of voice section. Interestingly, the model taking input features from the non-voice + 10% and 20% voice yields the outstanding EERs on all sets. It also outperforms the regular method using the full signal. Furthermore, the non-voice + 20% voice shows the best results for all classification metrics (accuracy, F-scores (0.5, 1, 2), and recall).

V. Discussion

Even though we can distinguish spoofs more precisely using non-voiced regions, some remaining issues and limitations should be discussed as follows.

[1]https://www.mathworks.com/help/audio/ref/detectspeech.html
[2]https://pypi.org/project/image-classifiers/

TABLE III: Analysis of voice and non-voice sections for feature extraction and classification on replay attack detection. Note that MFCC and ResNet-34 are feature and classifier of all conditions. The optimal section is the smallest EER.

Feature (MFCC)	EER (%)		Accuracy (%)	F0.5 (%)	F1 (%)	F2 (%)	Precision (%)	Recall (%)
	Dev	Eval	Eval					
Whole utterance	1.72	1.86	98.14	99.39	98.91	98.45	99.71	98.14
Non voice + 5% voice	1.63	1.89	98.11	99.38	98.90	98.42	99.70	98.10
Non voice + 10% voice	**1.43**	1.76	98.24	99.42	98.98	98.54	99.72	98.24
Non voice + 20% voice	1.53	**1.72**	**98.28**	**99.44**	**99.00**	**98.56**	**99.73**	**98.28**
Non voice + 30% voice	1.62	1.77	98.22	99.42	98.97	98.52	99.72	98.22
Non voice + 40% voice	1.71	1.77	98.23	99.42	98.97	98.52	99.72	98.22

Firstly, spoofing recorded by a *perfect* quality recorder is still difficult to distinguish. Fig. 5 shows an example of spoof attacks using a perfect recorder. We can see that the signals are almost identical. Such cases might cause the remaining errors in the proposed method.

Secondly, there are other forms of spoofing attacks, e.g., replacing non-voiced sections with genuine noises or adding silence and removing silences [10]. Such spoofing techniques might pose a threat to the proposed method. Nevertheless, we use some voice information from extended non-voiced boundaries instead of relying on only non-voiced regions. Hence, the proposed method should be robust against such spoofing techniques. However, we will improve our model by using continuity detection algorithms in future works. Consequently, replacing spoof noises with genuine noise will cause a discontinuity in the signal.

Thirdly, this study shows that instead of using the whole utterance for processing in the spoof detection scheme, only the important sections can gain more effectiveness. The results indicate that the optimal section combines the non-voice section and $10-20\%$ of voice. However, the precise portion of voice regions is still unclear. Thus, finding the more precise optimal voice percentage might decrease the EER, which can be improvised and improved further.

VI. Conclusion

This paper proposed a method for detecting replay attacks using the optimal sections of a speech signal. We investigated the difference between spoof and genuine on non-voice sections in both time domain and cepstral analysis. We then conducted the experiments using various combinations of voice and non-voice sections to find the optimal section. In this study, the spoof detection scheme took MFCC as a feature, and ResNet-34 was a classifier. The result suggests that the non-voice section contains essential information regarding the record and playback devices, and the combination of the non-voice and voice section from 10 to 20% is the optimal section. We incorporated the optimal section into the spoofing detection scheme. The proposed method yielded a 7.52% relative reduction in equal error rate compared to the baseline using the whole utterances. In future work, we will investigate whether or not our method can be applied to detect other types of spoofing attacks, such as voice conversion and deepfake speech.

References

[1] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, 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.

[2] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Galka, "Audio replay attack detection using high-frequency features." in *Interspeech*, 2017, pp. 27–31.

[3] P. Nagarsheth, E. Khoury, K. Patil, and M. Garland, "Replay attack detection using DNN for channel discrimination." in *Interspeech*, 2017, pp. 97–101.

[4] T. Gunendradasan, B. Wickramasinghe, P. N. Le, E. Ambikairajah, and J. Epps, "Detection of replay-spoofing attacks using frequency modulation features." in *Interspeech*, 2018, pp. 636–640.

[5] C.-I. Lai, A. Abad, K. Richmond, J. Yamagishi, N. Dehak, and S. King, "Attentive filtering networks for audio replay attack detection," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6316–6320.

[6] Y. Ren, Z. Fang, D. Liu, and C. Chen, "Replay attack detection based on distortion by loudspeaker for voice authentication," *Multimedia Tools and Applications*, vol. 78, no. 7, pp. 8383–8396, 2019.

[7] Y. Ye, L. Lao, D. Yan, and L. Lin, "Detection of replay attack based on normalized constant q cepstral feature," in *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*. IEEE, 2019, pp. 407–411.

[8] B. Chettri, E. Benetos, and B. L. Sturm, "Dataset artefacts in anti-spoofing systems: a case study on the asvspoof 2017 benchmark," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 3018–3028, 2020.

[9] Y. Wang, Y. Liu, P. Gao, and Y. Wang, "An experimental study on replay attack detection using spoofing clues from both voiced and non-voiced segments," in *2021 5th International Conference on Digital Signal Processing*, 2021, pp. 266–271.

[10] 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.

[11] M. Saranya, R. Padmanabhan, and H. A. Murthy, "Replay attack detection in speaker verification using non-voiced segments and decision level feature switching," in *2018 international conference on signal processing and communications (SPCOM)*. IEEE, 2018, pp. 332–336.

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

[13] B. Chettri, S. Mishra, B. L. Sturm, and E. Benetos, "Analysing the predictions of a CNN-based replay spoofing detection system," in *2018 IEEE spoken language technology workshop (SLT)*. IEEE, 2018, pp. 92–97.

[14] J. Alam and P. Kenny, "Spoofing detection employing infinite impulse response—constant q transform-based feature representations," in *2017 25Th european signal processing conference (EUSIPCO)*. IEEE, 2017, pp. 101–105.

[15] M. Hajipour, M. A. Akhaee, and R. Toosi, "Listening to sounds of silence for audio replay attack detection," in *2021 7th International Conference on Signal Processing and Intelligent Systems (ICSPIS)*. IEEE, 2021, pp. 1–6.

[16] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, pp. 101–114, 2020.

[17] Y. Zhang, W. Wang, and P. Zhang, "The Effect of Silence and Dual-Band Fusion in Anti-Spoofing System," in *Proc. Interspeech 2021*, 2021, pp. 4279–4283.

[18] R. Font, J. M. Espín, and M. J. Cano, "Experimental analysis of features for replay attack detection-results on the ASVspoof 2017 challenge." in *Interspeech*, 2017, pp. 7–11.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] T. Giannakopoulos, "A method for silence removal and segmentation of speech signals, implemented in MATLAB," *University of Athens, Athens*, vol. 2, 2009.

[21] M. Todisco, H. Delgado, and N. Evans, "Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[22] M. Singh and D. Pati, "Usefulness of linear prediction residual for replay attack detection," *AEU-International Journal of Electronics and Communications*, vol. 110, p. 152837, 2019.

[23] V. Tiwari, "MFCC and its applications in speaker recognition," *International journal on emerging technologies*, vol. 1, no. 1, pp. 19–22, 2010.