2023 18th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP) | 979-8-3503-7121-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/iSAI-NLP60301.2023.10354625

Spoof Detection using Voice Contribution on LFCC features and ResNet-34

Khaing Zar Mon
Sirindhorn International Institute of Technology,
Thammasat University
Pathumthani, Thailand
m6522040556@g.siit.tu.ac.th

Kasorn Galajit
NECTEC, National Science and Technology Development Agency,
Pathum Thani, Thailand
kasorn.galajit@nectec.or.th

Candy Olivia Mawalim
Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa 923–1292 Japan
candylim@jaist.ac.jp

Jessada Karnjana
NECTEC, National Science and Technology Development Agency,
Pathum Thani, Thailand
jessada.karnjana@nectec.or.th

Tsuyoshi Isshiki
Department of Information and Communication Engineering,
School of Engineering, Tokyo Institute of Technology
Tokyo, Japan
isshiki@ict.e.titech.ac.jp

Pakinee Aimmanee
Sirindhorn International Institute of Technology,
Thammasat University
Pathumthani, Thailand
pakinee@siit.tu.ac.th

***Abstract*—Biometric authentication, especially in speaker verification, has seen significant advancements recently. Despite these significant strides, compelling evidence highlights the ongoing vulnerability to spoofing attacks, requiring specialized countermeasures to detect various attack types. This paper specifically focuses on detecting replay, speech synthesis, and voice conversion attacks. In our spoof detection strategy, we employed linear frequency cepstral coefficients (LFCC) for front-end feature extraction and ResNet-34 for distinguishing between genuine and fake speech. By integrating LFCC with ResNet-34, we evaluated the proposed method using the ASVspoof 2019 dataset, PA (Physical Access), and LA (Logical Access). In our approach, we contrast using the entire utterance for feature extraction in both PA and LA datasets with an alternative method that extracts features from a specific percentage of the voice segment within the utterance for classification. In addition, we conducted a comprehensive evaluation by comparing our proposed method with the established baseline techniques, LFCC-GMM and CQCC-GMM. The proposed method demonstrates promising performance with an equal error rate (EER) of 3.11% and 3.49% for replay attacks (PA) in the development and evaluation datasets. For voice conversion and speech synthesis attacks (LA), the method achieves EERs of 0.16% in the development dataset and 6.89% in the evaluation dataset. The proposed method shows promising results in identifying spoof attacks for both PA and LA attacks.**

***Index Terms*—replay attack, speech synthesis, voice conversion, LFCC, ResNet-34, ASVspoof**

I. Introduction

Automatic speaker verification (ASV) uses voice features to authenticate identities by comparing a person's voice to a pre-enrolled template. ASV has drawn much attention due to its potential use in several fields, including access control, forensic analysis, and voice-based authentication systems. With advancements in technology and the increasing use of speaker verification, the biometric system is encountering numerous attacks aimed at circumventing its security measures [1].

Spoofing attacks deceive ASV systems with pre-recorded, synthesized, or modified voice samples using techniques like speech synthesis, voice conversion, and other sophisticated methods [1]. For example, replay attacks use pre-recorded speech, speech synthesis attacks use text-to-speech synthesis methods, and voice conversion attacks change the speaker's vocal qualities [2] [3]. Voice impersonation uses accurate matching the target speaker's manner, pitch, and other perceptible signals, creating the appearance of an original voice by the target speaker [4]. It is challenging to identify each of these attacks using a specialized set of signal processing techniques, feature extraction methods, and classification algorithms [1]. ASV systems are vulnerable to spoofing, necessitating specialized detection methods employing advanced signal processing, machine learning, and deep learning to identify fake voice signals through speech feature analysis [1] [2] [5].

Spoof detection is crucial for ASV systems to detect and eliminate spoofed speech. This paper is a part of the ASEAN IVO 2023 project, "Spoof Detection for Automatic Speaker Verification," which aims to enhance the security and reliability of speaker verification by effectively detecting spoofing attacks.

979-8-3503-7121-5/23/$31.00 ©2023 IEEE

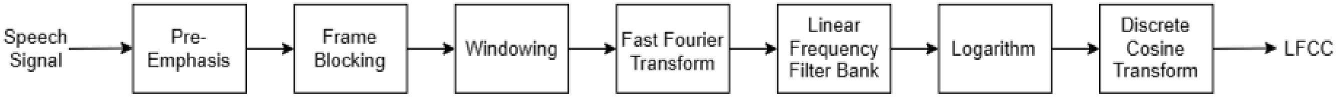

Fig. 1. LFCC block diagram.

Mills et al. proposed a method for detecting replay attacks by examining various combinations of voice and non-voice sections. They used the ResNet-34 model for classification and mel-frequency cepstral coefficients as features. They did the experiment based on the ASVspoof 2019: replay spoofing attacks in a physical access (PA) scenario. Their results of experiments run on ASVspoof 2019 challenge dataset showed that adding 10% and 20% voice to the non-voice regions produces the best performance. With a relative improvement of 7.52% and an error rate of 1.72%, the suggested technique shows notable improvements over the baselines [5].

According to Rosello et al., conformers performed well in the classification task of ASV anti-spoofing. The ASVspoof 2019 logical access database was used to assess the proposed system. The conformer encoder yielded promising results but did not surpass state-of-the-art anti-spoofing systems. The incorporation of phase features and time modeling enabled the gated recurrent convolutional neural network (GRCNN) to achieve as low as 3.85% EER and 0.0952 in tandem detection cost function (t-DCF) [6].

Previous studies have individually focused on detecting PA and LA attacks [5] [6]. In this article, we want to develop a unified approach for detecting all these attack types. To accomplish this, we chose LFCC as the front-end feature extraction method and ResNet-34 as the classification model.

LFCC was experimentally proven to be superior for female speech than male speech, as it effectively captures the spectral characteristics in the high formant frequency region [7]. Thus, when the dataset contains more female speakers than male speakers, it is the most suitable choice. By focusing on linear frequency resolution, it successfully captured the placements and shapes of spectral peaks in the synthesized speeches. It has shown promising results for detecting physical access (PA) and logical access (LA) attacks [8] [9] [10] [11]. Furthermore, it has demonstrated superior performance compared to other acoustic features, particularly for detecting unknown forgery types of attacks [12]. This observation served as motivation for us to integrate LFCC with the ResNet-34 model. ResNet was proven to be a powerful and effective tool for the detection of genuine and spoof speech [13]. Its effectiveness has been demonstrated in various research areas, including image processing, audio, speech signal processing, and spoof detection [14]. Through experimental evaluations on the ASVspoof2017 dataset, ResNet outperformed all other single-model systems, demonstrating its superiority in accurately detecting spoofed audio samples [13].

In this study, we aimed to propose an approach that integrates LFCC and ResNet-34 for spoof detection. Additionally, we conducted experiments to explore the potential benefits of employing different voice percentages to enhance the model's accuracy and efficacy in detecting replay attacks, speech synthesis and voice conversion attacks. To the best of our knowledge, previous studies have not explored the impact of incorporating varying percentages of voice to enhance the model's performance through the utilization of LFCC and ResNet-34 for spoof detection.

The rest of this paper is structured as follows: section II provides background information on LFCC and ResNet-34 architecture. In section III, we provide the details of the proposed scheme and discuss the methodology for utilizing LFCC features and ResNet-34 in our framework. We describe the experimental setup, including the database used for training and evaluation, the chosen evaluation metrics, and the presentation of results in section IV. In section V, we thoroughly examine and discuss the findings obtained from our experiments. Finally, in section VI, we present our conclusions, summarizing the key findings and highlighting the contributions of our research.

II. Background

In this section, basic background on two key components utilized in our study: LFCC and the ResNet-34 architecture, are provided.

A. Linear Frequency Cepstral Coefficients (LFCC)

To accurately detect replay attacks, voice conversion, and speech synthesis, it is crucial to use a robust audio feature extraction method. The LFCC feature extraction process, depicted in Fig. 1, comprises seven sequential steps:

(1) The input speech signal undergoes pre-emphasis, ensuring spectral flattening and reduced susceptibility to finite precision effects.

(2) The signal is divided into numerous frames.

(3) A Hamming window weighting is applied to each frame.

(4) A discrete Fourier transform (DFT) is executed on each weighted frame, generating short-time spectra.

(5) These spectra are then utilized to compute energies in the sub-bands of a linear filterbank.

(6) The base-10 logarithm of these energies results in a log-power spectrum.

(7) The log-power spectrum undergoes a discrete cosine transform (DCT), yielding the desired LFCC representation [8].

This process efficiently extracts LFCC features crucial for accurate spoof detection in automatic speaker verification systems.

B. Residual Network with 34 Layers (ResNet-34)

ResNet-34 refers to a specific variant of the Residual Neural Network (ResNet) architecture that consists of 34 layers. ResNet addresses the vanishing gradients problem by introducing shortcut connections with identity functions, facilitating training of deeper networks [13], [15]. The shortcut connection can be visualized using a block diagram, as depicted in Fig. 2.

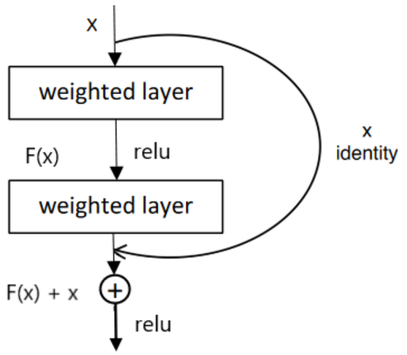

Fig. 2. "Basic-block" building block for ResNet.

The concept of residual learning is defined by the equation:

$$y = F(x, \{W_i\}) + x. \tag{1}$$

where x represents the input vector, y represents the output vector, and $F(x, \{W_i\})$ represents the residual mapping function. The goal of residual learning is to optimize the square weight matrix, W_i, so that the difference between the input and output (residual) approaches zero [15].

III. Proposed Method

In the spoof detection process, both feature extraction and classification play crucial roles. The first step involves extracting LFCC features from the audio files, which are then utilized as input for the subsequent classification stage. In the process of LFCC feature extraction, we captured the desired portion of the speech signal by varying percentages of the voice segment from the utterances. In selecting a specific segment from the utterance, we have to identify the part to include. Spoofed speech displays distinct characteristics in its early sections, possibly revealing clues associated with replay processes involving an attacker's microphone and loudspeaker [5]. Thus, we have chosen to incorporate the initial silence part and the head of the voice segment. Moreover, we aim to capture diverse acoustic patterns and variations by focusing on non-consecutive segments. As a result, we selected the following peices. First is initial silence from the beginning of the signal to the start of the voice part, second is the head, a specific percentage of the voice's initial portion. Third is tail, a specific percentage of the voice's final portion of the utterance. To detect the voice segment, we use the voice activity detection (VAD) to identify speech boundaries within the utterance, determining the starting and ending points. Afterward, we concatenate the initial silence and a specified percentage of the head and tail sections of the voice segment. We illustrated the comparison between the original and initial silence, along with voice segments extracted from 15% of the head and 15% of the tail of the utterances, as depicted in Fig. 3. The two utterances, one genuine and one fake, are from the same speaker, sentence, and under nearly identical conditions. Extracting specific percentages of the utterances highlights the differences between genuine and fake more prominently.

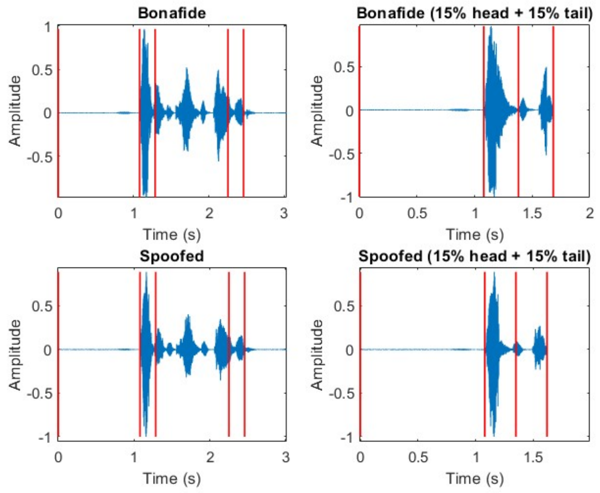

Fig. 3. On the left-hand side: original genuine and spoofed speech utterances. On the right-hand side: extracted voice segment, consisting of initial silence, 15% of the head and 15% of the tail from the original utterances.

For speech spoofing detection, the ResNet34 architecture is utilized. Additionally, the input dimension of the classifier is adjusted to optimize performance in the experiments, achieving an optimal size that enhances the overall results. In this case, the optimal input dimension for the PA dataset is set to 57 × 600, while for the LA dataset, it is adjusted to 57 × 746. The block diagram of our proposed method is illustrated in Fig. 4.

IV. Experimental and Results

A. Dataset and Experimental Setup

In our experiments, we utilized the well-known ASVspoof 2019 datasets. The ASVspoof 2019 challenge targets three major attack types: synthetic speech (SS), voice conversion (VC), and replay attacks. The dataset was divided into two subsets: logical access (LA), including spoof speech samples generated using text-to-speech (TTS) and voice conversion techniques, and physical access (PA), containing replayed speech recordings. Both subsets are further divided into training, development, and evaluation subsets. Within the LA dataset, spoofed speech signals were generated through the utilization of two voice conversion algorithms and four speech synthesis algorithms during the training and development subsets. Moreover, for the evaluation set, the spoofed data was generated by employing seven speech synthesis algorithms and six voice conversion spoofing algorithms [16]. Table I provides

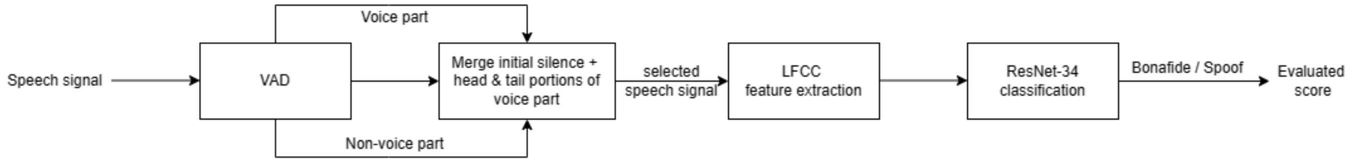

Fig. 4. Block diagram of the proposed method.

detailed information about the ASVspoof 2019 dataset, which was utilized in our study.

TABLE I
A SUMMARY OF ASVSPOOF CHALLENGE 2019 DATABASE

Dataset	Subset	Speaker		Utterance	
		Male	***Female***	***Bonafide***	***Spoof***
ASVspoof 2019 Logical Access	Train	8	12	2,580	22,800
	Development	8	12	2,548	22,296
	Evaluatioin	30	37	7,355	63,882
ASVspoof 2019 Physical Access	Train	8	12	5,400	48,600
	Development	8	12	5,400	24,300
	Evaluatioin	30	37	18,090	116,640

LFCC features were extracted with a 30 ms window and a 15 ms shift, employing a 1024-point Fourier transform and 70 filters. The resulting features consisted of 19 static cepstra along with energy, delta, and delta-delta coefficients. This comprehensive approach effectively captures frequencies ranging from 0 Hz to 8 kHz. Additionally, we applied voice detection to selectively extract LFCC features from a specific percentage of the voice segment.

For speech spoofing detection, we used the ResNet-34 architecture. Our model is trained using a combination of genuine and spoofed speech samples obtained from the ASVspoof 2019 training set. The training process employs the Adam optimizer with a learning rate of 0.0001. To optimize the model, we utilize sparse categorical cross-entropy as the loss function. During training, the model was trained for 50 epochs, with a batch size of 16. To evaluate the model's effectiveness, the development and evaluation sets, which contain previously unseen genuine and spoofed speech samples, were utilized. Evaluation metrics such as accuracy, equal error rate (EER), F1 score, and detection cost function (DCF) are employed to assess the model's performance in detecting speech spoofing.

B. Evaluation Metric

In our study, we assess the effectiveness of our experiments using four key evaluation metrics: equal error rate (EER), accuracy, detection cost function (DCF), and the F1 score. These metrics are calculated based on the counts of correctly identified spoofed speech samples and instances where genuine speech was incorrectly classified as spoofed.

The Equal Error Rate (EER) is a biometric security metric that signifies the point of balance where the system achieves equal probabilities of falsely accepting a non-matching sample and falsely rejecting a matching one [17], [18].

Accuracy is a measure of how accurately a system can classify or detect genuine and spoofed speech samples in a spoof detection system. It quantifies the percentage of correctly classified samples out of the total samples in the dataset.

The detection cost function (DCF) is defined based on the costs associated with misses (failure to detect the target) and false alarms (incorrectly detecting the target), along with the prior probability of the target speaker hypothesis [19].

The F1 score is the harmonic mean of precision and recall. Precision assesses the correctly classified positive samples among the predicted positives, while recall measures the correctly classified positive samples among the actual positives [20]. The F1 score offers a balanced model performance evaluation, especially in cases with class imbalance.

C. Result

We present the experimental results for the EER, accuracy, DCF, and F1 score based on the ASVspoof 2019 physical access (PA) and logical access (LA) datasets in Table II and Table III. The term "Mix Utterance" describes a combination of the initial silence and a specified percentage of both the head and tail of a voice segment, denoted as (H% head and T% tail). The experiments involved extracting LFCC features from both the entire audio file and specific percentages of voice segments. By varying the percentage, we can investigate the impact on the model's performance and effectiveness of different feature extraction strategies in detecting speech spoofing.

In Table II, we present the experimental results on the ASVspoof 2019 PA dataset. For the development dataset, the lowest EER is 3.11% with 97.22% accuracy, a DCF value of 2.78%, and an F1 score of 97.19%. These results were obtained using the initial silence part, along with 15% of voice from both the head and tail of the speech segment. For the evaluation dataset, an EER of 3.49%, an accuracy of 96.51%, DCF value of 3.51%, and F1 score of 96.50% were achieved by using the initial silence part, along with 5% of voice from the head and 5% of voice from the tail of the speech segment. Considering the entire utterance, the EER for the development dataset is 4.09%, and for the evaluation dataset, it is 4.37%. The accuracy achieved using the entire utterance is 96.32% for the development dataset and 95.35% for the evaluation dataset. Additionally, the DCF and F1 scores were 3.5%, 96.45% and 4.27%, 95.72% for the development and evaluation datasets, respectively.

TABLE II
EXPERIMENTAL RESULTS: ASVSPOOF 2019 **PA DATASETS** WITH WHOLE UTTERANCE AND VARYING PERCENTAGES OF EXTRACTED VOICE SEGMENTS

Feature (LFCC)			**EER (%)**		**Accuracy (%)**		**DCF (%)**		**F1 (%)**	
			dev	***eval***	***dev***	***eval***	***dev***	***eval***	***dev***	***eval***
Whole Utterance			4.09	4.37	96.32	95.35	3.50	4.27	96.45	95.72
Mix Utterance	***H***	***T***								
	5%	5%	3.58	***3.49***	96.32	***96.51***	3.68	***3.51***	96.29	***96.50***
	10%	10%	4.63	4.54	95.37	95.46	4.63	4.54	95.23	95.40
	15%	15%	***3.11***	3.78	***97.22***	96.32	***2.78***	3.68	***97.19***	96.31
	20%	20%	3.15	3.96	96.81	95.89	3.19	4.11	96.78	95.91
	25%	25%	3.93	3.51	96.69	96.49	3.31	3.52	96.62	96.45
	30%	30%	3.29	3.73	97.04	96.27	2.96	3.73	96.98	96.23
	35%	35%	4.73	3.96	95.39	96.10	4.61	3.90	95.42	96.11
	40%	40%	4.71	4.34	95.47	95.25	4.53	4.75	95.37	95.17

TABLE III
EXPERIMENTAL RESULTS: ASVSPOOF 2019 **LA DATASETS** WITH WHOLE UTTERANCE AND VARYING PERCENTAGES OF EXTRACTED VOICE SEGMENTS

Feature (LFCC)			**EER (%)**		**Accuracy (%)**		**DCF (%)**		**F1 (%)**	
			dev	***eval***	***dev***	***eval***	***dev***	***eval***	***dev***	***eval***
Whole Utterance			0.31	7.13	95.88	89.73	0.43	10.27	95.48	88.59
Mix Utterance	***H***	***T***								
	5%	5%	0.34	7.58	99.67	91.84	0.33	8.16	99.67	91.15
	10%	10%	0.35	7.32	95.68	92.30	0.43	7.70	95.50	91.70
	15%	15%	0.29	7.09	99.69	***93.11***	0.31	***6.89***	99.69	***92.64***
	20%	20%	0.43	10.55	99.24	91.70	0.76	8.30	99.24	91.05
	25%	25%	0.19	7.73	***99.81***	91.72	***0.19***	8.28	***99.82***	91.01
	30%	30%	0.25	7.83	95.64	91.89	0.44	8.11	95.45	91.22
	35%	35%	0.19	7.64	***99.81***	92.25	***0.19***	7.75	***99.82***	91.64
	40%	40%	***0.16***	***6.89***	99.79	91.84	0.21	8.16	99.79	91.15

We evaluated our experiments on the ASVspoof 2019 LA dataset, and the results are presented in Table III. By using the initial silence part and including 40% of the voice from both the head and tail of the speech segment, we achieved the lowest EER of 0.16% and 6.89% for the development and evaluation datasets, respectively. Our model demonstrated an impressive 99.81% accuracy, a DCF value of 0.19%, and an F1 score of 99.82% for the development dataset when considering "initial silence + (25% head + 25% tail) of voice" and "initial silence + (35% head + 35% tail) of voice" approaches. For the evaluation dataset, the accuracy achieved was 93.11%, with the lowest DCF value of 6.89%, and the highest F1 score of 92.64%, utilizing the initial silence part and 15% of the voice from the head and tail of the speech segment. In the development dataset, the entire utterance shows an EER of 0.31% with 95.88% accuracy. In contrast, the evaluation dataset has an EER of 7.13% and 89.73% accuracy. The DCF and F1 scores for development are 0.43% and 95.48%, while for evaluation, they are 10.27% and 88.59%, respectively.

The results from the development datasets of the LA dataset show better performance compared to the evaluation dataset. This disparity can be attributed to the composition of the datasets. The training and development data incorporate known attacks, which enable the model to learn from these specific attack types. In contrast, the evaluation data includes unknown attacks and two known attacks, which the model has not been explicitly exposed to during training. As a result, the model's performance may be relatively lower on the evaluation dataset, as it encounters attack types that it has not been trained extensively on.

V. DISCUSSIONS

Utilizing 15% of the voice segment demonstrates strong performance on the development dataset of the PA dataset, as indicated by EER, accuracy, DCF, and F1 score metrics. In the evaluation dataset, employing 5% of the voice segment results in low EER, high accuracy, low DCF value, and high F1 score. Although the use of the percentage strategy may not significantly improve performance for the PA dataset, it offers noteworthy benefits. Selectively utilizing specific speech segments optimizes computational resources and memory requirements, enhancing model efficiency. Thus, despite limited performance gains, the trade-off between improved computational efficiency and reduced memory demands remains favorable.

Analyzing the EER measure for the LA dataset in both the development and evaluation datasets revealed that using 40% of the speech data yielded the best outcomes. Furthermore, examining accuracy, DCF, and F1 scores showed that 35% and 25% of the speech data in the development dataset provided optimal results, while using only 15% of the speech data delivered the highest accuracy, DCF, and F1 scores for the evaluation dataset. These findings highlight the positive impact of extracting specific percentages of voice segments on model performance in the LA dataset. Selectively focusing on non-consecutive portions of voice segments at different percentages enables the capture of essential features for distinguishing between genuine and spoofed signals. These distinctions form the basis for informed countermeasure decisions, resulting in improved spoofed speech detection.

TABLE IV
EER(%) COMPARISON WITH LA AND PA OF ASVSPOOF 2019

Dataset	Baseline System	dev set	eval set
Logical Access	LFCC-GMM	2.71	8.09
	CQCC-GMM	0.43	9.57
	proposed method	***0.16***	***6.89***
Physical Access	LFCC-GMM	11.96	13.54
	CQCC-GMM	9.87	11.04
	proposed method	***3.11***	***3.49***

In our evaluation, we performed a comparative analysis between our proposed methods and two baseline countermeasures, LFCC-GMM and CQCC-GMM. The comparison was based on the EER performance metric, and the results are presented in Table IV. Across both the PA and LA datasets, our method exhibited superior performance compared to both baseline approaches.

VI. CONCLUSION

In this study, we introduced a spoof detection experiment utilizing LFCC as the front-end feature and the ResNet-34 model for classification. Our investigation included a comparison of performance between extracting features from the entire audio file and extracting a specific percentage of the voice segments. These features were then fed into the same architecture classifier, and their performance was evaluated. By utilizing the ASVspoof 2019 PA and LA datasets, we assessed the effectiveness of our proposed LFCC and ResNet-34 model, which demonstrated strong performance. These datasets encompassed a diverse range of attack types, such as replay attacks, voice synthesis, and voice conversion attacks. Additionally, we conducted experiments by extracting features using different percentages of the voice segments. We observed that varying percentages produced different results, helping us identify the optimal percentage that outperforms using the entire utterance. It is crucial to note that the optimal percentage may vary depending on the datasets and specific attack types. In our ongoing research, we will primarily focus on feature fusion and classifier integration to enhance our ability to distinguish genuine from spoofed speech in future work.

VII. ACKNOWLEDGE

This research is financially supported by Thailand Advanced Institute of Science and Technology (TAIST), National Science and Technology Development Agency (NSTDA), Tokyo Institute of Technology, Sirindhorn International Institute of Technology (SIIT), Thammasat University (TU) under the TAIST Tokyo Tech Program. The authors would also like to express their gratitude to the National Research Council of Thailand (NRCT) for providing scholarships to support this study. The ASEAN IVO (https://www.nict.go.jp/en/asean_ivo/index.html) project, spoof detection for automatic speaker verification, was involved in the production of the contents of this work and financially supported by NICT (http://www.nict.go.jp/en/index.html).

REFERENCES

[1] C. B. Tan, M. H. A. Hijazi, N. Khamis, P. N. E. B. Nohuddin, Z. Zainol, F. Coenen, and A. Gani, "A survey on presentation attack detection for automatic speaker verification systems: State-of-the-art, taxonomy, issues and future direction," *Multimedia Tools and Applications*, vol. 80, no. 21-23, pp. 32 725–32 762, 2021.

[2] J. Villalba and E. Lleida, "Detecting replay attacks from far-field recordings on speaker verification systems," in *Biometrics and ID Management: COST 2101 European Workshop, BioID 2011, Brandenburg (Havel), Germany, March 8-10, 2011. Proceedings 3*. Springer, 2011, pp. 274–285.

[3] M. Zhang, X. Wang, F. Fang, H. Li, and J. Yamagishi, "Joint training framework for text-to-speech and voice conversion using multi-source tacotron and wavenet," *arXiv preprint arXiv:1903.12389*, 2019.

[4] Y. Gao, R. Singh, and B. Raj, "Voice impersonation using generative adversarial networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2506–2510.

[5] A. G. Mills, P. Kaewcharuay, P. Sathirasattayanon, S. Duangpummet, K. Galajit, J. Karnjana, and P. Aimmanee, "Replay attack detection based on voice and non-voice sections for speaker verification," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 221–226.

[6] E. Roselló Casado, A. Gómez Alanís, M. Chica Villar, Á. M. Gómez García, J. A. González López, A. M. Peinado Herreros *et al.*, "On the application of conformers to logical access voice spoofing attack detection," 2022.

[7] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, and S. Shamma, "Linear versus mel frequency cepstral coefficients for speaker recognition," in *2011 IEEE workshop on automatic speech recognition & understanding*. IEEE, 2011, pp. 559–564.

[8] A. Chaiwongyen, K. Pinkeaw, W. Kongprawechnon, J. Karnjana, and M. Unoki, "Replay attack detection in automatic speaker verification based on resnewt18 with linear frequency cepstral coefficients," in *2021 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*. IEEE, 2021, pp. 1–5.

[9] S. Cui, B. Huang, J. Huang, and X. Kang, "Synthetic speech detection based on local autoregression and variance statistics," *IEEE Signal Processing Letters*, vol. 29, pp. 1462–1466, 2022.

[10] T. Arif, A. Javed, M. Alhameed, F. Jeribi, and A. Tahir, "Voice spoofing countermeasure for logical access attacks detection," *IEEE Access*, vol. 9, pp. 162 857–162 868, 2021.

[11] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," 2015.

[12] K. Ma, Y. Feng, B. Chen, and G. Zhao, "End-to-end dual-branch network towards synthetic speech detection," *IEEE Signal Processing Letters*, vol. 30, pp. 359–363, 2023.

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

[14] 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, p. e28, 2020.

[15] E. Long, *Slicer*. iUniverse, 2000.

[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, p. 101114, 2020.

[17] A. Muniasamy, "Applications of keystroke dynamics biometrics in online learning environments: A selective study," in *Biometric Authentication in Online Learning Environments*. IGI Global, 2019, pp. 97–121.

[18] J. McAvoy and D. Sammon, *Agile methodology adoption decisions: An innovative approach to teaching and learning*. EDSIG, 2005, vol. 16, no. 4.

[19] D. A. van Leeuwen and N. Brümmer, *An Introduction to Application-Independent Evaluation of Speaker Recognition Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 330–353.

[20] R. Kundu, "F1 score in machine learning: Intro & calculation," 2023.