Indonesian Speech Anti-Spoofing System: Data Creation and Convolutional Neural Network Models

Sarah Azka Arief
Institut Teknologi Bandung,
Jl. Ganesa No. 10, Bandung, Indonesia
sarahazkarief@gmail.com

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

Dessi Puji Lestari
Institut Teknologi Bandung,
Jl. Ganesa No. 10, Bandung, Indonesia
dessipuji@staff.stei.itb.ac.id

2024 11th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA) | 979-8-3315-2031-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICAICTA63815.2024.10763091

***Abstract*— Biometric systems are prone to spoofing attacks. While research in speech anti-spoofing has been progressing, there is a limited availability of diverse language datasets. This study aims to bridge this gap by developing an Indonesian spoofed speech dataset, which includes replay attacks, text-to-speech, and voice conversion. This dataset forms the foundation for creating an Indonesian speech anti-spoofing system. Subsequently, light convolutional neural network (LCNN) and residual network (ResNet) models, based on convolutional neural networks (CNN), were developed to evaluate the dataset. The input features used are linear frequency cepstral coefficients (LFCC). Both models demonstrate remarkably low minDCF and EER scores approaching zero. The results also exhibit exceptional scores with 4-fold cross validation, showing strong initial performance with no signs of overfitting. However, models trained solely on Common Voice or Prosa.ai datasets performed poorly in cross-source tests, suggesting generalization issues due to a lack of diversity in the dataset. This highlights the need for further improvement and continued research in Indonesian speech spoof detection.**

Keywords—Spoof speech detection, Indonesian, ResNet, LCNN, LFCC

I. INTRODUCTION

Technological advancements in neural networks have promoted biometric systems, such as Automatic Speaker Verification (ASV) systems [1]. These systems are vulnerable to spoofing attacks [2], where attackers use spoofed data to impersonate verified speakers, potentially leading to identity theft or data breaches [3]. Attacks can be categorized as physical access (PA) scenarios, including replay attacks, and logical access (LA) scenarios, including speech synthesis and voice conversion.

Although studies to develop countermeasures have been conducted throughout the years, as seen in previous ASVspoof challenges [4], [5], [6], [7], most are limited to certain languages due to dataset constraints. There remains a need for inclusivity for underrepresented languages in the field of research for spoof speech detection, such as Bahasa Indonesia. This paper aims to fill the research gap by developing a convolutional neural network-based system to detect spoofed speech in the Indonesian language.

Several studies on spoof speech detection for specific language domains have been conducted in neighboring languages. In Thailand, Galajit et al. constructed ThaiSpoof, a database for spoof detection in the Thai Language [8]. The spoofed data in the database were generated using text-to-speech (TTS) tools, fundamental frequency modifications, and pitch shifting. The utilization of the dataset was later shown using a convolutional neural network (CNN) model with linear frequency cepstral coefficient (LFCC) features. Similar research involves the creation of FMFCC-A by Zhang et al., where spoofed data are generated using 11 Mandarin TTS system and 2 voice conversion (VC) systems [9].

Our research focuses on the Indonesian language in hopes of developing a robust spoofed speech detection system. This study involves the development of a specialized dataset for Indonesian spoof detection, addressing the lack of resources in this area. Additionally, it explores the application of CNN-based models, particularly LCNN and ResNet, in the context of Indonesian spoof speech detection. This research seeks to determine the effectiveness of these models when applied to the newly developed dataset and to assess their ability to generalize across different data sources. These efforts contribute to advancing the understanding and technology of spoofed speech detection in Bahasa Indonesia, setting the stage for future research and applications in this field.

II. SPEECH SPOOF DETECTION

Automatic Speaker Verification (ASV) systems have played a significant role in identifying spoofed speech through implementing machine learning techniques to increase accuracy, ensuring the security of systems [10]. ASV systems can be integrated with countermeasure systems, which are mechanisms or techniques designed to protect ASV systems from spoof attacks such as spoof speech detection systems [6].

Spoofed speeches are used in spoofing attacks for fraud, threats, or spreading false information. Several techniques to generate spoofed speech includes speech synthesis, voice conversion, impersonation, and replay attack [11]. Spoofing attacks in ASV can be categorized into physical access (PA) scenarios such as replay attacks and logical access (LA) scenarios such as spoof attacks that are based upon machine-generated samples such as speech synthesis and voice conversion [12]. Effective countermeasure solutions are required to detect both types of attacks.

According to a study done by Mittal & Dua, early development of countermeasures involved classical machine learning approaches, but with the rise of deep learning, the focus has shifted to neural network-based approaches [10]. Convolutional neural network (CNN) models are favored for their minimal preprocessing requirements due to kernel usage [10] and their ability to automatically extract features from sound, capturing complex spatial and temporal variations in speech signals [9]. Two popular CNN-based architectures are residual networks (ResNet) [13] and light convolutional neural networks (LCNN) [14] [9]. Features, which are characteristics or attributes of speech signals, also play a fundamental role in affecting the accuracy and robustness of a spoof speech detection algorithm [15] with linear frequency cepstral coefficient (LFCC) being one of the most common features used in spoof speech detection with it being used as features for the baseline countermeasure systems for ASVspoof 2019 [6] and 2021 [7] challenges.

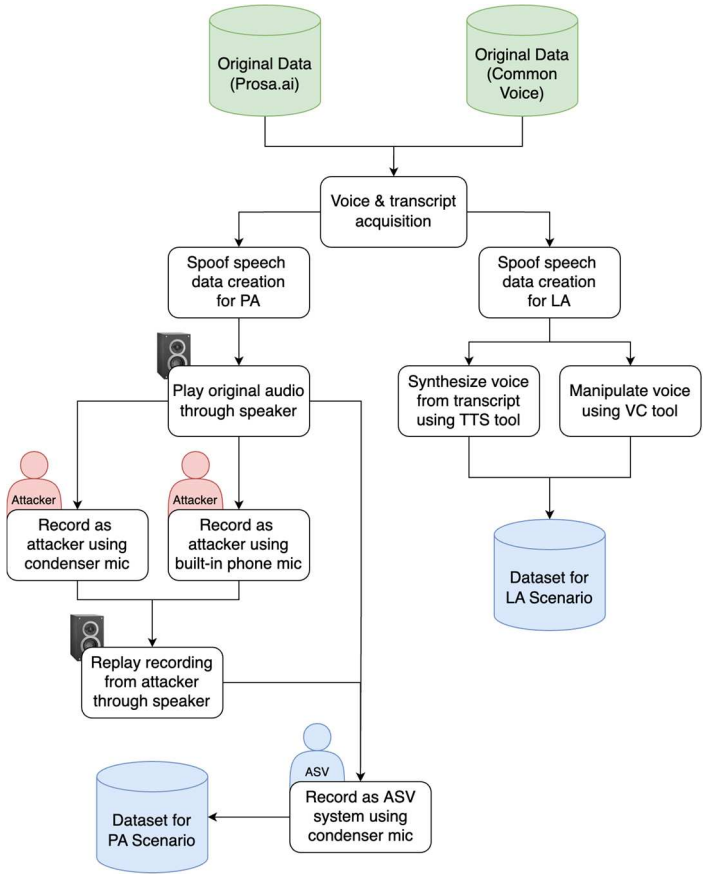


Fig. 1. General flow for the dataset development

TABLE I
BONA FIDE AND SPOOF AUDIO IN LA DATASET

Source	Type	Accent	Number of Utterances
Common Voice	Bona fide	multi	4,540
	Spoof (VC)	multi	90,800
	Spoof (TTS)	bbc	4,540
	Spoof (TTS)	ind	4,540
	Spoof (TTS)	jav	4,540
Prosa.ai	Bona fide	multi	2,000
	Spoof (VC)	multi	120,000
	Spoof (TTS)	bbc	2,000
	Spoof (TTS)	ind	2,000
	Spoof (TTS)	jav	2,000
Total			236,960

III. BUILDING THE DATASET

This section outlines the development of the dataset which is briefly visualized in Figure 1. The initial Bahasa Indonesia audio used to build the dataset is sourced from two existing datasets: an open-source dataset from Common Voice [16] and a proprietary dataset developed by Prosa.ai [17]. As the Common Voice dataset is made up of audio data from volunteers, the acoustic conditions vary from one audio file to another with the appearance of noise being unavoidable for some of the audio. In contrast, the Prosa.ai dataset consists of audio samples from 53 unique speakers with a balanced gender distribution of approximately 50% male and 50% female, which are studio recorded to ensure clear and consistent audio quality throughout each audio. The duration of individual audio samples from both Common Voice and Prosa.ai ranges from 3 to 60 seconds.

Through combining the data from Common Voice and Prosa.ai, we develop a dataset for the purpose of spoofed speech detection in Bahasa Indonesia. This dataset consists of bona fide and spoofed speech for both LA and PA scenarios. Tools and methods used to generate the spoofed speeches in the dataset include replay attack simulations, text-to-speech (TTS) for speech synthesis, and voice conversion (VC).

A. LA scenario

For the LA scenario, spoofed speech data are generated using the Massively Multilingual Speech (MMS) [18] model for TTS-based spoofs and the FreeVC [19] system for VC-based spoofs. The original data used to create the spoofed ones include multiple accents or variations within the Indonesian language (multi). The MMS model was used to create speech data for three major Indonesian accents: Indonesian (ind), Javanese (jav), and Bataknese (bbc). However, a limitation of the MMS model is its ability to generate speech for only one male speaker. On the other hand, the FreeVC system can change a speaker's voice without needing any text input, extracting key features from the original voice to create realistic sounding spoofed speech. Table I shows the generated Indonesian language spoof data along with the bona fide data.

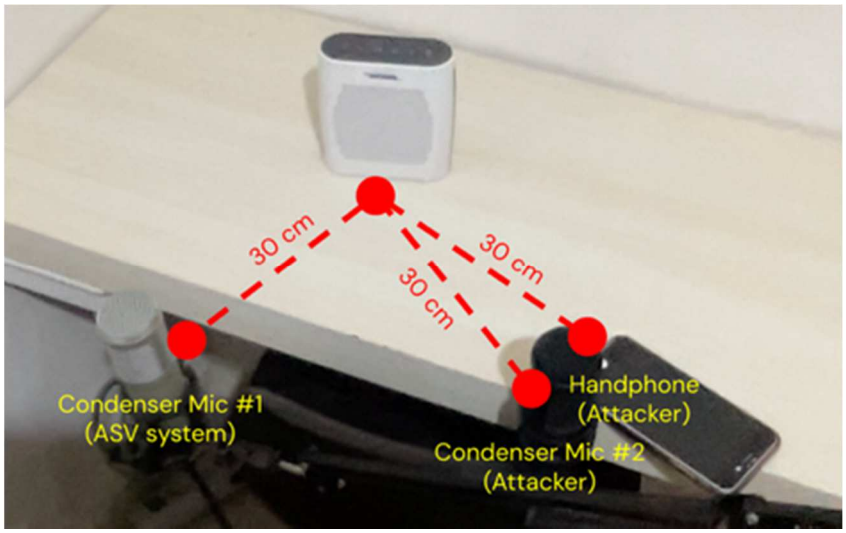


Fig. 2. Microphone and speaker placements for replay attack simulation

TABLE II
BONA FIDE AND SPOOF AUDIO IN PA DATASET

Source	Type	Number of Utterances
Common Voice	Bona fide	4,540
	Spoof (Handphone Mic)	4,540
	Spoof (Condenser Mic)	4,540
Prosa.ai	Bona fide	2,000
	Spoof (Handphone Mic)	2,000
	Spoof (Condenser Mic)	2,000
Total		19,620

B. *PA scenario*

For PA scenario, spoofed speech data are made by simulating replay attacks. The process involves playing original audio through a speaker and recording it simultaneously on three microphones: a condenser microphone for the ASV system, a condenser microphone for the attacker, and a built-in handphone microphone for the attacker. These microphones are connected to a digital audio workstation (DAW) for recording and are placed 30 centimeters (about 11.81 in) away from the speaker with placements seen in Figure 2.

The ASV system's recordings are used as bona fide data, while the attacker's recordings are replayed through a speaker and re-recorded by the ASV system's microphone, creating spoof audio samples for the PA scenario dataset. Table II shows the Indonesian language spoof dataset consisting of bona fide and spoof data for PA scenario.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. *Spoof detection models*

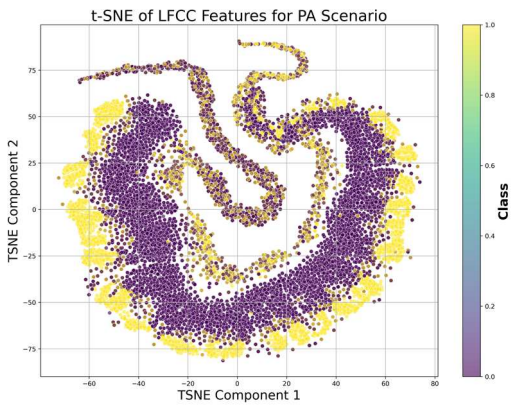
The choice of using convolutional neural network (CNN) based models is due to their ability to identify local patterns in data with minimal preprocessing, making them a widely used deep learning approach for spoof speech detection [10]. Models used to carry out the development of Indonesian spoof speech detection are residual networks (ResNet) and also light convolutional neural networks (LCNN). Linear frequency cepstral coefficients (LFCC) are chosen as the input feature, as it is one of the most common features used due to its performance in differentiating between bona fide and spoof speech [20].

The LCNN model is the baseline model established by ASVspoof 2021 [7], which also takes in LFCC as its input. We developed a ResNet model for spoof detection system, leveraging the framework proposed in [13]. For both models, the epoch is set to 50, batch size 16, and learning rate 0.001. Early stopping with patience of 5 is implemented for both ResNet and LCNN model. Scores used will be based off the metrics used and defined in ASVspoof 2024 which are minDCF and EER [21]. Experiments will be done separately for both PA and LA scenarios.

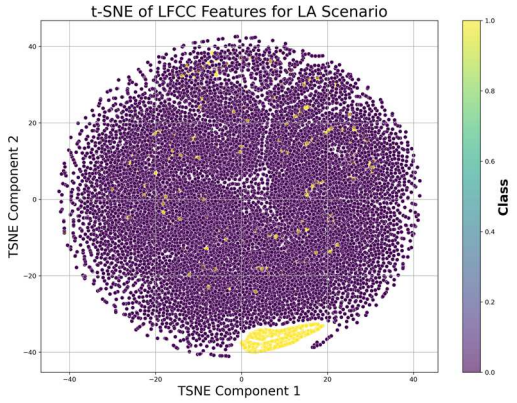
B. *Dataset partitioning*

In this stage, the datasets, comprising processed audio for both PA and LA scenarios, are divided into three subsets: training, development, and testing. Since the Common Voice dataset is already pre-split, this section focuses on the partitioning of the Prosa.ai data. Of the 50 speakers in the Prosa.ai dataset, 30 speakers are allocated to the training subset, 12 to the development subset, and 8 to the testing subset, ensuring no speaker overlaps between subsets. This allocation was chosen to achieve a balanced distribution for training, development, and testing subsets, specifically approximately 60% for training, 20% for development, and 20% for testing across both PA and LA scenario datasets. This distribution is particularly beneficial for ensuring robust model evaluation and reducing bias, which is crucial for spoofed speech detection systems where model accuracy is critical.

With the initial 50 speakers in the Prosa.ai dataset evenly divided by gender, gender balance is maintained across all subsets, with equal representation of male and female speakers. The training subset includes 15 male and 15 female speakers, the development subset includes 6 male and 6 female speakers, and the testing subset includes 4 male and 4 female speakers. Additionally, the balance among types of spoofed audio is ensured within each subset to prevent any category from dominating, thus reducing bias in model training and evaluation.



(a) PA scenario



(b) LA scenario

Fig. 3. 2D Projection of features using t-SNE

TABLE III
DISTRIBUTION FOR TRAIN, DEV, TEST SUBSET IN PA DATASET

Source	Type	Number of Utterances		
		Train	Dev	Test
Common Voice	Bona fide	3,055	780	705
	Spoof (Handphone)	3,055	780	705
	Spoof (Condenser)	3,055	780	705
Prosa.ai	Bona fide	1,200	400	400
	Spoof (Handphone)	1,200	400	400
	Spoof (Condenser)	1,200	400	400
Total		12,765	3,540	3,315
		19,620		

Table III shows the respective train, development, and test subset distributions for the PA scenario dataset, resulting in 65% for training, 18% for development, and 17% for testing. For the LA scenario, Table IV shows the final distributions for each subset, resulting in 60% for training, 20% for development, and 20% for testing.

C. *Experimental setup and metrics*

Both LFCC features from PA and LA datasets are visualized and projected in 2D using t-SNE to get a better understanding of the input of the models beforehand as seen in Figure 3. In Figure 3, yellow indicates data belonging to the bona fide class, while purple indicates data labeled as spoof. The substantial class overlap, and intricate patterns shown by the data points for PA scenario reveals a more complex and heterogeneous distribution compared to LA scenario's homogenous data points.

To fully grasp the performance of the model relative to the dataset, models are trained and validated using the 4-fold cross validation scheme. After that, models are rebuilt using the entire training and development subset and is later tested using three slightly different test sets: the entirety of the test subset, parts of the test subset that is made up of only Common Voice audio and parts of the test subset that is made up of only Prosa.ai audio. The scores received will be used for comparison alongside the average score from previously described 4-fold cross validation phase.

TABLE IV
DISTRIBUTION FOR TRAIN, DEV, TEST SUBSET IN LA DATASET

Source	Type	Accent	Number of Utterances		
			Train	Dev	Test
Common Voice	Bona fide	multi	3,055	780	705
	Spoof (VC)	multi	54,480	18,160	18,160
	Spoof (TTS)	bbc	3,055	780	705
	Spoof (TTS)	ind	3,055	780	705
	Spoof (TTS)	jav	3,055	780	705
Prosa.ai	Bona fide	multi	1,200	320	480
	Spoof (VC)	multi	72,000	24,000	24,000
	Spoof (TTS)	bbc	1,200	320	480
	Spoof (TTS)	ind	1,200	320	480
	Spoof (TTS)	jav	1,200	320	480
Total			143,500	46,560	46,900
			236,960		

To assess model generalization and dataset behavior from Common Voice and Prosa.ai, each model was trained and validated using training and development subsets exclusively from each source. Each trained model was then tested on test subsets of same-source (e.g., model trained using Common Voice only data is tested using parts of test subset that is only from Common Voice) and cross-source (e.g., model trained using Common Voice only data is tested using parts of test subset that is only from Prosa.ai).

Metrics used to evaluate the performance of the models are minimum detection cost function (minDCF) and equal error rate (EER). Both metrics are previously defined in ASVspoof Evaluation Plan 5 and used in ASVspoof 2024 [21] for speech deepfake detection in a stand-alone setting without speaker verification.

TABLE V
SCORES FOR AVERAGE FOLD AND TEST (PA)

Phase	minDCF		EER (%)	
	LCNN	ResNet	LCNN	ResNet
1st Fold	0.00682	0.02005	0.25119	0.73471
2nd Fold	0.02004	0.00000	0.90160	0.00000
3rd Fold	0.00000	0.00000	0.00000	0.00000
4th Fold	0.00000	0.00000	0.00000	0.00000
Average	0.00671	0.00501	0.28820	0.18368
Tested on Prosa.ai Only	0.00236	0.00943	0.25681	0.51363
Tested on Common Voice Only	0.00894	0.00000	0.31915	0.00000
Tested on Entirety of Test Subset	0.00684	0.00218	0.27716	0.10147

TABLE VI
SCORES FOR SAME/CROSS-SOURCE TEST (PA)

Trained On	Tested On	minDCF		EER (%)	
		LCNN	ResNet	LCNN	ResNet
Common Voice	Same-source	0.00695	0.00000	0.28369	0.00000
	Cross-source	1.00000	1.00000	80.8648	99.7432
Prosa.ai	Same-source	0.00000	0.00000	0.00000	0.00000
	Cross-source	1.00000	1.00000	99.2908	100.000

TABLE VII
SCORES FOR AVERAGE FOLD AND TEST (LA)

Phase	minDCF		EER (%)	
	LCNN	ResNet	LCNN	ResNet
1st Fold	0.00011	0.00000	0.00539	0.00000
2nd Fold	0.00002	0.00002	0.00109	0.00109
3rd Fold	0.00002	0.00002	0.00108	0.00108
4th Fold	0.00039	0.00009	0.05501	0.00434
Average	0.00014	0.00003	0.01564	0.00163
Tested on Prosa.ai Only	0.00047	0.00000	0.02359	0.00000
Tested on Common Voice Only	0.00158	0.00030	0.14491	0.01480
Tested on Entirety of Test Subset	0.00147	0.00000	0.15220	0.00000

TABLE VIII
SCORES FOR SAME/CROSS SOURCE TEST (LA)

Trained On	Tested On	minDCF		EER (%)	
		LCNN	ResNet	LCNN	ResNet
Common Voice	Same-source	0.00000	0.00000	0.00000	0.00000
	Cross-source	0.59712	1.00000	30.0000	100.000
Prosa.ai	Same-source	0.00008	0.00396	0.00393	0.20833
	Cross-source	0.00035	0.00623	0.01726	0.28488

D. Results and analysis

This subsection highlights the results of the experiment and analyzes the outcome. For PA, Table V summarizes the performance of both LCNN and ResNet on 4-fold cross validation and specific tests. Both models show excellent performance in the PA scenario, with minDCF and EER values reaching zero when evaluated using 4-fold cross validation. However, when tested on cross-source data (Table VI), both models perform poorly, with minDCF scores of 1 and EERs ranging from 81-100%, demonstrating poor generalization. The poor generalization in the PA scenario could be attributed to several factors:

- Differences in recording quality between studio-recorded Prosa.ai data and volunteer-recorded Common Voice data.
- Variations in recording settings, such as gain and stereo-to-mono conversion, which might remove critical features for spoof detection. These features could include subtle acoustic cues like room reverberation or microphone characteristics that are present in genuine recordings but may be altered or absent in spoofed ones.
- Limited data in the PA scenario dataset compared to the LA scenario, potentially restricting the models' exposure to diverse conditions.

For the LA scenario, Table VII displays both model's performance, including results from 4-fold cross-validation and specific tests. The results demonstrate excellent performance with minDCF and EER scores consistently approaching zero. However, in cross-source testing (as seen in Table VIII), models trained on the Common Voice dataset showed inferior performance, while models trained on the Prosa.ai dataset maintained good performance even when tested on Common Voice data. The hypothesis drawn from these results include:

- The characteristics of the data. LA scenario might benefit from the consistent quality of studio-recorded data (Prosa.ai), making it easier for models to learn distinguishing features between genuine and spoofed speech.
- In LA scenario, models likely use speaker-dependent characteristics from speech for discrimination. These features, such as fundamental frequency patterns or spectral envelope characteristics, are more consistent across different recording conditions. In contrast, PA scenario models rely more on detecting speech distortions introduced during replay attacks, which can vary significantly based on recording equipment and environment.
- The spoof audio in LA dataset might produce more consistent artifacts, while the ones in PA spoof might introduce more variable distortions depending on the playback and re-recording conditions.

The extremely low EER (0%) in same-source testing contrasted with very high EER (near 100%) in cross-source testing for some models raises concerns about the credibility of these results. This stark difference suggests potential overfitting to source-specific characteristics rather than learning generalizable features for spoof detection. Future

work should investigate this phenomenon and consider techniques to improve cross-source generalization.

V. Conclusion

This study focused on creating an Indonesian spoofed speech dataset to support the development of Indonesian speech anti-spoofing systems. The dataset was compiled from two sources, Common Voice and Prosa.ai, incorporating various accents and recording conditions for both logical access (LA) and physical access (PA) scenarios. The LA scenario dataset consists of text-to-speech-based spoofs generated using the Massively Multilingual Speech (MMS) model and voice-conversion-based spoofs generated using the FreeVC system. For PA scenario dataset, replay attacks were simulated using different microphone setups to generate the spoofed speeches.

Evaluations of light convolutional neural network (LCNN) and residual network (ResNet) models on this dataset showed remarkable performance, including 4-fold cross validation tests, indicating robust initial outcomes without overfitting concerns. However, the models exhibited significant generalization issues when trained on individual datasets and tested across sources, underscoring the need for diverse and high-quality datasets. Despite these challenges, the consistent quality of the Prosa.ai dataset facilitated better generalization, highlighting the importance of high-quality data in training.

Future work should focus on developing techniques to improve cross-source generalization and enhancing dataset diversity and quality to improve model generalization. To enhance the diversity of the PA dataset, future work could consider using methods that simulate room acoustics to generate additional pseudo data. This approach could potentially increase the robustness of models trained on this dataset. Additionally, exploring further countermeasures and hybrid approaches could improve the robustness of Indonesian speech anti-spoofing systems. This study contributes to the development of a foundational dataset and provides a basis for future research in Indonesian speech anti-spoofing.

Acknowledgment

This research is financially supported by PPMI ITB 2024. The authors would also like to express their gratitude to Prosa.ai for providing the dataset and computational resources essential for our experiments. Additionally, the ASEAN IVO (http://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 presentation and financially supported by NICT (http: //www.nict.go.jp/en/index.html).

References

[1] A. H. K. M. K. and P. S. Aithal, “Voice Biometric Systems for User Identification and Authentication – A Literature Review,” International Journal of Applied Engineering and Management Letters, pp. 198–209, Apr. 2022, doi: 10.47992/ijaeml.2581.7000.0131.

[2] X. Liu et al., “ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild,” IEEE/ACM Trans Audio Speech Lang Process, vol. 31, pp. 2507–2522, 2023, doi: 10.1109/TASLP.2023.3285283.

[3] A. Nautsch et al., “ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech,” IEEE Trans. Biom. Behav. Identity Sci., vol. 3 (2), pp. 252-264, Feb. 2021, doi: 10.1109/TBIOM.2021.3059479.

[4] Z. Wu et al., “ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge.” in Proceedings of INTERSPEECH, pp. 2037-2041, 2015, doi: 10.21437/INTERSPEECH.2015-462.

[5] Z. Wu et al., “ASVspoof: The automatic speaker verification spoofing and countermeasures challenge,” IEEE Journal on Selected Topics in Signal Processing, vol. 11, no. 4, pp. 588–604, Jun. 2017, doi: 10.1109/JSTSP.2017.2671435.

[6] X. Wang et al., “ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech,” Comput. Speech Lang., vol. 64, pp. 10101114, 2020, doi: 10.1016/J.CSL.2020.101114.

[7] J. Yamagishi et al., “ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection,” CoRR, vol. abs/2109.00537, 2021.

[8] K. Galajit et al., “ThaiSpoof: A Database for Spoof Detection in Thai Language” 2023 18th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), pp. 1-6, Nov. 2023, doi: 10.1109/iSAI-NLP60301.2023.10354956.

[9] Z. Zhang, Y. Gu, X. Yi, and X. Zhao, “FMFCC-A: A Challenging Mandarin Dataset for Synthetic Speech Detection,” in IWDW 2021, vol. 13180, pp. 117–131, 2021, doi: 10.1007/978-3-030-95398-0_9.

[10] A. Mittal and M. Dua, “Automatic speaker verification systems and spoof detection techniques: review and analysis,” Int J Speech Technol, vol. 25, no. 1, pp. 105–134, Mar. 2022, doi: 10.1007/s10772-021-09876-2.

[11] C. B. Tan et al., “A survey on presentation attack detection for automatic speaker verification systems: State-of-the-art, taxonomy, issues and future direction,” Multimed Tools Appl, vol. 80, no. 21–23, pp. 32725–32762, Sep. 2021, doi: 10.1007/s11042-021-11235-x.

[12] A. Khan, K. M. Malik, J. Ryan, and M. Saravanan, “Battling voice spoofing: a review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures,” Artif Intell Rev, vol. 56, pp. 513–566, Oct. 2023, doi: 10.1007/s10462-023-10539-8.

[13] M. Alzantot, Z. Wang, and M. B. Srivastava, “Deep residual neural networks for audio spoofing detection,” in Proceedings of INTERSPEECH, 2019, pp. 1078–1082. doi: 10.21437/Interspeech.2019-3174.

[14] R. Białobrzeski, M. Kośmider, M. Matuszewski, M. Plata, and A. Rakowski, “Robust Bayesian and light neural networks for voice spoofing detection,” in Proceedings of INTERSPEECH, 2019, pp. 1028–1032. doi: 10.21437/Interspeech.2019-2676.

[15] Z. Wu et al., “ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge.” [Online]. Available: http://www.festvox.org/

[16] “Common Voice.” Accessed: Jul. 22, 2024. [Online]. Available: https://commonvoice.mozilla.org/en

[17] “Prosa AI | Indonesian Natural Language Processing Solutions.” Accessed: Jul. 22, 2024. [Online]. Available: https://prosa.ai/

[18] V. Pratap et al., “Scaling Speech Technology to 1,000+ Languages,” CoRR, vol. abs/2305.13516, 2023, doi: 10.48550/ARXIV.2305.13516.

[19] J. Li, W. Tu, and L. Xiao, “Freevc: Towards High-Quality Text-Free One-Shot Voice Conversion,” ICASSP, pp. 1–.5, 2023, doi: 10.1109/ICASSP49357.2023.10095191.

[20] M. Todisco, H. Delgado, and N. Evans, “A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients,” in Odyssey 2016: Speaker and Language Recognition Workshop, International Speech Communication Association, 2016, pp. 283–290. doi: 10.21437/Odyssey.2016-41.

[21] H. Delgado et al., “ASVspoof 5 Evaluation Plan *,” 2024, Accessed: Jul. 23, 2024. [Online]. Available: http://www.asvspoof.org/