SYNTHETIC SPEECH DETECTION USING RAWNET WITH PATHOLOGICAL FEATURES

Louis Caesa Kesuma[1], Dessi Puji Lestari[1], Candy Olivia Mawalim[1, 2]

[1]Institut Teknologi Bandung, Indonesia
[2]Japan Advanced Institute of Science and Technology, Japan
13521069@std.stei.itb.ac.id,
dessipuji@itb.ac.id,
candylim@jaist.ac.jp

ABSTRACT

Realistic synthetic voices resembling genuine ones pose new challenges for Automatic Speaker Verification (ASV) systems. This study develops three RawNet variants (RawNet1, RawNet2, and RawNet3) integrated with pathological features such as jitter, shimmer, and glottal-to-noise excitation ratio (GNE). The dataset includes genuine and synthetic Indonesian speech, aiming to strengthen research on underrepresented languages. Feature integration was explored at both feature and architectural level. Results show that combined models outperform the baseline, with architectural-level integration achieving the best performance, particularly in handling unprecedented synthetic voices. These findings highlight that pathological features can improve the accuracy and robustness of synthetic speech detection in Indonesian.

***Index Terms*—** Synthetic Voice, RawNet, Pathological Features, Automatic Speaker Verification (ASV), Indonesian

1. INTRODUCTION

Artificial intelligence has become increasingly prevalent across social and technological domains. One notable advancement is synthetic speech, which brings benefits in applications such as virtual assistants but also introduces serious security challenges for Automatic Speaker Verification (ASV) systems [18]. A major threat is the misuse of synthetic voices for fraud, identity theft, and other malicious purposes. Synthetic voices have become highly natural and difficult to distinguish from genuine speech due to rapid progress in speech synthesis, creating an urgent need for robust detection systems [13, 19].

Recent studies propose using pathological features which are commonly linked to disordered speech, to improve fake voice detection [2, 3]. In parallel, raw features focused models such as RawNet [8, 9, 10, 11] have gained popularity for its ability to process the abstractness of raw features, achieving strong results in ASVspoof challenges [17]. However, further exploration of RawNet variants remains necessary, especially in combination with newer techniques such as pathological features.

Another challenge lies in dataset bias. Existing corpora focus mainly on English and Mandarin, limiting performance when models are applied to other languages [14]. This underlines the lack of research and development on synthetic speech detection in Indonesian, making it crucial to develop both datasets and models tailored to this language.

This study addresses these gaps by developing an Indonesian dataset of genuine and synthetic speech and evaluating RawNet models enriched with pathological features. The findings aim to improve detection accuracy and robustness against synthetic speech in low-resource language contexts.

2. RELATED WORK

2.1. Speech-Pathological Features

Pathological features are audio features often associated with voice disorders [2]. These features share similarities with those extracted from synthetic speech, making them potential discriminative factors in detecting synthetic voices. Below are several variations of pathological features along with their explanations.

- Jitter: variation in the period of the voice waveform, reflecting frequency stability.
- Shimmer: variation in amplitude, indirectly reflecting vocal effort.
- Harmonic-to-Noise Ratio (HNR): quantifies the balance between harmonic and noise components.
- Cepstral Harmonic-to-Noise Ratio (CHNR): similar to HNR but emphasizes energy differences between the full spectrum and noise components.
- Normalized Noise Energy (NNE): measures additional noise based on the ratio of noise energy to total signal energy per frame.

- Glottal-to-Noise Excitation Ratio (GNE): represents irregular noise (turbulence) in the signal.

The potential of pathological features for synthetic speech detection using the ASVspoof 2019 dataset [16]. Beyond standalone evaluation, they tested various combinations of pathological features and ultimately achieved the best performance when these were integrated with derivative features and additional techniques, suggesting that pathological features hold promise particularly when combined with other methods. Specifically, they used ten segmental pathological features, which is a combination of jitter (local, PPQ3, and PPQ5), shimmer (local, APQ3, APQ5, and APQ11), CHNR, NNE, GNE, and HNR.

2.2. Raw Waveform

Raw features such as the raw waveform are capable of representing a broader spectrum, particularly in capturing temporal dynamics where handcrafted features often fail [17]. Moreover, the use of raw features reduces technical constraints in speech research, thereby providing greater opportunities for integration with various other approaches [9].

2.3. RawNet

RawNet is a speaker verification system that performs verification directly from raw audio without prior acoustic feature extraction [8]. Over the course of its development, several iterations of RawNet have been introduced, which are RawNet1, RawNet2, and RawNet3.

2.3.1. RawNet1

Table 1. RawNet1 architecture [9]

Layer	Input: 59049 samples	Output shape
Strided-conv	Conv(3, 3, 128) BN LeakyReLU	(19683, 128)
Res block	Conv(3, 1, 128) BN LeakyReLU Conv(3, 1, 128) BN LeakyReLU Maxpooling(3) × 2	(2187, 128)
Res block	Conv(3, 1, 256) BN LeakyReLU Conv(3, 1, 256) BN LeakyReLU Maxpooling(3) × 4	(27, 256)
GRU	GRU(1024)	(1024)
Speaker Embedding	FC(128)	(128)
Output	FC(1211)	(1211)

RawNet1 is the first iteration of RawNet, developed specifically for speaker verification [14]. RawNet1 consists of several layers, as shown in Table 1.

2.3.2. RawNet2

RawNet2 is the updated version of RawNet1. The main difference between RawNet1 and RawNet2 is the filter layer. RawNet2 used sinc filter, while RawNet1 used strided convolutional in the filter layer. The architecture of RawNet2 is shown in Table 2. The use of raw audio enables RawNet2 to learn patterns and cues that would remain undetected when using more traditional methods [17].

Table 2. RawNet2 architecture [10]

Layer	Input: 59049 samples	Output shape
Fixed Sinc filters	Conv(251, 1, 128) Maxpooling(3) BN & LeakyReLU	(19683, 128)
Res block	BN & LeakyReLU Conv(3, 1, 128) BN & LeakyReLU Conv(3, 1, 128) Maxpooling(3) FMS × 2	(2187, 128)
Res block	BN & LeakyReLU Conv(3, 1, 256) BN & LeakyReLU Conv(3, 1, 256) Maxpooling(3) FMS × 4	(27, 256)
GRU	GRU(1024)	(1024)
Speaker Embedding	FC(1024)	(1024)

2.3.3. RawNet3

RawNet3 is the improved version of RawNet2, featuring architectural modifications, a self-supervised training scheme, and the use of parameterized filter bank for direct speech signal processing [11]. The architecture of RawNet3 is illustrated in Fig. 1 and Fig. 2.

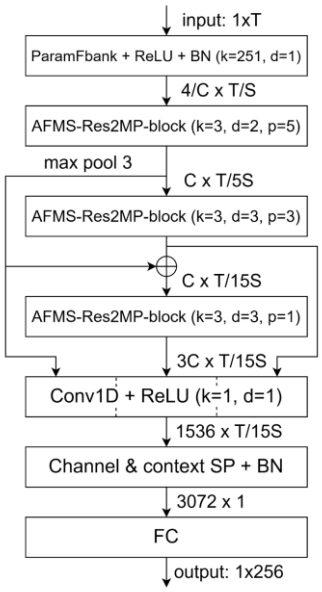

Figure 1. RawNet3 architecture [11]

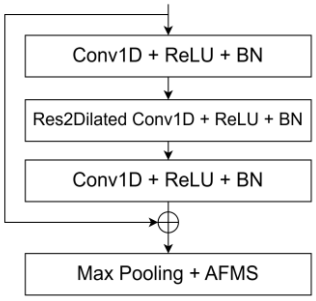

Figure 2. AFMS-Res2MP-block in RawNet3 [11]

3. PROPOSED METHOD

Experiments did not employ the full set of ten segmental pathological features described in [16]. Instead, only a subset was utilized: jitter (local, PPQ3, and PPQ5), shimmer (local, APQ3, APQ5, and APQ11), and GNE. These features were extracted using the Python libraries praat-parselmouth and openDBM. To enrich the representation, first- and second-order derivatives (Δ and ΔΔ) were also included, resulting in a total of 24 features rather than the original 30.

RawNet1, RawNet2, and RawNet3 were evaluated with and without pathological feature integration. Baseline models were compared against two integration schemes: feature-level (GF) and architectural-level (GA). For the baseline configuration, the input dimension was fixed at 64,000, corresponding to 4 seconds of speech sampled at 16 kHz.

3.1. GF models

Feature-level integration was performed by concatenating the raw waveform features with the pathological features. This approach does not modify the RawNet architecture and can be represented with a simple illustration, as shown in Fig.3. The only change is in the input dimension, which increases from [64000, 1] to [64000 + 24, 1] due to the addition of 24 pathological features.

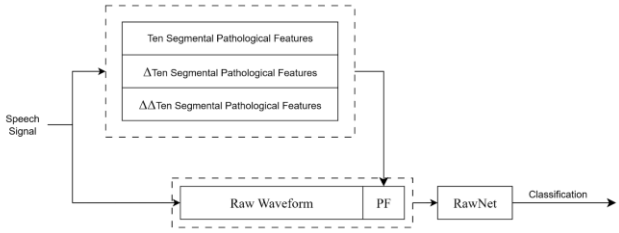

Figure 3. RawNet GF architecture

3.2. GA models

Architectural-level integration does not substantially alter the overall RawNet design, but only modifies the dimensionality of the fully connected (FC) layer. This approach is intended to enhance semantic performance, as RawNet is primarily designed to process raw features and may not be well suited to directly handle pathological features or other hand-crafted features in the same manner. Instead, this method enables RawNet to perform classification by combining processed raw features with externally processed pathological features. Unlike GF, however, the input dimensions of the FC layer differ across RawNet architectures after the inclusion of pathological features. The general architecture of RawNet with GA integration is shown in Fig. 4.

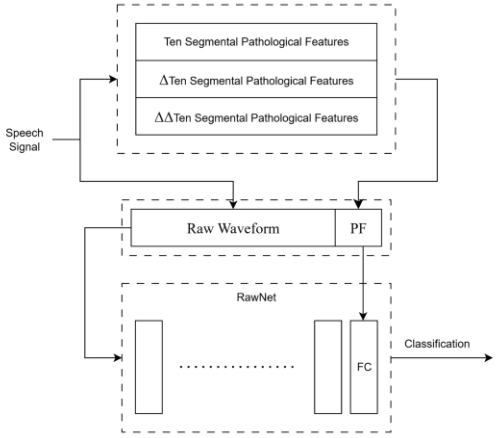

Figure 4. RawNet GA architecture

4. EXPERIMENTAL SETUP

Experiments in this study were conducted using a modified version of RawNet1, RawNet2, and RawNet3. Code used for experiments is available at `https://github.com/Ainzw0rth/RawNets-PF.git`

4.1. Dataset

The real speech dataset employed in this study comprises utterances from Common Voice and Prosa.ai. Synthetic speech was generated from this dataset using several TTS systems, including Google TTS [6], Facebook MMS [15], VITS [12], ElevenLabs [5], and DupDub [4]. Furthermore, FreeVC [7] was applied to combine real and synthetic data in order to construct a more realistic synthetic voice dataset. The diversity of technologies used was influenced by datasets in similar studies [1, 13, 14].

Table 3. Details of each data source for both real and synthetic speech

Category	Source	Total	Avg Duration
Real	Common Voice	5,000	~ 3.76 s
	Prosa	2,000	~ 5.15 s
Synthetic (TTS)	MMS	7,000	~ 3.89 s
	VITS	7,000	~ 3.28 s
	Google TTS	7,000	~ 3.84 s
	ElevenLabs	349	~ 3.68 s
	DupDub	40	~ 4.0 s
	DupDub (UnID)	40	~ 4.21 s
Synthetic (VC)	FreeVC dan Google TTS	3,469	~ 3.84 s
	FreeVC dan MMS	3,469	~ 3.89 s
	FreeVC dan VITS	3,469	~ 3.27 s
	FreeVC dan ElevenLabs	349	~ 3.66 s
	FreeVC dan DupDub	20	~ 4.04 s
Total		39,205	~ 3.89 s

The resulting datasets were evaluated under three experimental scenarios: seen, unseen-in-dataset (UID), and unseen-not-in-dataset (UnID). The seen scenario assesses model performance when both the TTS system and transcript appear in the training set. The UID scenario evaluates detection of synthetic speech generated by previously unseen TTS systems while maintaining transcripts from the training set. The UnID scenario measures model robustness when both the TTS system and transcripts are entirely absent from the training data.

In the seen scenario, the dataset was partitioned into 70% training, 15% validation, and 15% testing, with the same real speech test set also incorporated into the UID and UnID scenarios for consistency in evaluation. The details of each data source, both real and synthetic, are summarized in Table 3, while the data distribution across the three scenarios is presented in Table 4. For each audio recording, 4-second segments were extracted using a sliding window with a 1-second offset; segments shorter than 2 seconds were discarded. These 4-second segments were then used to extract both pathological features and raw waveform representations.

Table 4. Details of scenario data distribution

Scenario	Category	Source
Seen	Real	Common Voice
		Prosa
	Synthetic (TTS)	MMS
		VITS
	Synthetic (VC)	FreeVC and Google TTS
		FreeVC and MMS
		FreeVC and VITS
UID	Synthetic (TTS)	Google TTS
		ElevenLabs
		DupDub
	Synthetic (VC)	FreeVC and ElevenLabs
		FreeVC and DupDub
	Real	Common Voice
		Prosa
UnID	Synthetic (TTS)	DupDub (UnID)
	Real	Common Voice
		Prosa

Table 5. Model average performance across all scenarios. Highlighted results indicate that the model variation outperform the baseline and the other version of integration (RawNet1 will be compared with RawNet1-GF and RawNet1-GA, RawNet2 with RawNet2-GF and RawNet2-GA, and RawNet3 with RawNet3-GF and RawNet3-GA)

Metrik	Baseline			GF			GA		
	RawNet1	RawNet2	RawNet3	RawNet1-GF	RawNet2-GF	RawNet3-GF	RawNet1-GA	RawNet2-GA	RawNet3-GA
Accuracy	87.10%	95.73%	90.77%	86.59%	**97.06%**	90.09%	**93.00%**	96.82%	**94.06%**
Balanced Accuracy	73.69%	80.86%	77.58%	73.41%	82.24%	77.61%	**80.55%**	**83.31%**	**81.01%**
Precision	83.14%	92.61%	86.22%	82.79%	**94.98%**	85.57%	**88.62%**	94.43%	**89.99%**
Recall	100.00%	99.96%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
F1-Score	89.18%	95.97%	91.64%	88.87%	**97.37%**	91.13%	**93.37%**	97.05%	**94.31%**
F2-Score	94.76%	98.28%	96.15%	94.57%	**98.91%**	95.87%	**97.05%**	98.78%	**97.51%**
EER	3.87%	8.34%	**6.20%**	4.09%	**4.48%**	7.77%	**3.68%**	6.37%	7.57%

5. RESULTS AND ANALYSIS

Table 5 indicates that integrating pathological features into each RawNet model yields improved performance compared to the baseline. Furthermore, architectural-level integration (GA) generally outperforms feature-level integration (GF), with GA achieving the best results across all scenarios. This performance difference establishes GA as the most effective model, even among the integrated variants. The recall, which reaches 100% for all models (except the RawNet2 baseline), indicates that the developed models never misclassified genuine speech as synthetic when the genuine speech had been learned during training.

Table 5 concludes that integrating RawNet with pathological features not only improves performance but also enhances model robustness compared to the baseline. Among all models, RawNet2-GF and RawNet2-GA achieve the best performance, with the GA variant in general showing superior results over both the baseline and GF configurations. This finding is consistent with the approach in [16], which proposed separating raw and artificial features into two distinct pipelines and merging them at a later stage. The performance differences are likely due to the semantic design of RawNet, which is optimized for raw feature processing rather than handcrafted features such as pathological features.

6. CONCLUSION

Synthetic speech detection can be effectively performed using RawNet combined with pathological features. RawNet2-GA demonstrated the best performance in all scenarios, making it the most robust model. Incorporating pathological features at the architectural level, where raw waveform features are first processed before integration, was generally more effective than feature-level fusion. Across all RawNet variants, the addition of pathological features consistently improved both accuracy and robustness, enabling better detection of synthetic speech from previously unseen sources. Model performance and robustness can be further enhanced by training with synthetic speech data from a broader range of sources.

Future work should incorporate a wider range of TTS and VC technologies to increase the variability of synthetic speech data, thereby improving the model's ability to handle diverse real-world attacks. Expanding the scenarios and the quality of speech are also recommended to further evaluate model robustness. For models combining raw and handcrafted features, integration is suggested after raw features have been processed, as this approach has shown superior performance. Additionally, future studies may also explore the use of ten segmental pathological features as mentioned in [2].

7. REFERENCES

[1] S. A. Arief, "Deteksi suara palsu berbahasa Indonesia berbasis convolutional neural network," UPT Perpustakaan ITB, 2024.

[2] A. Chaiwongyen, S. Duangpummet, J. Karnjana, W. Kongprawechnon, and M. Unoki, "Potential of speech-pathological features for deepfake speech detection," IEEE Access, vol. 12, pp. 121958–121970, 2024. [Online]. Available: https://doi.org/10.1109/ACCESS.2024.3447582

[3] H. Delgado et al., "ASVspoof 5 evaluation plan," ASVspoof Consortium, 2024. [Online]. Available: http://www.asvspoof.org/

[4] DupDub, "About DupDub," [Online]. Available: https://www.dupdub.com/about

[5] ElevenLabs, "ElevenLabs documentation," [Online]. Available: https://www.elevenlabs.io

[6] Google Cloud, "Text-to-speech documentation," [Online]. Available: https://cloud.google.com/text-to-speech

[7] L. Jingyi, T. Weiping, and X. Li, "FreeVC: Towards high-quality text-free one-shot voice conversion," arXiv preprint arXiv:2210.15418, 2022. [Online]. Available: https://arxiv.org/pdf/2210.15418

[8] Jungjee, “RawNet,” GitHub repository, 2024. [Online]. Available: https://github.com/Jungjee/RawNet

[9] J.-w. Jung, H.-S. Heo, J.-h. Kim, H.-j. Shim, and H.-J. Yu, “RawNet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification,” arXiv preprint arXiv:1904.08104, 2019. [Online]. Available: https://arxiv.org/abs/1904.08104

[10] J.-w. Jung, S.-b. Kim, H.-j. Shim, J.-h. Kim, and H.-J. Yu, “Improved RawNet with feature map scaling for text-independent speaker verification using raw waveforms,” arXiv preprint arXiv:2004.00526, 2020. [Online]. Available: https://arxiv.org/abs/2004.00526

[11] J.-w. Jung, Y. J. Kim, H.-S. Heo, B.-J. Lee, Y. Kwon, and J. S. Chung, “Pushing the limits of raw waveform speaker recognition,” arXiv preprint arXiv:2203.08488, 2022. [Online]. Available: https://arxiv.org/abs/2203.08488

[12] J. Kim, J. Kong, and J. Son, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” arXiv preprint arXiv:2106.06103, 2021. [Online]. Available: https://arxiv.org/abs/2106.06103

[13] 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, 2022. [Online]. Available: https://doi.org/10.1007/s10772-021-09876-2

[14] N. M. Müller et al., “MLAAD: The multi-language audio anti-spoofing dataset,” arXiv preprint arXiv:2401.09512, 2024. [Online]. Available: http://arxiv.org/abs/2401.09512

[15] V. Pratap, Q. Xu, A. Sriram, et al., “Massively multilingual speech,” arXiv preprint arXiv:2305.13516, 2023. [Online]. Available: https://arxiv.org/pdf/2305.13516

[16] D. Salvati, C. Drioli, and G. L. Foresti, “A late fusion deep neural network for robust speaker identification using raw waveforms and gammatone cepstral coefficients,” Expert Syst. Appl., vol. 222, p. 119750, 2023. [Online]. Available: https://doi.org/10.1016/j.eswa.2023.119750

[17] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, “End-to-end anti-spoofing with rawnet2,” in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2021. [Online]. Available: https://arxiv.org/abs/2011.01108

[18] 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 Appl., vol. 80, no. 21–23, pp. 32725–32762, 2021. [Online]. Available: https://doi.org/10.1007/s11042-021-11235-x

[19] Y. Zhang, F. Jiang, and Z. Duan, “One-class learning towards synthetic voice spoofing detection,” IEEE Signal Process. Lett., vol. 28, pp. 937–941, 2021.