Deepfake-speech Detection with Pathological Features and Multilayer Perceptron Neural Network

Anuwat Chaiwongyen*†, Suradej Duangpummet‡,
Jessada Karnjana‡, Waree Kongprawechnon†, and Masashi Unoki*
* Japan Advanced Institute of Science and Technology, Ishikawa, Japan
† Sirindhorn International Institute of Technology, Thammasat University, Pathumthani, Thailand
‡ NECTEC, National Science and Technology Development Agency, Pathumthani, Thailand
E-mail: {anuwat, unoki}@jaist.ac.jp, {suradej.dua, jessada.kar}@nectec.or.th, waree@siit.tu.ac.th

Abstract—**Deepfake speech, a misuse of speech technology, is of great concern since it seems natural and is difficult to detect. Although many methods using various speech features have been proposed, deepfake-speech detection accuracy must be improved, especially in real-world scenarios. Therefore, this paper presents a method for detecting deepfake speech on the basis of pathological features used by pathologists for assessing voice quality. The six-pathological features, including jitter, shimmer, harmonics-to-noise ratio, cepstral-harmonics-to-noise ratio, normalized noise energy, and glottal-to-noise excitation ratio, are fed to a multilayer perceptron neural network. We evaluated the proposed method using the Audio Deep Synthesis Detection Challenge dataset. The results indicate that the proposed model can be used for detecting deepfake speech. The proposed method's accuracy, precision, recall, and F1-score were over 98% on the development set, and it outperformed the baseline method on the adaptation set.**

I. Introduction

Deepfake speech is a misuse of speech technologies, such as voice conversion or text-to-speech techniques [1], [2], to synthesize fake speech. Since deepfake speech seems natural and is challenging to detect, it poses a significant threat to economies and societies. For example, criminals used deepfake speech to impersonate a CEO's voice and successfully swindled over USD $243,000$ [3]. Automatic speaker verification to authenticate personal voice is also vulnerable to such attack [4].

Several methods have been proposed to detect deepfake speech [5]–[9]. Many classifiers have been utilized, such as the Gaussian mixture model (GMM), deep neural networks (DNNs) [10], recurrent neural networks (RNNs) [11], convolution neural networks (CNN) [12], and residual neural network (ResNet) [13]. Also, many speech features, including spectrograms, linear-frequency cepstral coefficients (LFCCs) [14], mel-frequency cepstral coefficients [15], constant-Q transform [16], and constant-Q cepstral coefficients [17], have been used. For example, Yi [18] and Wang [19] independently proposed a method based on GMM that takes LFCCs as the input feature [20].

These features are represented in phase, power spectrum, and cepstral coefficients. However, to the best of our knowledge, there is a lack of attention on pathological features in deepfake-speech detection. Since pathological features are widely used to analyze voice disorders, we hypothesize that the voice quality of deepfake speech can be considered as a disordered voice. Hence, pathological features might be clues for deepfake-speech detection.

We, therefore, investigated six pathological features, including jitter, shimmer, harmonics-to-noise ratio (HNR), cepstral-harmonics-to-noise ratio (CHNR), normalized noise energy (NNE), and glottal-to-noise excitation ratio (GNE). These features are fed to a classifier for detecting deepfake speech.

The rest of this paper is organized as follows. Section 2 briefly describes the pathological features mentioned above. Section 3 presents the proposed method. Sections 4 and 5 present the evaluation of the proposed method, results, and discussion. Finally, Section 6 summarizes this work.

II. Pathological Features

Pathological features can be used to distinguish between normal and pathological voices [21] and diagnose diseases such as Parkinson's disease [22] neck and head cancers [23]. This study investigated whether the following pathological features can be used to recognize the deepfake-speech signal.

A. Jitter features

Jitter is the measure of the cycle-to-cycle variations of the fundamental frequency [24], [25]. As the characteristics of jitter can be identified by several methods, this work focuses on four types as follows.

1) Jitter (local): Jitter ($local$) is the average absolute difference between consecutive periods divided by the average period, that is:

$$\text{Jitter } (local) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100, \qquad (1)$$

where T_i represents the extracted f_0 period lengths, and N is the number of extracted f_0 periods [25].

2) Jitter (rap): Jitter (rap) is the average absolute difference between a period of its average and its two neighbors, divided by the average period. It is defined as:

$$\text{Jitter } (rap) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |T_i - (\frac{1}{3}\sum_{i=i-1}^{i+1} T_i)|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100. \quad (2)$$

3) Jitter (ppq5): Jitter ($ppq5$) is the average absolute difference between a period and its average and its five closest neighbors, divided by the average period. It is defined as:

$$\text{Jitter } (ppq5) = \frac{\frac{1}{N-1}\sum_{i=2}^{N-2} |T_i - (\frac{1}{5}\sum_{i=i-2}^{i+2} T_i)|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100. \quad (3)$$

4) Jitter (ppq55): Jitter ($ppq55$) is the average absolute difference between a period and its average and its 55 closest neighbors, divided by the average period. It is defined as:

$$\text{Jitter } (ppq55) = \frac{\frac{1}{N-1}\sum_{i=27}^{N-27} |T_i - (\frac{1}{55}\sum_{i=i-27}^{i+27} T_i)|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100. \quad (4)$$

B. Shimmer features

Shimmer is a variation of a signal in amplitude that results from irregular vocal fold vibrations. There are various ways to identify shimmer characteristics. We focused on two types of shimmer features.

1) Shim (local): Shim ($local$) refers to the average of absolute differences between the source-signal-amplitude-related in each index (A_i) and its next neighbor (A_{i+1}), divided by the average of the signal amplitudes. It is defined as:

$$\text{Shim } (local) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i}, \quad (5)$$

where N is the number of fundamental frequency periods, and A_i denotes the signal amplitude at index i.

2) Shim (x-point amplitude perturbation quotients): Shim x-point amplitude perturbation quotients (APQx) is also defined similarly as Shim ($local$). However, it considers the absolute difference between the amplitude of each index and its $x-1$ closest neighbors. It is defined as:

$$\text{Shim } (\text{APQ}x) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |A_i - (\frac{1}{x}\sum_{n=i-m}^{i+m} A_n)|}{\frac{1}{N}\sum_{i=1}^{N} A_i}, \quad (6)$$

where $m = \frac{x-1}{2}$.

C. Harmonics-to-noise ratio (HNR)

The HNR is a measure of the proportion of the harmonic and noise components of speech. The noise (ι_{En}) is computed as the energy of the residual produced after subtracting the average waveform from each individual cycle. The harmonic energy (γ_{En}) is determined as the energy of an average waveform of a frame pitch built synchronously around ten consecutive glottal cycles. Hence, this feature requires an earlier f_0 estimation [26]. The HNR is defined as:

$$\text{HNR} = 20 \log \frac{\gamma_{En}}{\iota_{En}}. \quad (7)$$

The HNR is calculated for each frame of analysis. The output HNR is the average of each frame.

D. Cepstral-harmonics-to-noise ratio (CHNR)

The CHNR is used to calculate HNR as the level difference between the total energy of the spectrum and noise energy, with the noise component being considered as the energy that cannot be related to the spectrum of the original signal [26]. The CHNR-calculation process is illustrated in Fig. 1.

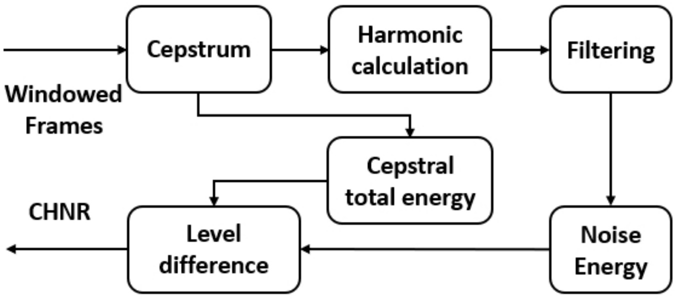

Fig. 1. CHNR calculation process [26].

E. Normalized noise energy (NNE)

NNE is another feature used to quantify the amount of additive noise, and it is defined as the ratio of the energy of the noise to the total energy of the signal for each frame of analysis [26]. The NNE-calculation process is illustrated in Fig. 2.

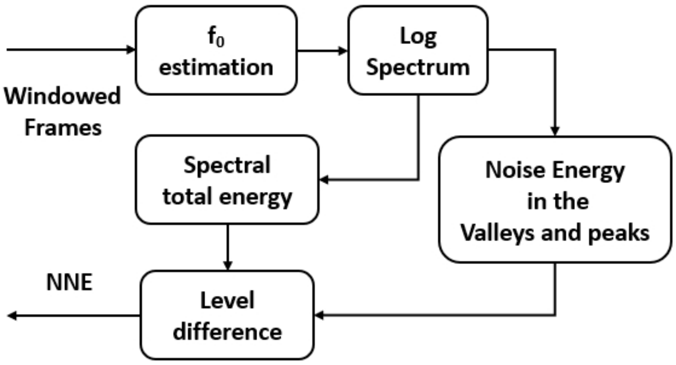

Fig. 2. NNE calculation process [26].

F. Glottal-to-noise excitation ratio (GNE)

The GNE is used to describe turbulent noise while disregarding modulation effects [27]. It is assumed that glottal pulses produce a simultaneous and synchronous excitation of multiple frequency channels, which is indicated by the correlation between Hilbert envelopes of multiple frequency bands [26]. The GNE-calculation process is illustrated in Fig. 3.

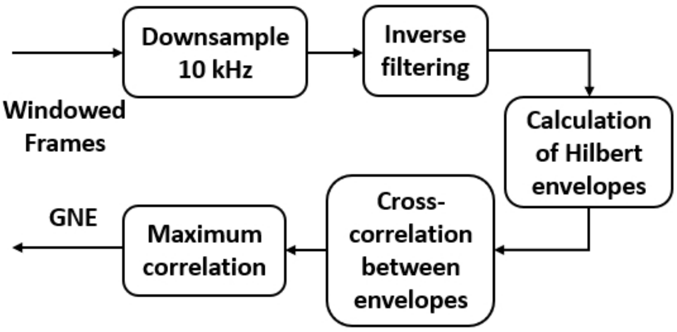

Fig. 3. GNE calculation process [26].

III. Proposed Method

The proposed method detects deepfake speech using a multilayer perceptron (MLP) neural network that takes a combination of the above pathological features as its input, as shown in Fig. 5.

A. Feature analyses

We first investigated those features that can be used to distinguish between genuine and fake speech signals. We conducted an experiment by using the dataset from the Audio Deep Synthesis Detection (ADD) 2022 Challenge [18].

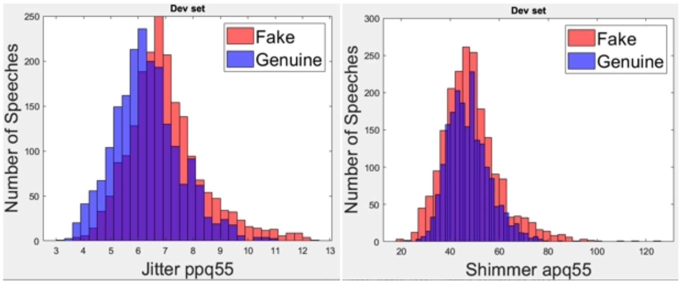

Fig. 4. Histograms of Jitter ($apq55$) and Shim (apq55) in the development set.

Figure 4 shows the histograms of the difference between genuine and fake speech signals for each pathological feature that is not useful in detecting fake speech signals. We randomly selected $2,000$ samples of both genuine and fake speech signals.

Figure 6 shows histograms of the difference between genuine and fake speech signals. for each pathological feature used for detecting fake speech in this paper.

For distinguishing genuine and fake speech, we found that Jitter ($local$), Jitter (rap), Jitter ($ppq5$), CHNR, NNE, and GNE help discriminate between genuine and fake speech signals.

Shim ($local$), Shim (APQ3), Shim (APQ5), Shim (APQ11), and HNR are less effective in discriminating between genuine and fake speech signals, as shown in Fig. 6.

On the other hand, it was found that Jitter ($ppq55$) and Shim (APQ55) are unsuitable for the discrimination because their genuine and fake speech histograms overlap, as shown in Fig. 4.

TABLE I
Number of utterances in ADD 2022 dataset [18]

Dataset	Number of utterances		
	Fake	Genuine	Total
training set	24,072	3,012	27,084
dev. set	26,017	2,307	28,324
adp. set	700	300	1,000

Therefore, we used Jitter ($local$), Jitter (rap), Jitter ($ppq5$), Shim ($local$), Shim (APQ3), Shim (APQ5), Shim (APQ11), HNR, CHNE, GNE, and NNE with an MLP neural network for detecting deepfake speech.

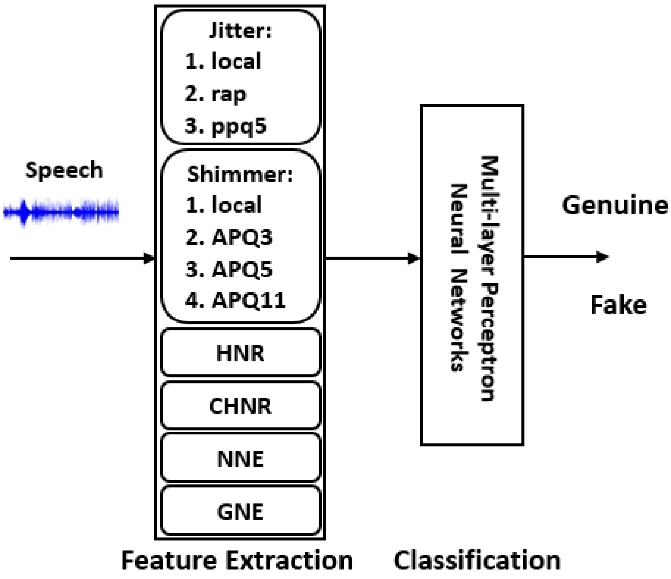

Fig. 5. Proposed method.

B. Experimental setup

The ADD 2022 dataset was used to evaluate the performance of the proposed method. The dataset is divided into three sets: training, development, and adaptation, as shown in Table I. From our observation, both training and development sets have a high signal-to-noise ratio (SNR), whereas the adaptation set has a low SNR with real-world background noise, such as background music and people chatting. In this research, our method was trained from the training set. We then evaluated the proposed method from the development and adaptation sets. For jitter and shimmer extraction, we used the perturbation analysis code implemented in MATLAB [28]. For HNR, CHNR, NNE, and GNE extraction, we used the AVCA-ByO MATLAB toolbox [26].

For classification, we applied the MLP neural network, 3 layers, 11 input features, 11 hidden-layer nodes, and one output layer. The number of training epochs was set to $1,000$. The model used an Adam optimizer, and a learning rate was set to 0.0001. The total number of parameters of our model were only 288.

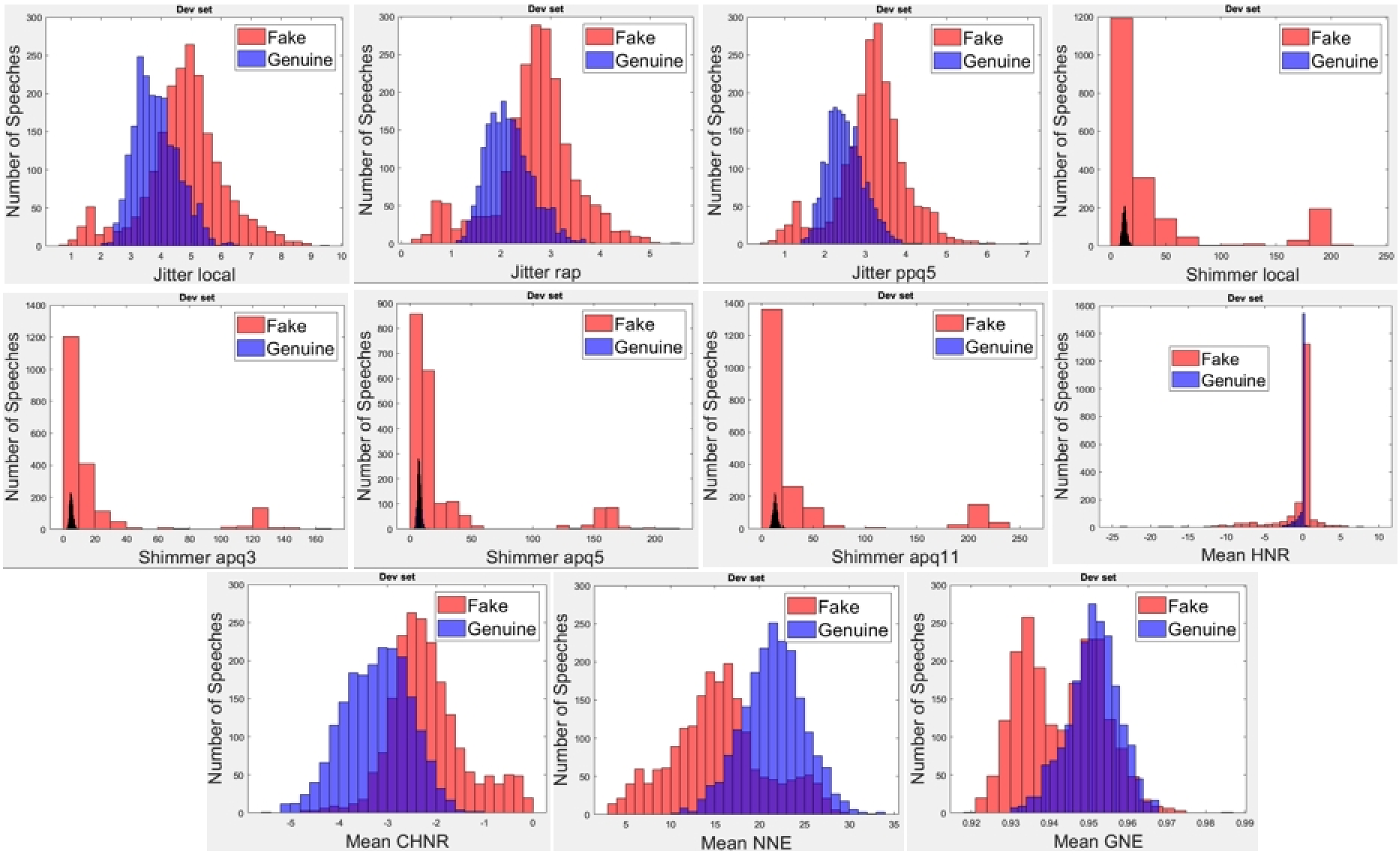

Fig. 6. Histograms of pathological features used in the development set.

IV. Results

We compared our method with two methods using the LFCC feature with two classifiers: GMM and CNN. The method using gammatone cepstral coefficients (GTCCs) with Resnet34 was also compared [29]. The results are shown in Table II. They indicate that even though the accuracy, precision, recall, and F1-score of the proposed method were slightly lower than those of LFCC with GMM, LFCC with CNN, and GTCC with ResNet34 in the development set, the accuracy, recall, and F1-score of our method were better than those of the three other methods in the adaptation sets. It means that the proposed method may be robust against speech signals with background noise, even though the proposed method has never been trained with speech signals with background noise before.

It can be seen from the table that using the pathological features alone is not good in terms of balanced accuracy. However, the proposed method can be improved further by incorporating these pathological features with others, e.g., features obtained from a CNN that takes LFCCs as its input, as shown in Fig. 7. The 11 pathological features and the flattened feature obtained from the CNN are combined and fed to an MLP neural network. The results indicate that the efficiency improved in both the development set and adaptation set, as shown by the fifth method of Table II. It can be concluded that the flattened feature from the CNN with LFCCs may marginally enhance the performance of the proposed method for the clean speech signal but considerably improve the balanced accuracy and precision for the noisy signal. However, such improvement costs a poorer recall.

V. Discussion

Although our method could successfully detect deepfake speech in the development set, the following problems and limitations should be discussed. First, the experimental results indicate that the proposed method could equally efficiently detect deepfake speech in the development set without background noise compared to the other methods regarding the accuracy, precision, recall, and F1 score. However, its balanced accuracy is poorer to some extent. The reason might be that the unbalanced dataset was used to train the proposed model. The investigation of the proposed method to be trained with a balanced dataset will be explored further. Second, the feature size of our proposed method is relatively small in comparison with the other methods. For example, for the same utterance, the size of the pathological features was only 11, whereas these sizes for GTCC with ResNet34, LFCC with CNN, and LFCC with GMM were larger, which are 60×128, 60×128, and 57×100, respectively. Thus, the pathological features have discrimination potential for recognizing the deepfake speech. Suppose they are used well and smartly with other conven-

TABLE II
RESULTS OF PROPOSED METHODS COMPARED WITH EXISTING METHODS

Method	Dev. set evaluation (%)					Adp. set evaluation(%)				
	Accuracy	Precision	Recall	F1-score	Balanced accuracy	Accuracy	Precision	Recall	F1-score	Balanced accuracy
1. LFCC with GMM[18]	**99.99**	**99.99**	**100**	**99.99**	**99.97**	68.80	57.73	96.63	72.04	76.38
2. LFCC with CNN	99.92	99.98	99.94	99.96	99.84	73.00	95.74	64.29	78.81	78.81
3. GTCC with ResNet34	99.96	99.98	99.98	99.98	99.86	65.10	59.71	86.19	70.55	68.69
4. Pathological features with MLP (Proposed method)	98.09	98.47	99.47	98.97	91.02	76.30	75.81	**97.14**	**85.16**	62.40
5. Pathological features, LFCC and CNN features with MLP (Proposed method)	99.76	99.92	99.82	99.87	99.45	**79.90**	**96.46**	74.00	83.75	**83.83**

tional features. In that case, they might reduce misclassification in decision-making in some cases with higher uncertainty. Third, the performances of our proposed method, LFCC with GMM, and GTCC with ResNet34, gradually degraded in noisy environments because all methods, trained from only the training set, were fragile under noisy conditions in the adaptation set. One possible way that we can try to solve this problem is to apply augmentation techniques so that the model can be trained with a noisy version of clean speech. Finally, in this study, we experimentally explore the usefulness of the pathological features by combining them with some conventional features, as shown in the fifth row of Table II. The overall performance is remarkably improved, especially in the balanced accuracy and precision. However, there is a trade-off between those improved aspects and recall. The reason for the recall decrease will be studied further. Also, the features used to combine with the pathological features in this study were chosen to prove the concept. They are not crafted in such a way that they go hand in hand with the pathological features. Thus, feature selection is left from this study and will be done in the future.

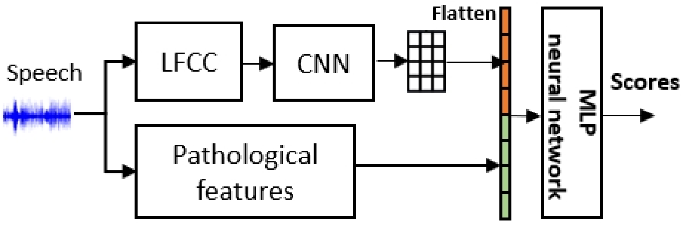

Fig. 7. Structure of proposed method using pathological features with LFCC and CNN feature, and MLP neural networks.

VI. CONCLUSION

This paper highlighted the study of pathological features to detect deepfake speech. We conducted experiments to investigate pathological features of genuine and fake speech. We then proposed a method that uses three jitter features, four shimmer features, HNR, CHNR, NNE, and GNE, for deepfake-speech detection. These pathological features, consisting of only 11 numbers, were classified using an MLP neural network to determine whether the speech was genuine or fake. The proposed method was evaluated on the basis of the ADD 2022 Challenge. The results indicate that the proposed method could effectively detect deepfake speech. Although the efficiency of the proposed method was still lower than the baseline in the development set, it was better than the baseline in the adaptation set. Thus, these pathological features can consistently account for synthetic voices and be used for deepfake-speech detection. In addition, we can improve the performance of the proposed method by combining these pathological features with some conventional ones (e.g., flattened features obtained from a CNN that takes LFCCs as its input) and then classifying the integrated features with an MLP neural network or other classifiers.

ACKNOWLEDGMENT

This work was supported by a Grant-in-Aid for Scientific Fund for the Promotion of Joint International Research (Fostering Joint International Research (B)) (20KK0233), I-O DATA foundation, and KDDI foundation (Research Grant Program). This work was supported by a grant from the SIIT-JAIST-NSTDA Dual Doctoral Degree Program. Also, this work is the output of the ASEAN IVO (http://www.nict.go.jp/en/asean_ivo/index.html) project, titled Spoof Detection for Automatic Speaker Verification, and financially supported by NICT (http://www.nict.go.jp/en/index.html).

REFERENCES

[1] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[2] Y. Ren, Y. Ruan, X. Tan, *et al.*, "Fastspeech: Fast, robust and controllable text to speech," *Advances in neural information processing systems*, vol. 32, 2019.

[3] K. Hartmann and K. Giles, "The next generation of cyber-enabled information warfare," in *2020 12th International Conference on Cyber Conflict (CyCon)*, IEEE, vol. 1300, 2020, pp. 233–250.

[4] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *speech communication*, vol. 66, pp. 130–153, 2015.

[5] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," *arXiv preprint arXiv:1907.00501*, 2019.

[6] M. India, P. Safari, and J. Hernando, "Self multi-head attention for speaker recognition," *arXiv preprint arXiv:1906.09890*, 2019.

[7] R. Yan, C. Wen, S. Zhou, T. Guo, W. Zou, and X. Li, "Audio deepfake detection system with neural stitching for add 2022," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 9226–9230.

[8] Z. Lv, S. Zhang, K. Tang, and P. Hu, "Fake audio detection based on unsupervised pretraining models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9231–9235.

[9] J. M. Martıén-Doñas and A. Álvarez, "The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9241–9245.

[10] S. Duraibi, W. Alhamdani, and F. T. Sheldon, "Replay spoof attack detection using deep neural networks for classification," in *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, IEEE, 2020, pp. 170–174.

[11] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132 306, 2020.

[12] J. Wu, "Introduction to convolutional neural networks," *National Key Lab for Novel Software Technology. Nanjing University. China*, vol. 5, no. 23, p. 495, 2017.

[13] M. Shafiq and Z. Gu, "Deep residual learning for image recognition: A survey," *Applied Sciences*, vol. 12, no. 18, p. 8972, 2022.

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

[15] H. S. Kumbhar and S. U. Bhandari, "Speech emotion recognition using MFCC features and LSTM network," in *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, IEEE, 2019, pp. 1–3.

[16] J. Yang and R. K. Das, "Low frequency frame-wise normalization over constant-Q transform for playback speech detection," *Digital Signal Processing*, vol. 89, pp. 30–39, 2019.

[17] M. Todisco, H. Delgado, and N. W. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients.," in *Odyssey*, vol. 2016, 2016, pp. 283–290.

[18] J. Yi, R. Fu, J. Tao, *et al.*, "Add 2022: The first audio deep synthesis detection challenge," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022.

[19] X. Wang, J. Yamagishi, M. Todisco, *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101 114, 2020.

[20] S. S. Nidhyananthan and R. S. S. Kumari, "Language and text-independent speaker identification system using GMM," *WSEAS Transactions on Signal processing*, vol. 9, no. 4, pp. 185–194, 2013.

[21] A. Sasou, "Automatic identification of pathological voice quality based on the GRBAS categorization," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2017, pp. 1243–1247.

[22] D. Meghraoui, B. Boudraa, T. Merazi-Meksen, and P. G. Vilda, "A novel pre-processing technique in pathologic voice detection: Application to parkinson's disease phonation," *Biomedical Signal Processing and Control*, vol. 68, p. 102 604, 2021.

[23] R. Islam, M. Tarique, and E. Abdel-Raheem, "A survey on signal processing based pathological voice detection techniques," *IEEE Access*, vol. 8, pp. 66 749–66 776, 2020. DOI: 10.1109/ACCESS.2020.2985280.

[24] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and shimmer measurements for speaker recognition," in *8th Annual Conference of the International Speech Communication Association; 2007 Aug. 27-31; Antwerp (Belgium).[place unknown]: ISCA; 2007. p. 778-81.*, International Speech Communication Association (ISCA), 2007.

[25] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal acoustic analysis – jitter, shimmer and HNR parameters," *Procedia Technology*, vol. 9, pp. 1112–1122, 2013, CENTERIS 2013 - Conference on ENTERprise Information Systems / ProjMAN 2013 - International Conference on Project MANagement/ HCIST 2013 - International Conference on Health and Social Care Information Systems and Technologies, ISSN: 2212-0173. DOI: https://doi.org/10.1016/j.protcy.2013.12.124. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212017313002788.

[26] J. Gómez-Garcıéa, L. Moro-Velázquez, J. D. Arias-Londoño, and J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. part iii: Review of acoustic modelling strategies," *Biomedical Signal Processing and Control*, vol. 66, p. 102 049, 2021.

[27] D. Michaelis, T. Gramss, and H. W. Strube, "Glottal-to-noise excitation ratio–a new measure for describing pathological voices," *Acta Acustica united with Acustica*, vol. 83, no. 4, pp. 700–706, 1997.

[28] M. Vashkevich, A. Petrovsky, and Y. Rushkevich, "Bulbar als detection based on analysis of voice perturbation and vibrato," in *2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, 2019, pp. 267–272. DOI: 10.23919/SPA.2019.8936657.

[29] H. Choudhary, D. Sadhya, and V. Patel, "Automatic speaker verification using gammatone frequency cepstral coefficients," in *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, IEEE, 2021, pp. 424–428.