IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 20 June 2024, accepted 15 August 2024, date of publication 22 August 2024, date of current version 10 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3447582

RESEARCH ARTICLE

Potential of Speech-Pathological Features for Deepfake Speech Detection

ANUWAT CHAIWONGYEN[1,2], SURADEJ DUANGPUMMET[3], JESSADA KARNJANA[3], WAREE KONGPRAWECHNON[2], (Member, IEEE), AND MASASHI UNOKI[1], (Member, IEEE)

[1]Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923-1292, Japan
[2]Sirindhorn International Institute of Technology, Thammasat University, Khlong Nueng, Pathum Thani 12120, Thailand
[3]National Electronics and Computer Technology Center (NECTEC), National Science and Technology Development Agency, Khlong Nueng, Pathum Thani 12120, Thailand

Corresponding author: Masashi Unoki (unoki@jaist.ac.jp)

This work was supported in part by the Grant-in-Aid for Scientific Fund for the Promotion of Joint International Research [Fostering Joint International Research (B)] under Grant 20KK0233, in part by the Grant-in-Aid for Transformative Research Areas (A) under Grant 23H04344, in part by the Sirindhorn International Institute of Technology (SIIT)-Japan Advanced Institute of Science and Technology (JAIST)-National Science and Technology Development Agency (NSTDA) Dual Doctoral Degree Program and the Thammasat University Research Fund under Grant TUFT 83/2566, in part by the National Institute of Information and Communications Technology (NICT), and in part by the ICT Virtual Organization of ASEAN Institutes and NICT (ASEAN IVO) project titled 'Spoof Detection for Automatic Speaker Verification'(www.nict.go.jp/en/asean_ivo).

ABSTRACT There is a great concern regarding the misuse of deepfake speech technology to synthesize a real person's voice. Therefore, developing speech-security systems capable of detecting deepfake speech remains paramount in safeguarding against such misuse. Although various speech features and methods have been proposed, their potential for distinguishing between genuine and deepfake speech remains unclear. Since speech-pathological features with deep learning are widely used to assess unnaturalness in disordered voices associated with voice-production mechanisms, we investigated the potential of eleven speech-pathological features for distinguishing between genuine and deepfake speech, i.e., jitter (three types), shimmer (four types), harmonics-to-noise ratio, cepstral-harmonics-to-noise ratio, normalized noise energy, and glottal-to-noise excitation ratio. This paper proposes a method of combining two models on the basis of two different dimensions of speech-pathological features to greatly improve the effectiveness of deepfake speech detection, along with mel-spectrogram features, to enhance detection efficiency. We evaluated the proposed method on the datasets of the Automatic Speaker Verification Spoofing and Countermeasures Challenges ASVspoof 2019 and 2021. The results indicate that the proposed method outperforms the baselines in terms of accuracy, recall, F1-score, and F2-score, achieving 95.06, 99.46, 97.30, and 98.59%, respectively, on the ASVspoof 2019 dataset. It also surpasses the baselines on the ASVspoof 2021 dataset in terms of recall, F1-score, F2-score, and equal error rate, achieving 99.96, 96.65, 98.18, and 15.97%, respectively.

INDEX TERMS Deepfake speech detection, speech-pathological feature, jitter and shimmer, glottal-to-noise, harmonics-to-noise ratio, cepstral-harmonics-to-noise ratio, normalized noise energy.

I. INTRODUCTION

Deepfake speech refers to a synthesized human voice generated using advanced voice conversion and text-to-speech techniques [1], [2]. It finds applications in various domains, such as audiobooks, customer services, and virtual assistants. However, the misuse of deepfake speech poses a significant threat to economies and societies. For instance, criminals have exploited deepfake speech to impersonate a CEO's voice, successfully defrauding over USD 243, 000 [3]. Therefore, detecting deepfake speech is crucial for fraud protection and ensuring the reliability of automatic speaker verification (ASV) systems.

Detecting deepfake speech has involved using several advanced techniques primarily focusing on two approaches: creating efficient classifiers [4], [5], [6] and exploring

The associate editor coordinating the review of this manuscript and approving it for publication was Lin Wang.

acoustic features [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17]. In the first approach, various classifiers have been used, including Gaussian mixture models (GMMs) [18], deep neural networks [19], recurrent neural networks (RNNs) [20], convolution neural networks (CNNs) [21], and residual neural networks (ResNets) [22]. The selection of these classifiers might depend on the characteristics and dimensions of the features. For example, features with small dimensions are suitable for traditional machine-learning models, while those with large dimensions are better handled with deep-learning models, such as CNNs, RNNs, and ResNets [23].

The second approach is focused on using speech and acoustic features as front-end features [24]. Numerous features have been used for detecting deepfake speech, including spectrograms, linear-frequency cepstral coefficients (LFCCs) [25], mel-frequency cepstral coefficients [26], constant-Q transform [27], and constant-Q cepstral coefficients [28]. For example, Yi [29] and Wang [30] independently proposed deepfake-detection methods using LFCCs with GMM. These features are represented in phase, power spectrum, and cepstral coefficients. These features, however, were used without thoroughly clarifying their potential for distinguishing between genuine and deepfake speech.

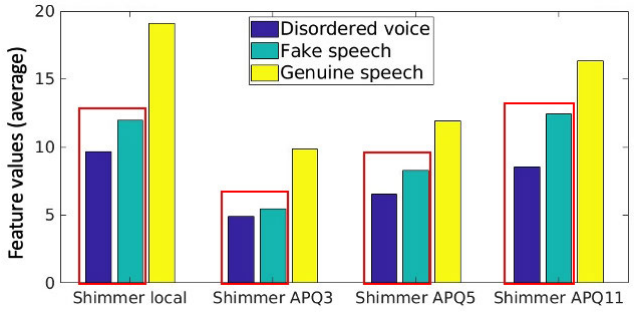

FIGURE 1. Relationship between disordered voice and deepfake speech.

Speech-pathological features, on the other hand, have been introduced to detect the unnatural characteristics of synthesized audio [17], [31]. Speech-pathological features are crucial components closely intertwined with the complex human speech-production mechanisms, representing relevant acoustic, phonatory, and aerodynamic parameters. Speech-language pathologists and otolaryngologists typically use these features to distinguish between normal and disordered voices [32]. In combination with machine-learning algorithms, speech-pathological features are also used in automatic voice assessment and evaluation systems. These systems assist healthcare professionals and medical doctors in classifying, diagnosing, assessing the severity of, and identifying the types of voice disorders [33], [34], [35]. However, the study of pathological features for deepfake speech detection is limited. A method proposed by Kai et al. uses only a few features and is used for fake audio detection [17]. Therefore, a comprehensive investigation of the potential of speech-pathological features in distinguishing between genuine and deepfake speech is necessary.

The motivation for using speech-pathological features to detect deepfake speech is that both fake speech and speech affected by voice disorders can sound different from typical, natural speech. They share common acoustic variations, such as changes in pitch, loudness, and overall quality, which make them sound unnatural.

The relationship of speech-pathological features derived from disordered voice (Hyperkinetic dysarthria) [36], deepfake, and genuine speeches is investigated. Figure 1 shows an example of shimmer features. We can observe that the shimmer (*local*), shimmer (*APQ*3), and shimmer (*APQ*5) exhibit notable distinctions. The feature values of disordered voice and deepfake speech are close to each other, whereas the feature values of genuine speech are different. Therefore, these speech-pathological features, particularly the shimmer features, might be crucial indicators for detecting deepfake speech so that we investigate the potential of the 11 speech-pathological features in more detail in Section II.

This paper focuses on two research questions: whether speech-pathological features can be used to detect deepfake speech and which features contribute to the detection process. To effectively detect such unnaturalness in deepfake-speech signals, we propose a method that combines two models: one based on ten segmental speech-pathological features with their first- and second-order derivatives, denoted as Δ and $\Delta\Delta$, respectively, and the other based on the mel-spectrogram. ResNet-18 models are used as classifiers, outputting the deepfake score through score fusion.

The novelty and main contributions of this paper are as follows:

- The speech-pathological features of jitter, shimmer, harmonics-to-noise ratio (HNR), cepstral-harmonics-to-noise ratio (CHNR), normalized noise energy (NNE), and glottal-to-noise excitation ratio (GNE), have potential to distinguish between genuine and deepfake speech, similar to how medical professionals identify speech disorders in patients.
- The effectiveness of these features to distinguish between genuine and deepfake speech can be enhanced using the segmental frames of analysis technique.
- The proposed method combines two models based on two different dimensions of speech-pathological features to greatly improve the effectiveness of deepfake speech detection.

This paper is organized as follows. Section II briefly introduces speech-pathological features. Section III presents the proposed deepfake-speech-detection method and the analysis of the potential of speech-pathological features for distinguishing between genuine and deepfake speech. Section IV describes the experimental setup and experiments, and Section V presents the evaluation results. Our findings

and limitations are also discussed. Finally, Section VI summarizes the key points of this paper.

II. SPEECH-PATHOLOGICAL FEATURES

Speech-pathological features are typically used to distinguish between normal and pathological voices [37] to diagnose diseases such as Parkinson's disease [38], neck and head cancers [39], and organic pathologies [40]. This section describes deriving the speech-pathological features, which have the potential to be used for distinguishing between genuine and deepfake speech.

A. JITTER FEATURES

Jitter measures the variation in the period from cycle to cycle of a fundamental frequency (F_0) waveform. References [41] and [42], as shown in Fig. 2. Since Jitter can be defined with several methods, we focused on three definitions as follows.

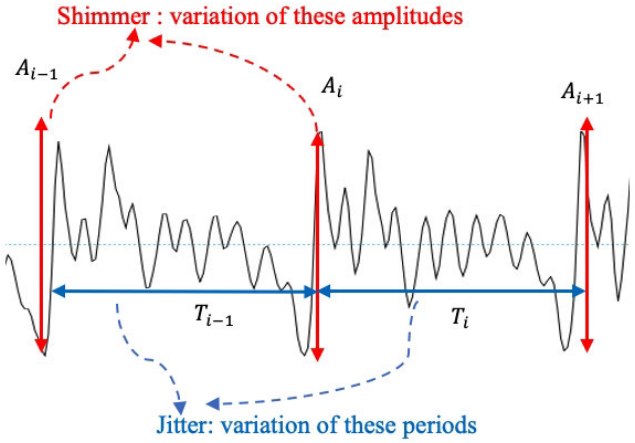

FIGURE 2. Jitter and shimmer concept illustration.

1) JITTER (*local*)

Jitter (*local*) is the percentage of the average absolute difference between consecutive periods divided by the average period, i.e.,

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

where T_i is the period lengths of the extracted F_0, and N is the number of F_0 periods [42].

2) JITTER (*PPQ*3)

Jitter (*PPQ*3), also known as jitter rap, is the percentage of the average absolute difference between a period and the average of that period with its two neighbors divided by the average period. It is defined as [42]:

$$\text{Jitter } (PPQ3) = \frac{\frac{100}{N-1}\sum_{i=1}^{N-1}|T_i - (\frac{1}{3}\sum_{i=i-1}^{i+1} T_i)|}{\frac{1}{N}\sum_{i=1}^{N} T_i}. \tag{2}$$

3) JITTER (*PPQ*5)

Jitter (*PPQ*5) is the percentage of the average absolute difference between a period, and the average of that period with its four neighbors, divided by the average period. It is defined as [42]:

$$\text{Jitter } (PPQ5) = \frac{\frac{100}{N-1}\sum_{i=2}^{N-2}|T_i - (\frac{1}{5}\sum_{i=i-2}^{i+2} T_i)|}{\frac{1}{N}\sum_{i=1}^{N} T_i}. \tag{3}$$

B. SHIMMER FEATURES

Shimmer measures the amplitude variation of a F_0 waveform, resulting from irregular vocal-fold vibrations, as shown in Fig. 2. Jiang et al [43] demonstrated that shimmer has significant differences in speaking styles. This feature can be used to assess the vocal quality and potentially indicate a voice disorder [42]. Since there are various ways to identify shimmer characteristics, we focused on the following two types of shimmer.

1) SHIMMER (*local*)

Shimmer (*local*) refers to the percentages of the average of absolute differences between the source-signal amplitude related to each index (A_i) and its next neighbor (A_{i+1}) divided by the average of the signal amplitudes. It is defined as [42]:

$$\text{Shimmer } (local) = \frac{\frac{100}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i}, \tag{4}$$

where N is the number of F_0 periods, and A_i denotes the signal amplitude at index i.

2) SHIMMER (x-POINT AMPLITUDE PERTURBATION QUOTIENTS)

Shimmer x-point amplitude perturbation quotients, shimmer ($APQx$), are defined similarly to shimmer (*local*). However, it takes into account the absolute difference between the amplitude of each index (A_i) and an average of the x-point closest neighbors around A_i. It is defined as [42]:

$$\text{Shimmer } (APQx) = \frac{\frac{100}{N-m+1}\sum_{i=m}^{N-m}|A_i - (\frac{1}{x}\sum_{n=i-m}^{i+m} A_n)|}{\frac{1}{N}\sum_{i=1}^{N} A_i}, \tag{5}$$

where $m = \frac{x-1}{2}$. We investigated three x-point shimmer features: $APQ3$, $APQ5$, and $APQ11$.

C. HARMONICS-TO-NOISE RATIO (HNR)

The HNR is a metric that quantifies the balance between the harmonic and noisy elements present in speech. Calculating the noise component (ι_{En}) involves computing the energy of the residual signal obtained by subtracting the average waveform from each cycle. The harmonic energy (γ_{En}) is derived from the energy of an average waveform created from a frame pitch that is synchronized with approximately ten consecutive glottal cycles. Therefore, this feature relies on an

earlier estimation of F_0 [44]. HNR is defined as:

$$\text{HNR} = 20 \log \frac{\gamma_{En}}{\iota_{En}}. \quad (6)$$

D. CEPSTRAL-HARMONICS-TO-NOISE RATIO (CHNR)

The CHNR is used to compute the HNR by quantifying the disparity in energy levels between the overall spectrum and energy attributed to noise. In this context, noise energy represents the portion of energy that cannot be attributed to the original signal's spectrum [44]. The CHNR calculation procedure is illustrated in Fig. 3.

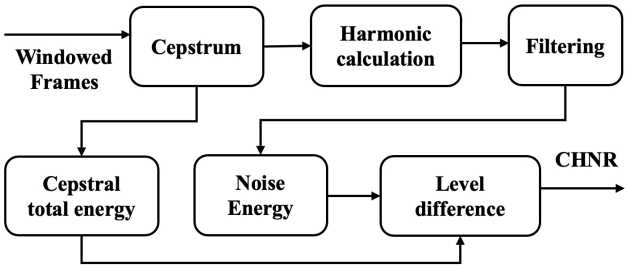

FIGURE 3. CHNR calculation procedure [44].

E. NORMALIZED NOISE ENERGY (NNE)

NNE measures the extended additive noise and is determined by comparing the noise's energy to the overall energy of the signal within each analyzed frame [44]. The NNE calculation procedure is illustrated in Fig. 4.

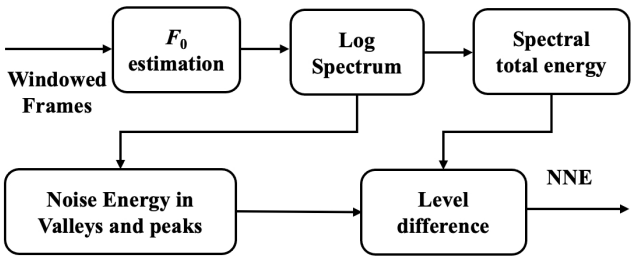

FIGURE 4. NNE calculation procedure [44].

F. GLOTTAL-TO-NOISE EXCITATION RATIO (GNE)

The GNE characterizes turbulent noise in speech, disregarding modulation effects [45]. Glottal pulses are assumed to generate simultaneous and synchronous excitation across multiple frequency channels, as evidenced by the correlation observed in the Hilbert envelopes of these distinct frequency bands [44]. The calculation procedure of the GNE is illustrated in Fig. 5.

G. MEL-SPECTROGRAM

The Mel-spectrogram is widely used in various types of speech-signal processing such as speech recognition and speaker identification [46], [47], [48]. It has also been used for detecting pathological voices [49], [50]. The

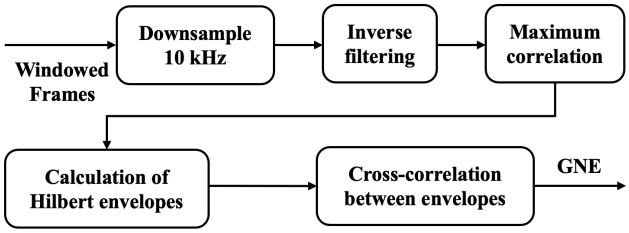

FIGURE 5. GNE calculation procedure [44].

mel-spectrogram is derived through the following steps [51]. An input speech signal is divided into short, overlapping windows. A fast Fourier transform is then applied to convert a time-domain signal into a frequency spectrum. Finally, the mel-frequency filter bank is used to convert a linear frequency scale into the mel-frequency scale. The mel-spectrogram calculation procedure is illustrated in Fig. 6.

FIGURE 6. Mel-spectrogram calculation procedure [44].

III. DEEPFAKE SPEECH DETECTION BASED ON SPEECH-PATHOLOGICAL FEATURES

A. FEATURE ANALYSES

Two preliminary studies are conducted to investigate the potential of speech-pathological features for distinguishing between genuine and deepfake speech. First, we analyze the fundamental effectiveness of each pathological feature on the basis of only their average values. Those speech-pathological features are then incorporated into a basic classifier, which is a multi-layer perceptron neural network, as shown in Fig. 8.

Jitter and shimmer are first derived using the instantaneous robust algorithm for pitch tracking (IRAPT) [52], while the HNR, CHNR, NNE, and GNE are extracted using the AVCA-ByO toolbox [44]. We randomly select 1,000 genuine and 1,000 fake speech signals from the 2019 version of the ASVspoof dataset, as detailed in Table 1. The speech signals are set to 4 s, with a sample rate of 16 kHz. Note that to ensure all signals are 4 s long, signals shorter than 4 s are repeated from the beginning, whereas signals longer than 4 s are truncated.

Figure 7 shows the box plots illustrating the differences of each pathological feature in distinguishing between genuine and fake speech signals. The line inside each box represents the median of each feature. The box itself represents the interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3). The bottom of the box represents Q1, and the top represents Q3 of each feature. The dashed lines extend from the top and bottom of the box to indicate the range of each feature, extending 1.5 times the

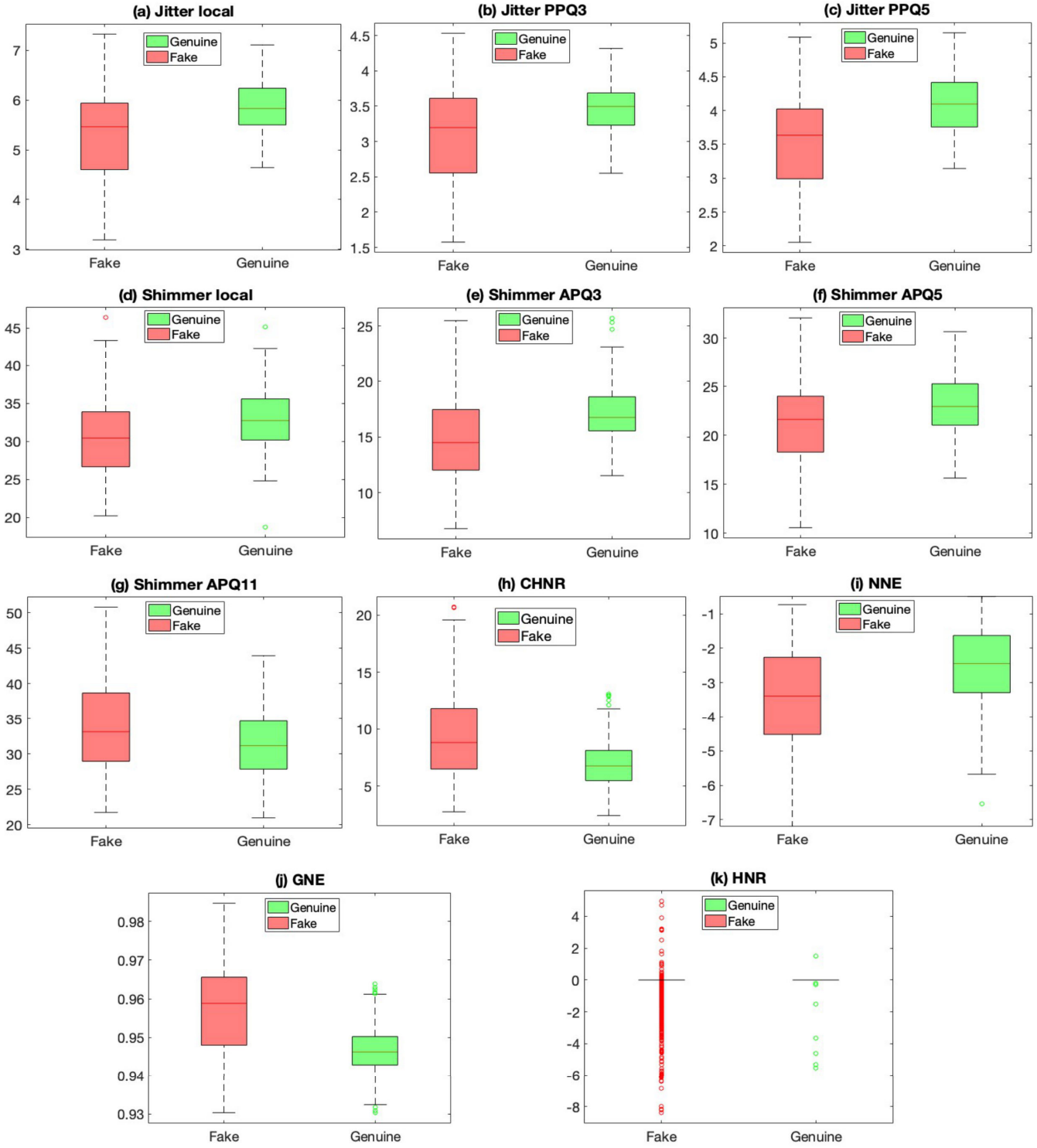

FIGURE 7. Box plots of speech-pathological features derived from 1,000 signals of both genuine (green) and fake (red) speech: (a) jitter (*local*), (b) jitter (*PPQ*3), (c) jitter (*PPQ*3), (d) shimmer (*local*), (e) shimmer (*APQ*3), (f) shimmer (*APQ*3), (g) (APQ11), (h) CHNR, (i) NNE, (j) GNE, and (k) HNR.

IQR from the top and bottom of the box. Data points beyond these dashed lines are considered outliers for each feature.

The results suggest that Jitter (*PPQ*5), the CHNR, and GNE are useful for distinguishing between genuine and deepfake speech signals. Jitter (*local*), jitter (*PPQ*3), jitter (*PPQ*5), shimmer (*local*), shimmer (*APQ*3), shimmer (*APQ*5), shimmer (*APQ*11), and NNE are less effective in distinguishing between genuine and deepfake speech signals. The HNR may be unsuitable for this because the median of both fake and genuine signals is mostly zero, and their distribution does not adhere to a normal distribution.

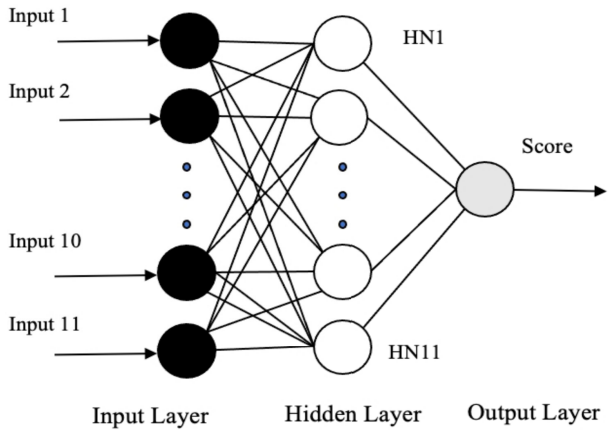

FIGURE 8. Multi-layer perceptron neural network.

A neural network is then used as a classifier. The structure of the classifier comprises one node in the input layer, one node in the hidden layer, and one node in the output layer. The hidden layer is activated with the ReLU function, and the sigmoid function is the activation function in the output layer. The training configurations of the classifier consisted of a maximum of 100 epochs, learning rate of 0.0001, and batch size of 128. The loss function was binary cross-entropy, and the Adam optimizer was used. The training set from the ASVspoof 2019 dataset was used for training, while the development set was used for evaluation.

TABLE 1. Number of utterances in the ASVspoof 2019 and ASVspoof 2021 datasets [24], [53].

Dataset		Number of utterances		
		Genuine	spoofed	Total
ASVspoof 2019	Training	2,580	22,800	25,380
	Development	2,548	22,296	24,844
	Evaluation	7,355	64,578	71,933
ASVspoof 2021	Evaluation	18,452	163,114	181,566

Table 2 presents the effectiveness of the speech-pathological features when used with the neural network. In addition to basic evaluation metrics such as accuracy, recall, and F1-score, an F2-score was also used. The dataset exhibits a high imbalance, where the positive class (spoofed or fake speech) is dominant. For the deepfake-detection task, our aim was to correctly detect as many positive samples as possible rather than solely maximizing the number of correct classifications. The F2-score is appropriate for this scenario since reducing the false negative rate is more important than reducing the false positive rate.

The results suggest that speech-pathological features, excluding the HNR, have the potential to effectively detect deepfake speech. Shimmer (*APQ*3) and GNE are particularly notable, while the HNR performs the worst. The last two rows of Table 2 show the comparison between using a combination of speech-pathological features with and without the HNR. These results indicate that using 10 speech-pathological features (without the HNR) outperforms the use of all 11 features. This method provides 89.94% accuracy and scores greater than 90% for precision, recall, F1-score, and F2-score. Thus, we conclude that only ten pathological features, excluding the HNR, with a simple classifier can effectively detect deepfake speech.

B. SEGMENTAL FRAMES OF ANALYSIS

Although the average of speech-pathological features has the potential to distinguish between genuine and deepfake speech, it might be inadequate. For instance, if the disparity between genuine and fake speech lies in consistency, with approximately 70%–80% consistency, while the remaining portions of the speech exhibit significant fluctuations, the average between genuine and fake speech becomes inconsequential. Therefore, instead of deriving the average of speech-pathological features from the speech signal as the conventional does, we propose deriving speech-pathological features by using the segmented frames of analysis technique.

The process of deriving speech-pathological features using segmented frames of analysis is illustrated in Fig. 9. The process involves receiving a speech signal and segmenting it into frames. The speech-pathological feature is then extracted from each frame. This derivation process starts from the first frame to the M-th frame. Thus, the number of speech-pathological features depends on the number of frames.

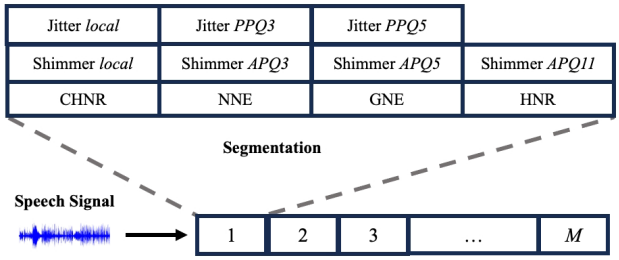

FIGURE 9. Segmental frames of analysis of speech-pathological features.

We evaluated the effectiveness of applying segmented frames of analysis for speech-pathological features. Each feature is derived on a frame-by-frame basis, using a window frame of 50 ms with an overlap of 25 ms. Thus, for a 4-s signal with a sampling rate of 16 kHz, each frame contains 800 samples of data. The data in frame 1 is fed into the feature extraction process, where feature values are obtained and stored in an array at position 1. This feature extraction process is then repeated for frames 2, 3, and so forth, until the last frame, which in this case is the 159th frame. Consequently, for each feature, the data for a single voice is located at position 159. The total number of segmented frames depends on the duration of the speech and the window frame size. These features are inputted into a neural network similar to the previous study. The classifier model consists of three layers: an input layer with 159 nodes corresponding to the new dimension of the feature, a hidden layer with 159 nodes,

TABLE 2. Results from applying an average of speech-pathological features with a neural network on the development set of ASVspoof 2019.

Speech-pathological features	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)
Jitter (*local*)	68.19	54.56	90.93	71.70	80.18	74.74
Jitter (*PPQ*3)	61.15	60.28	92.94	61.31	73.93	65.68
Jitter (*PPQ*5)	66.24	55.53	91.24	69.01	78.57	72.54
Shimmer (*local*)	75.33	51.90	90.17	81.38	85.56	83.00
Shimmer (*APQ*3)	85.64	53.17	90.37	94.03	92.16	**93.27**
Shimmer (*APQ*5)	58.16	50.25	89.82	60.20	78.08	65.45
Shimmer (*APQ*11)	**86.81**	52.00	90.13	**95.97**	**92.87**	94.60
CHNR	84.21	54.61	90.67	91.84	91.25	91.61
NNE	73.51	**62.84**	**92.95**	76.26	83.78	79.11
GNE	85.56	59.75	91.73	92.21	91.97	92.11
HNR	21.11	55.92	99.74	12.13	21.63	14.71
Average of 10 features (except HNR)	74.60	61.20	91.10	79.00	84.00	81.20
Combining 10 features (except HNR)	89.94	61.82	92.04	97.20	94.55	96.12
Combining all 11 features	89.17	60.61	91.81	96.54	94.12	95.66

and a single node for the output. The hidden layer is activated with the ReLU function, whereas the output layer is activated with the sigmoid function. The classifier's training settings included up to 100 epochs, a learning rate set to 0.0001, and a batch size of 128. Binary cross-entropy served as the loss function, and the Adam optimizer was used.

Table 3 lists the results of applying segmental frames of analysis with the speech-pathological features. The results indicate that extending the dimensions of ten speech-pathological features, excluding HNR, through segmental frames of analysis significantly improves performance compared with using the average method (as shown in Table 2) as follows: accuracy from 74.60 to 87.79%, recall from 79.00 to 95.70%, F1-score from 84.00 to 93.60%, and F2-score 81.20 to 95.00%.

C. PROPOSED METHOD

Although the ten segmental speech-pathological features are effective for distinguishing between genuine and deepfake speech, there is still room for improvement. Therefore, our method combines two models to enhance the effectiveness of deepfake speech detection: 1) $PF+\Delta+\Delta\Delta$ with ResNet-18, and 2) mel-spectrogram with ResNet-18. These two models are integrated using score fusion.

The proposed method is illustrated in Fig. 10. The method involves using $PF+\Delta+\Delta\Delta$ with the ResNet-18 model as the primary model, while the mel-spectrogram with the ResNet-18 model is the secondary model. If the prediction score from the primary model exceeds a predetermined threshold of 0.5, it is considered the final decision. However, if the score is below the threshold, the final score is determined by averaging the outputs from both the primary and secondary models.

A ResNet [54], used as a classifier, is an effective deep neural network architecture that addresses the vanishing gradient problem, wherein the gradients during backpropagation become excessively small. Numerous studies have leveraged ResNet in audio and speech-signal processing [55], [56], [57], including detecting synthetic speech [14], [58], [59]. The learning of the residual function of the residual block, which incorporates an intermediate input into the output of a sequence of convolutional blocks, is defined as:

$$y = \mathcal{F}(x) + x, \tag{7}$$

where **x** and **y** denote the input and output from the previous layer, respectively, and $\mathcal{F}(x)$ is a component of a CNN comprising several convolutional blocks. Residual blocks are available across multiple layers, ranging from 10 to over 100 layers, with each layer containing a distinct number of residual blocks. However, excessive features were not deemed necessary for this study. Thus, we decided to use 18 residual layers, i.e., the ResNet-18 model, as the classifier.

IV. EXPERIMENTS

The experiments were conducted to analyze the ten segmental speech-pathological features, their first-order derivative, and second-order derivative, denoted as PF, Δ, and $\Delta\Delta$, respectively, and evaluate the efficiency of the proposed method in detecting deepfake speech. The datasets, metrics, and experimental setup are described as follows.

A. DATASETS AND METRICS

We used the datasets of the ASVspoof 2019 [24] and ASVspoof 2021 challenges [60] to evaluate the performance of the proposed method. The ASVspoof is a series of bi-annual, competitive challenges where the goal is to develop countermeasures capable of distinguishing between genuine and spoofed or deepfake speech since 2015. The ASVspoof 2019 is the first edition focusing on countermeasures for logical access related to spoofing attacks on speech synthesized by using text-to-speech and voice conversion techniques. The dataset is divided into three subsets: training set, development set, and evaluation set.

Similarly, the ASVspoof 2021 challenge extends the 2019 challenge. The evaluation set aims to assess the

TABLE 3. Results of using segmental frames of analysis of speech-pathological features with neural networks on development set of ASVspoof 2019.

Speech-pathological feature	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)
Jitter (*local*)	85.95	64.57	92.78	91.46	91.11	91.72
Jitter (*PPQ*3)	85.44	64.13	92.70	90.94	91.81	91.27
Jitter (*PPQ*5)	87.44	59.37	91.60	94.68	93.12	94.05
Shimmer (*local*)	89.84	60.61	91.79	97.39	94.50	96.21
Shimmer (*APQ*3)	90.30	63.21	92.31	97.27	94.73	96.25
Shimmer (*APQ*5)	89.61	60.74	91.83	97.07	94.38	95.97
Shimmer (*APQ*11)	88.24	54.27	90.54	97.03	93.68	95.66
CHNR	**90.93**	**64.69**	92.60	97.70	**95.08**	**96.63**
NNE	88.67	61.57	92.06	95.67	93.80	94.91
GNE	88.62	54.49	89.98	**98.47**	93.95	96.61
HNR	21.16	55.95	**99.74**	12.78	21.70	14.77
Average of 10 features (except HNR)	87.79	60.07	92.00	95.70	93.60	95.00

TABLE 4. Ablation study of segmental frames of analysis of speech-pathological features with ResNet-18 on ASVspoof 2019 dataset.

Excluded Feature (9×159)	Development set (%)							Evaluation set (%)						
	Accuracy	Balanced accuracy	Precision	Recall	F1-score	F2-score	EER	Accuracy	Balanced accuracy	Precision	Recall	F1-score	F2-score	EER
Jitter (*local*)	96.09	82.16	96.10	99.67	97.86	98.95	7.57	93.66	78.47	95.42	97.61	96.50	97.16	**15.15**
Jitter (PPQ3)	96.49	84.49	96.61	99.59	98.07	98.97	6.83	92.94	**70.72**	**93.74**	98.73	96.17	97.69	10.45
Jitter (PPQ5)	95.99	82.42	96.17	99.49	97.80	98.81	7.18	92.81	83.15	96.61	**95.32**	95.96	**95.57**	11.64
Shimmer (*local*)	95.71	**80.68**	**95.80**	99.59	97.66	98.81	7.03	93.27	71.22	93.83	99.01	96.35	97.92	10.97
Shimmer (APQ3)	95.82	81.45	95.97	99.52	97.71	98.79	7.50	93.75	84.01	96.73	96.29	96.51	96.37	10.18
Shimmer (APQ5)	96.55	86.57	97.09	99.13	98.10	98.71	8.16	94.46	81.17	95.98	97.92	96.94	97.53	9.35
Shimmer (APQ11)	95.82	88.78	95.81	99.70	97.72	98.90	6.43	93.06	71.47	93.89	98.68	96.23	97.68	12.37
CHNR	**95.06**	84.88	96.81	**97.70**	**97.26**	**97.53**	**8.43**	**92.35**	79.25	95.71	95.76	**95.74**	95.75	10.88
NNE	95.93	82.42	96.18	99.42	97.78	98.76	8.35	93.45	76.74	95.04	97.80	96.40	97.24	11.54
GNE	96.10	82.00	96.06	99.74	97.87	98.99	6.12	94.28	81.36	96.05	97.64	96.84	97.31	9.80
All Features (10×159)	96.67	85.81	96.60	99.47	98.17	98.95	8.89	91.36	72.99	94.33	96.14	95.23	96.77	11.33

robustness of channel variation of the detection. The statistical information of both ASVspoof 2019 and ASVspoof 2021 datasets is shown in Table 1. The training set was used to train the models, while the development and evaluation sets were used for evaluation.

Various metrics were used to assess the performance of deepfake speech detection, i.e., accuracy, balanced accuracy, precision, recall, F1-score, F2-score, and the equal error rate (EER). Most metrics follow the evaluation as in the two ASVspoof challenges [24], [30].

B. FEATURE EXTRACTION

The segmental speech-pathological features we used for this study were derived using the following methods. For jitter and shimmer, we used the IRAPT algorithm [52]. For the HNR, CHNR, NNE, and GNE, we used the AVCA-ByO toolbox [44]. The segmental speech-pathological features were derived from a speech signal of 4 s with a sampling rate of 16 kHz. The length of window frames was set to 50 ms with an overlap of 25 ms. Consequently, the total frames of a speech signal were 159. The segmental features of ten of the pathological features were concatenated. Hence, the dimension of these features was 10×159. The design of the features is illustrated in Fig. 10.

The mel-spectrogram and LFCC were derived using the *torchaudio library* [61]. The dimensions of the mel-spectrogram are 80×401, while the dimensions of the LFCC are 60×265. The LFCC and its first-order and second-order derivatives were baseline features in ASVspoof 2019 [24].

C. CLASSIFIER

The ResNet-18 models were utilized as a classifier. The training process consisted of 100 epochs, a learning rate of 0.0001, and a batch size of 32. The Adam optimizer was employed. The binary cross-entropy between the predictions and the targets was used as the loss function. The output score was computed using the output of the "fake" node at the last fully connected layer before the sigmoid operation. The ResNet-18 used has a size of 140 MB. For the implementation, the specifications of the computer are as follows: the graphics card is an NVIDIA GeForce RTX 3090 with 24 GB of memory, the CPU is an Intel Core i9-10920X 3.50 GHz, and the system memory is 128 GB. The computational time for feature extraction and prediction is 1.37 seconds per speech sample.

V. RESULTS AND DISCUSSION

Five experiments were conducted to investigate the potential of speech-pathological features and the proposed method for deepfake speech detection. The first three experiments were

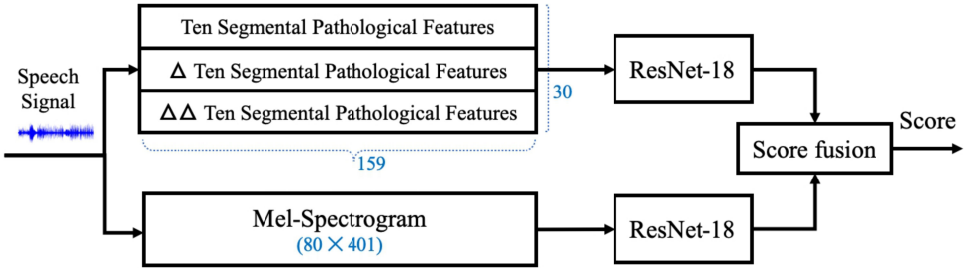

FIGURE 10. The proposed method, combining of (1) ten pathological-segment features with their first-order and second-order derivatives with ResNet-18 and (2) mel-spectrogram with ResNet-18, through score fusion.

presented in Section III. An ablation study and thorough evaluations were also conducted.

The first experiment involved analyzing whether using only the average value of each speech-pathological feature could distinguish between genuine and deepfake speech. The results are presented in Table 2. Our findings suggest that the ten speech-pathological features, excluding the HNR, have the potential to distinguish between genuine and deepfake speech.

The second experiment involved combining the ten speech-pathological features with a simple classifier, i.e., a three-layer neural network. The results suggest that the efficiency of speech-pathological features in almost all evaluation metrics improved, as indicated in Table 2.

The third experiment involved using segmental frames of analysis instead of relying on a single average value for each speech signal. This technique increases the feature dimension, allowing for higher resolution analysis. The comparisons between standard speech-pathological features and those features using the segmental frames of analysis are shown in Table 2 and Table 3, respectively. It was found that applying segmental frames of analysis can significantly enhance the overall efficiency of the speech-pathological features for deepfake speech. Specifically, comparing segmented features with the original calculation, accuracy, recall, F1-score, and F2-score showed substantial improvements, increasing by 13.19, 16.79, 9.60, and 13.80%, respectively. However, the HNR still performed poorly, which was similar to its average performance. Therefore, it was omitted from the features.

The fourth experiment was an ablation study of the proposed features, as shown in Table 4. ResNet-18 was a classifier, and the datasets were the development and evaluation sets of the ASVspoof 2019 dataset. In this study, one speech-pathological feature was removed at a time to assess the importance and potential of each feature for deepfake speech detection. The results of the baselines, which use all speech-pathological features, are presented in the last row. The findings indicate that the CHNR is the most important feature since its removal leads to the lowest performance in terms of accuracy, recall, F1-score, F2-score, and EER on the development set. These trends were also observed in the accuracy and F1-score on the evaluation set.

The fifth and final experiment involved extensive evaluations of the proposed method, as shown is Tables 5 and 6. We evaluated the proposed ten segmental speech-pathological features (PF), the first order derivative of PF (Δ), the second order derivative of PF ($\Delta\Delta$), and the combinations of them. The mel-spectrogram and LFCC were the baseline features.

Table 5 presents the experimental results obtained using the ASVspoof 2019 dataset. When comparing the LFCC and mel-spectrogram on the development set, these two features were comparable. However, on the evaluation set, the mel-spectrogram was better than the LFCC in terms of accuracy 94.36%, recall 96.36%, F1-score 96.84%, F2-score 98.36%, and EER 8.44%, while the LFCC was slightly better only in terms of balanced accuracy 89.94% and precision 98.27%. The reason is that the LFCC correctly detects genuine speech better than the mel-spectrogram but correctly detects deepfake speech less effectively than the mel-spectrogram on an imbalanced dataset. Therefore, the mel-spectrogram showed significantly better results than LFCC. When comparing the efficiency of the mel-spectrogram in both development and evaluation sets, the results were similar, with high accuracy, balanced accuracy, F1-score, and F2-score and low EER.

The third to fifth rows display the results of PF, Δ, and $\Delta\Delta$ with dimensions of 10×159. In the comparison between Δ and $\Delta\Delta$, the results indicate that Δ outperforms $\Delta\Delta$ in almost all metrics on both the development and evaluation sets, except for EER. However, it's important to note that the difference in EER between Δ and $\Delta\Delta$ is less significant. Nonetheless, the method using PF outperforms both methods with Δ and $\Delta\Delta$. Thus, PF is considered to be the most contributing feature among them in terms of performance.

The results of the combinations of the segmental speech-pathological features: $PF + \Delta$, $PF + \Delta\Delta$, and $\Delta + \Delta\Delta$, each with a dimension of 20×159 are presented in the sixth to the eighth rows. The row no. 9 presents the results from the combination of $PF + \Delta + \Delta\Delta$, which has a dimension of 30×159.

When comparing the combination of $PF + \Delta\Delta$ with $PF + \Delta + \Delta\Delta$, the results indicate that $PF + \Delta + \Delta\Delta$ was better in terms of accuracy, recall, F1-score, and F2-score on both datasets. The differences in the rest of the metrics are not

TABLE 5. Comparison of the proposed method with methods using different features and feature combinations on the ASVspoof 2019 dataset.

Method	Development set (%)							Evaluation set (%)						
	Accuracy	Balanced accuracy	Precision	Recall	F1-score	F2-score	EER	Accuracy	Balanced accuracy	Precision	Recall	F1-score	F2-score	EER
1. LFCC (60×265)	**96.86**	85.02	96.69	99.91	**98.27**	99.25	4.57	90.10	**89.94**	**98.27**	90.15	94.23	91.73	10.06
2. Mel-spectrogram (80×401)	96.79	84.41	96.56	**99.99**	98.24	**99.28**	**3.30**	94.36	86.71	97.33	96.36	96.84	98.36	**8.44**
3. *PF* (10×159)	96.67	**85.81**	96.60	99.47	98.17	98.95	8.89	91.36	72.99	94.33	96.14	95.23	96.77	11.33
4. Δ (10×159)	94.47	75.94	94.83	99.26	96.91	98.34	15.10	92.69	72.64	94.17	97.91	96.01	97.14	13.39
5. $\Delta\Delta$ (10×159)	93.63	72.28	94.08	99.14	96.54	98.09	15.37	91.65	70.17	93.67	97.26	95.43	96.52	12.69
6. *PF*+Δ (20×159)	95.81	82.25	96.15	99.31	97.70	98.60	6.78	92.77	72.63	94.16	98.01	96.05	97.21	12.44
7. *PF*+$\Delta\Delta$ (20×159)	96.59	85.03	**96.72**	99.57	98.13	98.99	6.55	93.59	74.65	94.55	98.64	96.65	97.80	10.34
8. Δ+$\Delta\Delta$ (20×159)	93.63	70.80	93.76	99.53	96.56	98.32	9.80	92.48	71.91	94.02	97.82	95.89	97.04	12.86
9. *PF*+Δ+$\Delta\Delta$ (30×159)	96.72	85.05	96.71	99.74	98.20	99.17	8.72	93.96	73.86	94.36	99.19	96.71	98.51	10.22
10. **Proposed method**	96.75	83.82	96.43	**99.99**	98.17	99.25	8.62	**95.06**	77.70	97.30	**99.46**	**97.30**	**98.59**	10.19

TABLE 6. Comparison of results obtained from the proposed method and the baselines on the ASVspoof 2021 dataset.

Method	Evaluation set (%)						
	Accuracy	Balanced accuracy	Precision	Recall	F1-score	F2-score	EER
1. LFCC (60×265)	85.22	**83.44**	**97.58**	85.68	91.24	87.82	16.55
2. Mel-spectrogram (80×401)	92.50	66.37	92.96	99.16	95.96	97.86	20.92
3. *PF*+Δ+$\Delta\Delta$ (30×159)	**92.60**	67.61	93.12	99.09	96.01	97.84	**15.97**
4. **Proposed method**	91.87	59.97	91.69	**99.96**	**96.65**	**98.18**	**15.97**

significant; this is because $PF+\Delta+\Delta\Delta$ has more dimensions than $\Delta + \Delta\Delta$. Among the ten segmental speech-pathological features listed from the third to the ninth rows, $PF + \Delta + \Delta\Delta$ was the most effective at detecting fake speech. The results of the $PF + \Delta + \Delta\Delta$ on the development and evaluation sets are quite similar. Its efficiency was high in terms of accuracy, recall, F1-score, and F2-score. The rest of the metrics were also similar, except for the balanced accuracy, which differed significantly.

In comparison with the LFCC and mel-spectrogram, the findings indicate that $PF + \Delta + \Delta\Delta$ performed better than using the LFCC in terms of accuracy 93.96%, recall 99.19%, F1-score 96.71%, and F2-score 98.51%. Conversely, $PF + \Delta + \Delta\Delta$ marginally underperformed relative to the mel-spectrogram. However, these differences are not statistically significant, as they are less than 1%, except the EER.

These results highlight two interesting aspects: (1) the dimensionality of the features and (2) classification of speech as genuine or synthetic. Since the dimensions of $PF + \Delta + \Delta\Delta$ are relatively small, i.e., 30×159, compared with those of the LFCC and mel-spectrogram, i.e., 60×265 and 80×401, respectively. However, the efficiency of them was comparable. Thus, it might be possible to enhance the performance of the proposed method by extending its resolution, such as reducing the length of window frames. These results also indicate that the mel-spectrogram was more effective for correctly detecting genuine speech, whereas $PF + \Delta + \Delta\Delta$ was more effective for correctly detecting fake speech.

The results on the ASVspoof 2019 evaluation set indicate that the proposed method is comparable in efficiently detecting deepfake speech to the mel-spectrogram in terms of accuracy, recall, F1-score, and F2-score. However, its balanced accuracy exhibits a degree of decline.

Table 6 lists the experimental results on ASVspoof 2021. In the comparison between LFCC and mel-spectrogram, the results indicate that the mel-spectrogram provided better results than LFCC in terms of accuracy 95.50%, recall 99.16%, F1-score 95.96%, and F2-score 97.86%. $PF + \Delta + \Delta\Delta$ slightly outperformed the mel-spectrogram in terms of accuracy 92.60%, balanced accuracy 67.61%, precision 93.12%, F1-score 96.01%, and particularly the EER 15.97%. However, $PF + \Delta + \Delta\Delta$ exhibited only a slight decrease compared with the mel-spectrogram in terms of recall and F2-score.

When $PF+\Delta+\Delta\Delta$ is combined with the mel-spectrogram and ResNet-18, i.e., the proposed method, the results indicate that the performance of the proposed method surpasses that

of the individual components in terms of recall 99.96%, F1-score 96.65%, and F2-score 98.18%. However, balanced accuracy and precision showed a decrease. The reason for this is that both $PF + \Delta + \Delta\Delta$ and the mel-spectrogram exhibited similar characteristics, resulting in high performance in correctly detecting fake speech but lower performance in correctly detecting genuine speech.

When comparing the performance of LFCC and the proposed method in terms of EER, the results showed that both are comparable, with 16.55% and 15.97%, respectively. However, in an in-depth analysis, we found that the false positive rate of LFCC is approximately three times lower than that of the proposed method. In contrast, the false negative rate of the proposed method is significantly lower than that of LFCC, by about sixteen times. This may be due to the high ratio of the imbalanced dataset, with more fake speech than genuine, causing the balanced accuracy of the two models to be significantly different. To improve the efficiency of distinguishing between genuine and fake speech, further studies will be investigated.

Although the proposed method, which combines these two models, did not improve in terms of all metrics, it showed high recall rates. The advantages of high recall are crucial for preventing unauthorized access and impersonation. In tasks involving sensitive scenarios in which unauthorized access carries a significant cost, prioritizing high recall is crucial for deepfake speech detection.

As evident from the third row of Table 6, the accuracy, balanced accuracy, and precision exhibited a slight decrease compared with the results obtained on the ASVspoof 2019 dataset. The effectiveness of $PF + \Delta + \Delta\Delta$ has limitations in detecting synthetic audio in environments involving communication over telephony and Voice over Internet Protocol (VoIP) networks, particularly due to various coding and transmission effects [53]. We will further investigate this scenario.

VI. CONCLUSION

We have proposed a deepfake-detection method that is based on speech-pathological features. We introduced speech-pathological features that are typically used to detect disordered voices for detecting deepfake speech. This is similar to medical professionals diagnosing speech disorders in patients. Eleven speech-pathological features were investigated, including three jitter features: jitter (*local*), jitter (*PPQ*3), jitter (*PPQ*5); four shimmer features: shimmer (*local*), shimmer (*APQ*3), shimmer (*APQ*5), shimmer (*APQ*11); HNR; CHNR; NNE; and GNE. We found that the standard derivation of these features, a single average value for each, effectively distinguished between genuine and deepfake speech. Applying a simple classifier to these speech-pathological features, excluding the HNR, could effectively detect deepfake speech. We introduced segmental frames of analysis for deriving these speech-pathological features to enhance the overall performance. The proposed method combines two models using ten segmental speech-pathological features and the mel-spectrogram for detecting deepfake speech. The proposed method was evaluated and compared with two baselines using datasets from ASVspoof 2019 and 2021. The results indicated that the proposed method outperformed the baselines, achieving a recall of 99.46 and 99.96% on the evaluation sets of ASVspoof 2019 and ASVspoof 2021, respectively.

However, a limitation of this research is the computational time required for feature extraction across all segmental frames during the analysis of speech-pathological features. This takes approximately 1.25 s, whereas it takes only approximately 11.20 ms for mel-spectrograms. In the future, we will evaluate the proposed method on diverse datasets, including those from noisy environments, and investigate speech-pathological features such as sampling rate, frame length, and speech length in more detail. We will also explore additional speech-pathological features.

REFERENCES

[1] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Commun.*, vol. 88, pp. 65–82, Apr. 2017.
[2] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech: Fast, robust and controllable text to speech," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 3171–3180.
[3] K. Hartmann and K. Giles, "The next generation of cyber-enabled information warfare," in *Proc. 12th Int. Conf. Cyber Conflict (CyCon)*, vol. 1300, May 2020, pp. 233–250.
[4] W. Ge, M. Panariello, J. Patino, M. Todisco, and N. Evans, "Partially-connected differentiable architecture search for deepfake and spoofing detection," 2021, *arXiv:2104.03123*.
[5] A. Luo, E. Li, Y. Liu, X. Kang, and Z. J. Wang, "A capsule network based approach for detection of audio spoofing attacks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6359–6363.
[6] Z. Wang, S. Cui, X. Kang, W. Sun, and Z. Li, "Densely connected convolutional network for audio spoofing detection," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2020, pp. 1352–1360.
[7] M. R. Kamble, H. B. Sailor, H. A. Patil, and H. Li, "Advances in anti-spoofing: From the perspective of ASVspoof challenges," *APSIPA Trans. Signal Inf. Process.*, vol. 9, no. 1, p. 2, 2020.
[8] Z. Zhang, X. Yi, and X. Zhao, "Fake speech detection using residual network with transformer encoder," in *Proc. ACM Workshop Inf. Hiding Multimedia Secur.*, Jun. 2021, pp. 13–22.
[9] J. Yang and R. K. Das, "Long-term high frequency features for synthetic speech detection," *Digit. Signal Process.*, vol. 97, Feb. 2020, Art. no. 102622.
[10] K. Zaman, M. Sah, C. Direkoglu, and M. Unoki, "A survey of audio classification using deep learning," *IEEE Access*, vol. 11, pp. 106620–106649, 2023.
[11] K. Li, Y. Wang, M. L. Nguyen, M. Akagi, and M. Unoki, "Analysis of amplitude and frequency perturbation in the voice for fake audio detection," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2022, pp. 929–936.
[12] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," 2019, *arXiv:1907.00501*.
[13] M. India, P. Safari, and J. Hernando, "Self multi-head attention for speaker recognition," 2019, *arXiv:1906.09890*.
[14] R. Yan, C. Wen, S. Zhou, T. Guo, W. Zou, and X. Li, "Audio deepfake detection system with neural stitching for ADD 2022," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 9226–9230.
[15] Z. Lv, S. Zhang, K. Tang, and P. Hu, "Fake audio detection based on unsupervised pretraining models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 9231–9235.

[16] J. M. Martín-Doñas and A. Álvarez, "The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 ADD challenge," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 9241–9245.

[17] K. Li, X. Lu, M. Akagi, and M. Unoki, "Contributions of jitter and shimmer in the voice for fake audio detection," *IEEE Access*, vol. 11, pp. 84689–84698, 2023.

[18] Y. Wang, W. Chen, J. Zhang, T. Dong, G. Shan, and X. Chi, "Efficient volume exploration using the Gaussian mixture model," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 11, pp. 1560–1573, Nov. 2011.

[19] S. Duraibi, W. Alhamdani, and F. T. Sheldon, "Replay spoof attack detection using deep neural networks for classification," in *Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, Dec. 2020, pp. 170–174.

[20] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Phys. D, Nonlinear Phenomena*, vol. 404, Mar. 2020, Art. no. 132306.

[21] J. Wu, "Introduction to convolutional neural networks," *Nat. Key Lab Novel Softw. Technol. Nanjing Univ. China*, vol. 5, no. 23, p. 495, 2017.

[22] M. Shafiq and Z. Gu, "Deep residual learning for image recognition: A survey," *Appl. Sci.*, vol. 12, no. 18, p. 8972, Sep. 2022.

[23] Z. Almutairi and H. Elgibreen, "A review of modern audio deepfake detection methods: Challenges and future directions," *Algorithms*, vol. 15, no. 5, p. 155, May 2022.

[24] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "ASVspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 3, no. 2, pp. 252–265, Apr. 2021.

[25] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Sep. 2015, pp. 2087–2091.

[26] H. S. Kumbhar and S. U. Bhandari, "Speech emotion recognition using MFCC features and LSTM network," in *Proc. 5th Int. Conf. Comput., Commun., Control Autom. (ICCUBEA)*, Sep. 2019, pp. 1–3.

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

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

[29] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, S. Liang, S. Wang, S. Zhang, X. Yan, L. Xu, Z. Wen, and H. Li, "ADD 2022: The first audio deep synthesis detection challenge," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 9216–9220.

[30] X. Wang et al., "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Comput. Speech Lang.*, vol. 64, Nov. 2020, Art. no. 101114.

[31] K. Kuligowska, P. Kisielewicz, and A. Włodarz, "Speech synthesis systems: Disadvantages and limitations," *Int. J. Eng. Technol.*, vol. 7, no. 2, pp. 234–239, May 2018.

[32] V. Dellwo, M. Huckvale, and M. Ashby, "How is individuality expressed in voice? An introduction to speech production and description for speaker classification," in *Speaker Classification I: Fundamentals, Features, and Methods*. Berlin, Germany: Springer, 2007, pp. 1–20.

[33] S.-H. Fang, Y. Tsao, M.-J. Hsiao, J.-Y. Chen, Y.-H. Lai, F.-C. Lin, and C.-T. Wang, "Detection of pathological voice using cepstrum vectors: A deep learning approach," *J. Voice*, vol. 33, no. 5, pp. 634–641, Sep. 2019.

[34] Z. Xie, C. Gadepalli, F. Jalalinajafabadi, B. M. G. Cheetham, and J. J. Homer, "Machine learning applied to GRBAS voice quality assessment," *Adv. Sci., Technol. Eng. Syst. J.*, vol. 3, no. 6, pp. 329–338, 2018.

[35] T. Kojima, S. Fujimura, K. Hasebe, Y. Okanoue, O. Shuya, R. Yuki, K. Shoji, R. Hori, Y. Kishimoto, and K. Omori, "Objective assessment of pathological voice using artificial intelligence based on the GRBAS scale," *J. Voice*, vol. 38, no. 3, pp. 561–566, May 2024.

[36] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000.

[37] A. Sasou, "Automatic identification of pathological voice quality based on the GRBAS categorization," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2017, pp. 1243–1247.

[38] 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," *Biomed. Signal Process. Control*, vol. 68, Jul. 2021, Art. no. 102604.

[39] R. Islam, M. Tarique, and E. Abdel-Raheem, "A survey on signal processing based pathological voice detection techniques," *IEEE Access*, vol. 8, pp. 66749–66776, 2020.

[40] S. R. Kadiri and P. Alku, "Analysis and detection of pathological voice using glottal source features," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 367–379, Feb. 2020.

[41] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and shimmer measurements for speaker recognition," in *Proc. Int. Speech Commun. Assoc. (ISCA)*, Aug. 2007, pp. 778–781.

[42] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal acoustic analysis—Jitter, shimmer and HNR parameters," *Proc. Technol.*, vol. 9, pp. 1112–1122, Jan. 2013.

[43] J. J. Jiang, D. B. Wexler, I. R. Titze, and S. D. Gray, "Fundamental frequency and amplitude perturbation in reconstructed canine vocal folds," *Ann. Otol., Rhinol. Laryngol.*, vol. 103, no. 2, pp. 145–148, Feb. 1994.

[44] J. A. Gómez-García, 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," *Biomed. Signal Process. Control*, vol. 66, Apr. 2021, Art. no. 102049.

[45] 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.

[46] L. Trinh Van, T. Dao Thi Le, T. Le Xuan, and E. Castelli, "Emotional speech recognition using deep neural networks," *Sensors*, vol. 22, no. 4, p. 1414, Feb. 2022.

[47] X. Wang, F. Xue, W. Wang, and A. Liu, "A network model of speaker identification with new feature extraction methods and asymmetric BLSTM," *Neurocomputing*, vol. 403, pp. 167–181, Aug. 2020.

[48] A. Meghanani, C. S. Anoop, and A. G. Ramakrishnan, "An exploration of log-mel spectrogram and MFCC features for Alzheimer's dementia recognition from spontaneous speech," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Jan. 2021, pp. 670–677.

[49] L. Geng, Y. Liang, H. Shan, Z. Xiao, W. Wang, and M. Wei, "Pathological voice detection and classification based on multimodal transmission network," *J. Voice*, vol. S0892-1997, no. 22, Dec. 2022, Art. no. 00370-8, doi: 10.1016/j.jvoice.2022.11.018.

[50] X. Peng, H. Xu, J. Liu, J. Wang, and C. He, "Voice disorder classification using convolutional neural network based on deep transfer learning," *Sci. Rep.*, vol. 13, no. 1, p. 7264, May 2023.

[51] L. Rabiner and R. Schafer, *Theory and Applications of Digital Speech Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 2010.

[52] M. Vashkevich, A. Petrovsky, and Y. Rushkevich, "Bulbar ALS detection based on analysis of voice perturbation and vibrato," in *Proc. Signal Process., Algorithms, Architectures, Arrangements, Appl. (SPA)*, Sep. 2019, pp. 267–272.

[53] H. Delgado, N. Evans, T. Kinnunen, K. A. Lee, X. Liu, A. Nautsch, J. Patino, M. Sahidullah, M. Todisco, X. Wang, and J. Yamagishi, "ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," 2021, *arXiv:2109.00535*.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[55] M. Yousefi and J. H. L. Hansen, "Speaker conditioning of acoustic models using affine transformation for multi-speaker speech recognition," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2021, pp. 283–288.

[56] M. B. Shafeen and S. Vijayan, "Parkinson's disease prognosis using the ResNet-50 model from speech features," in *Proc. Int. Conf. Innov. Sci. Technol. Sustain. Develop. (ICISTSD)*, Aug. 2022, pp. 282–286.

[57] A. Kumar, S. S. Mahmoud, Y. Wang, S. Faisal, and Q. Fang, "A comparison of time-frequency distributions for deep learning-based speech assessment of aphasic patients," in *Proc. 15th Int. Conf. Human Syst. Interact. (HSI)*, Jul. 2022, pp. 1–5.

[58] M. H. Rahman, M. Graciarena, D. Castan, C. Cobo-Kroenke, M. McLaren, and A. Lawson, "Detecting synthetic speech manipulation in real audio recordings," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2022, pp. 1–6.

[59] Z. Chen, Z. Xie, W. Zhang, and X. Xu, "ResNet and model fusion for automatic spoofing detection," in *Proc. Interspeech*, Aug. 2017, pp. 102–106.
[60] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, "ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection," 2021, *arXiv:2109.00537*.
[61] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

ANUWAT CHAIWONGYEN received the B.S. degree in technical education (computer engineering) from the Rajamangala University of Technology Lanna, Chiang Mai, Thailand, in 2008, and the M.Eng. degree in computer engineering from Thammasat University, Thailand, in 2017. He is currently pursuing the Ph.D. degree with the Sirindhorn International Institute of Technology (SIIT), Thammasat University, and Japan Advanced Institute of Science and Technology (JAIST), Japan. From 2008 to 2020, he was a Research Assistant with the National Electronics and Computer Technology (NECTEC). His research interests include speech spoofing detection, machine learning, and deep neural networks.

SURADEJ DUANGPUMMET received the B.Eng. degree in mechatronics engineering from the Pathumwan Institute of Technology, in 2002, the M.Eng. degree in mechatronics engineering from the Asian Institute of Technology, Thailand, in 2011, the first Ph.D. degree (Hons.) in information science from Japan Advanced Institute of Science and Technology (JAIST), in 2021, under the supervision of Prof. Masashi Unoki, and the second Ph.D. degree in engineering and technology from the Sirindhorn International Institute of Technology, Thammasat University, in 2023. He has been with the National Electronics and Computer Technology Center (NECTEC), since 2002, where he is currently a Researcher. His research interests include speech and acoustic signal processing, pattern recognition, artificial intelligence, neural networks, and machine learning.

JESSADA KARNJANA received the B.Eng. degree in electronics engineering from the King Mongkut's Institute of Technology Ladkrabang, Thailand, in 1999, the M.Eng. degree in microelectronics from Asian Institute of Technology, Thailand, in 2006, the first Ph.D. degree in information science from Japan Advanced Institute of Science and Technology, Japan, in 2016, and the second Ph.D. degree in engineering and technology from the Sirindhorn International Institute of Technology, Thammasat University, in 2017. He has been with the National Electronics and Computer Technology Center (NECTEC), Thailand, since 1999. His research interests include data analysis and machine learning, reasoning with uncertainty, signal processing, and wireless sensor networks.

WAREE KONGPRAWECHNON (Member, IEEE) received the B.Eng. degree (Hons.) in electrical engineering from Chulalongkorn University, Thailand, the M.Eng. degree in control engineering from Osaka University, Japan, in 1995, and the Ph.D. degree in engineering (mathematical engineering and information physics) from The University of Tokyo, Japan, in 1999. Since 1999, she has been a Lecturer with the School of Information, Computer, and Communication Technology, Sirindhorn International Institute of Technology, Thammasat University, Thailand, where she is currently an Associate Professor. Her research interests include theory in H∞ control, control theory, robust control, system identification, adaptive control, learning control, neural networks, and fuzzy control. Since 2019, she has been a Steering Committee Member of Asian Control Association.

MASASHI UNOKI (Member, IEEE) received the M.S. and Ph.D. degrees in information science from Japan Advanced Institute of Science and Technology (JAIST), in 1996 and 1999, respectively. He was a Japan Society for the Promotion of Science (JSPS) Research Fellow, from 1998 to 2001. He was a Visiting Researcher with the ATR Human Information Processing Laboratories, from 1999 to 2000, and the Centre for the Neural Basis of Hearing (CNBH), Department of Physiology, University of Cambridge, from 2000 to 2001. Since 2001, he has been a Faculty Member with the School of Information Science, JAIST, where he is currently a Full Professor. His current research interests include auditory-motivated signal processing and the modeling of auditory systems. He is a member of the Research Institute of Signal Processing (RISP), the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, the Acoustical Society of America (ASA), the Acoustical Society of Japan (ASJ), and the International Speech Communication Association (ISCA). He received the Sato Prize from ASJ, in 1999, 2010, and 2013, for an Outstanding Paper, and the Yamashita Taro "Young Researcher" Prize from the Yamashita Taro Research Foundation, in 2005.