IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 June 2023, accepted 29 July 2023, date of publication 3 August 2023, date of current version 15 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3301616

RESEARCH ARTICLE

Contributions of Jitter and Shimmer in the Voice for Fake Audio Detection

KAI LI[1], XUGANG LU[2], (Member, IEEE), MASATO AKAGI[1], (Life 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]Advanced Speech Technology Laboratory, National Institute of Information and Communications Technology, Kyoto 619-0289, Japan

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

This work was supported in part by the ASEAN IVO (http://www.nict.go.jp/en/asean_ivo/index.html) Project, titled Spoof Detection for Automatic Speaker Verification; in part by the National Institute of Information and Communications Technology (NICT) (http://www.nict.go.jp/en/index.html); in part by the SCOPE Program of Ministry of Internal Affairs and Communications under Grant 201605002; in part by the KDDI Foundation (Research Grant Program); and in part by the Fund for the Promotion of Joint International Research (Fostering Joint International Research) under Grant 20KK0233.

ABSTRACT Fake audio detection (FAD) aims to identify fraudulent speech generated through advanced speech-synthesis techniques. Most current FAD methods rely solely on a deep neural network (DNN) framework with either speech waveforms or commonly used acoustic features to extract high-level representations, overlooking the analysis of prosody differences between genuine and fake speech. Prosody carries important cues about the naturalness of speech and emotional content, which can be leveraged in the detection of fake audio. This paper explicitly investigates the differences in prosody information between genuine and fake speech represented by the jitter and shimmer features. On the basis of our investigation, we found strong evidence that obvious differences exist in the level of jitter and shimmer between fake and real speech, particularly on the shimmer feature that has a large dynamic variation for fake speech. To ensure accurate estimation of F_0 for better jitter and shimmer feature representations, we propose using two additional F_0 estimation methods, YIN and SWIPE, in place of the IRAPT algorithm in the feature extraction process. Moreover, we design a DNN-FAD system by explicitly combining the shimmer and Mel-spectrogram features. The effectiveness of the proposed method for FAD is evaluated in the datasets of Audio Deep Synthesis Detection (ADD) 2022 and 2023 challenges. The experimental results show that both the static and dynamic continuous shimmer features, especially that extracted with the YIN and SWIPE algorithms, can provide complementary knowledge to the traditional spectrum-based FAD systems. The optimal results effectively reduce the equal error rate from 41.29 % to 35.77 % in the ADD2023 challenge, achieving a relative improvement of 13.37 %.

INDEX TERMS Fake audio detection, amplitude perturbation, frequency perturbation, jitter and shimmer features, prosody information.

I. INTRODUCTION

The primary objective of fake audio detection (FAD) [1] is to identify fraudulent speech generated through advanced voice conversion (VC) [2], [3] or text-to-speech (TTS) [4], [5], [6] technologies. FAD technologies can be used to safeguard automatic speaker verification (ASV) systems from the risks posed by spoofing attacks. In recent years, numerous advanced methods for FAD have emerged. Most of them focus on two aspects, one is focusing on designing effective deep model architectures, and the other is focusing on exploring different types of acoustic features.

In [7], [8], and [9], a light-convolution neural network (LCNN) with a max-feature-map activation function, proposed for face verification [10], was utilized. The advantages of LCNN include a reduced number of trainable parameters and faster computation speed. In addition, the light convolutional gated recurrent neural network (LC-GRNN),

The associate editor coordinating the review of this manuscript and approving it for publication was Manuel Rosa-Zurera.

variants of a squeeze-excitation network (SENet) [11], and ResNet [12] were used as a deep-feature extractor [13] to defend against spoofing attacks. Many more works related to the back-end classifiers have also been reported [14], [15], [16], [17], [18], [19], [20]. Most architectures of DNNs are based on CNN or RNN modules because a CNN can efficiently extract features, while a RNN can effectively detect long-term dependencies in time variances. However, the performance of these methods highly depends on the discrimination of the front-end input.

Concerning acoustic features, traditional methods for extracting front-end features in FAD primarily rely on digital signal processing algorithms to extract spectra, phase, and other acoustic characteristics [7], [21]. Among these, spectrograms [8], linear frequency cepstral coefficients (LFCC) [22], and constant Q cepstrum coefficients (CQCC) [23] are widely used acoustic features. Spectrograms are commonly used as input features in CNN-based classifiers. On the other hand, CQCC features utilize a constant Q transform (CQT) instead of the short-time Fourier transform (STFT) to process speech signals. They have demonstrated superior performance to the commonly used mel-frequency cepstral coefficients (MFCCs) [22], [24].

However, these methods often do not explicitly analyze the distinctions between genuine and fake speech. Usually, the distinctions between the two types of speech stem from the challenging issue of unnaturalness in speech synthesis [25]. Moreover, the unnaturalness in synthesized speech is often caused by the limitations in capturing and reproducing rich and diverse prosody information. Prosody related to representations of non-linguistic information of voice is a key issue for solving the unnaturalness of synthesized speech. Therefore, exploring prosody differences between genuine and fake speech holds great promise in providing discriminative information for FAD.

Prosody refers to the melodic and rhythmic aspects of speech, including variations in pitch, loudness, duration, and intonation [26]. Jitter and shimmer are acoustic measures that provide information about the stability and irregularities in vocal fold vibration and intensity [27], [28]. These measures can be related to prosody because variations in vocal stability and irregularities can affect the melodic aspects of speech, such as pitch and loudness modulation. Previous studies have demonstrated the efficacy of these features in characterizing voices with pathological prosody [29], [30], [31]. It is reasonable to regard jitter and shimmer as valuable features for distinguishing between genuine and fake speech.

The accuracy of fundamental frequency (F_0) estimation directly affects the effectiveness of jitter and shimmer features [27]. Typically, Instantaneous Robust Algorithm for Pitch Tracking (IRAPT) [32] is used in the extraction process. The IRAPT algorithm precisely estimates instantaneous pitch values and demonstrates resistance to rapid frequency modulations. Although designed to be robust, the IRAPT algorithm may perform less accurately and robustly for complex application scenarios, such as the FAD task.

The YIN algorithm has gained significant popularity [33]. It is an effective approach based on the well-known autocorrelation method, incorporating several modifications. A notable advantage of the YIN algorithm is its unrestricted frequency search range, making it suitable for high-pitched speech and music. The improved version called Probabilistic YIN (pYIN) was presented in [34]. For speech and music, another pitch estimator known as Sawtooth Waveform-Inspired Pitch Estimator (SWIPE) has been developed [35]. SWIPE estimates the F_0 by matching the spectrum of the input signal with that of a sawtooth waveform. The SWIPE algorithm and its variation, SWIPE', are effective in reducing subharmonic errors commonly observed in other pitch estimation algorithms. According to the results of natural speech reported in [32], the YIN algorithm is suitable for male speech while SWIPE' is suitable for female speech. Both hold significant potential for enhancing the effectiveness of jitter and shimmer features in the FAD task.

We aim to explicitly investigate the differences in prosody features, encoded in the jitter and shimmer, between fake and genuine speech and incorporate these features in a DNN-based FAD system for improving performance. Toward this end, we previously proposed using statistical analysis methods to identify the most promising features. In accordance with the statistical results, both the static and dynamic continuous shimmer features are then selected for integration into a light convolutional neural network bidirectional long short-term memory (LCNN-BLSTM)-based FAD system. However, in our previous work [36], we utilized a less accurate F0 estimation algorithm that limited the effectiveness of the shimmer features. Additionally, the optimal combination weights between the Mel-spectrogram and shimmer features were not thoroughly discussed, potentially leading to performance degradation due to inconsistencies in the dynamic range of different features.

To overcome these remaining problems, in this paper, we propose using the F_0 estimation algorithms, YIN and SWIPE, instead of the commonly used IRAPT. Various combination weights are tested for optimally integrating shimmer and spectrum-based features. We evaluate the effectiveness of our proposed method for FAD in the datasets of Audio Deep Synthesis Detection (ADD) 2022 and 2023 Challenges.

II. JITTER AND SHIMMER

Jitter and shimmer features represent variations in F_0 and amplitude of adjacent glottis periods, respectively. They reflect the characteristics of amplitude and frequency perturbation (AFP). To illustrate the disparity in AFP between genuine and fake speech, particularly under degraded speech quality, two segments of genuine and fake speech were chosen with identical linguistic content (/i/). These segments are depicted in Figure 1. To visualize the difference in AFP, an amplitude normalization operation was conducted, scaling the amplitudes to the range of [0,1]. From Figure 1, it is evident that the stability of the fake-speech segment notably decreased in terms of both amplitude and frequency/period.

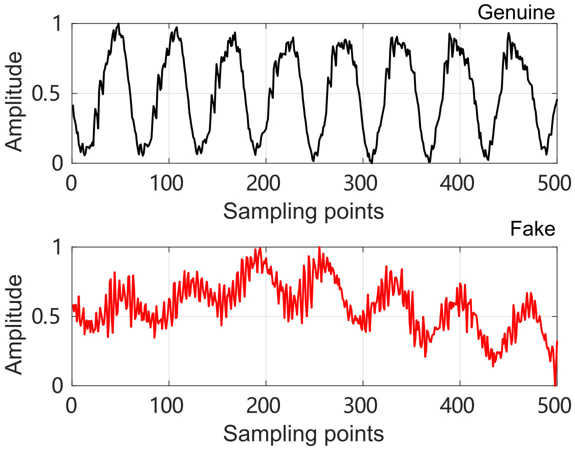


FIGURE 1. Comparison of differences among genuine and fake speech waveforms. These segments retain the same linguistic content (/i/). The sampling frequency used for the comparison is 16 kHz.

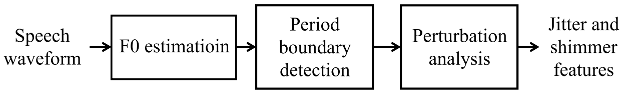


FIGURE 2. Extraction process of jitter and shimmer features.

As shown in Fig. 2, the extraction process of jitter and shimmer involves three essential steps [28]. First, the F_0 is estimated using general F_0 estimation methods. The estimated F_0 contour is used as a "reference" signal for further period detection. Therefore, the accuracy of F_o estimation directly affects the effectiveness of jitter and shimmer features. Second, the boundary of each fundamental period is detected using waveform matching with a phase constraint algorithm. Lastly, jitter and shimmer are calculated by considering several adjacent periods. In this section, we roughly introduce three state-of-the-art methods for estimating F_0: IRAPT [32], YIN [34], and SWIPE [35]. Subsequently, we provide the calculation method for the jitter and shimmer features.

A. F_0 ESTIMATION ALGORITHMS

Choosing an appropriate F_0 estimation algorithm entails considering several trade-offs. These include the upper and lower bounds of the F_0 search range, time and frequency resolution, robustness, computational complexity, and delay.

1) IRAPT

The main target of the IRAPT algorithm [32] is to estimate the instantaneous pitch values accurately, particularly in scenarios where there are rapid frequency modulations or noisy conditions. The IRAPT algorithm utilizes a robust framework that is less sensitive to rapid frequency changes and noise. Although designed to be robust, the IRAPT algorithm may still be influenced by certain artifacts or specific types of noise, which can affect the accuracy of pitch estimation. In addition, the algorithm may perform less accurately for extreme pitch ranges, where the instantaneous pitch values exhibit significant variations.

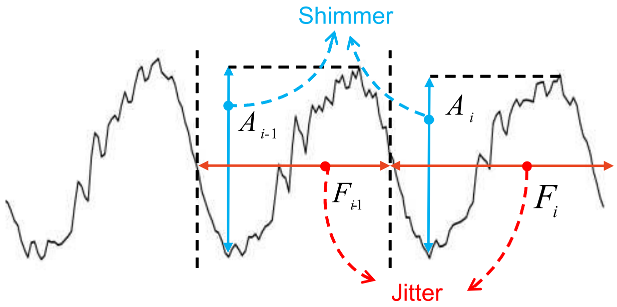


FIGURE 3. Schematic diagram of calculation of jitter and shimmer. A_i refer to the amplitude of the ith period, and F_i refer to the frequency of the ith period.

2) YIN

The YIN algorithm is based on the concept of the auto-correlation function and is particularly effective in handling non-periodic and noisy signals. The YIN algorithm provides an effective method for F_0 estimation, particularly in speech and music signals, with notable advantages in terms of computational efficiency and noise robustness. However, its applicability may be limited in certain scenarios with complex harmonic content or overlapping sounds.

3) SWIPE

The main theory behind the SWIPE algorithm is inspired by the sawtooth waveform, which is known for its periodic nature. The algorithm utilizes a comb-filtering approach to identify the F_0 by searching for the best match between the input signal and a series of synthetic sawtooth waveforms with varying periods. The key idea is to find the period that produces the highest correlation or similarity measure between the synthetic waveform and the signal being analyzed. The SWIPE algorithm offers a robust and efficient approach to F_0 estimation, particularly in scenarios involving speech and music signals.

B. CALCULATION OF JITTER FEATURES

Jitter, as depicted in Fig. 3, measures the variability of the fundamental period between consecutive periods, representing short-term variations rather than voluntary changes in F_0. In this paper, the jitter is utilized to provide some information related to the stability of speech-synthesis systems. Building upon the findings in [28], the following jitter features are considered:

(i) average and continuous differences of jitter between consecutive periods ($AJ1/CJ1$);

(ii) relative average and continuous perturbation of jitter, which evaluates the smoothness of period duration over 3 adjacent periods ($AJ2/CJ2$) [37];

(iii) average and continuous period-perturbation quotient of jitter, which quantifies the pitch period variability over 5 consecutive periods ($AJ3/CJ3$);

(iv) average and continuous frequency perturbation quotient of jitter ($AJ4/CJ4$), which aims to eliminate the influence of frequency "drift" and provide a more

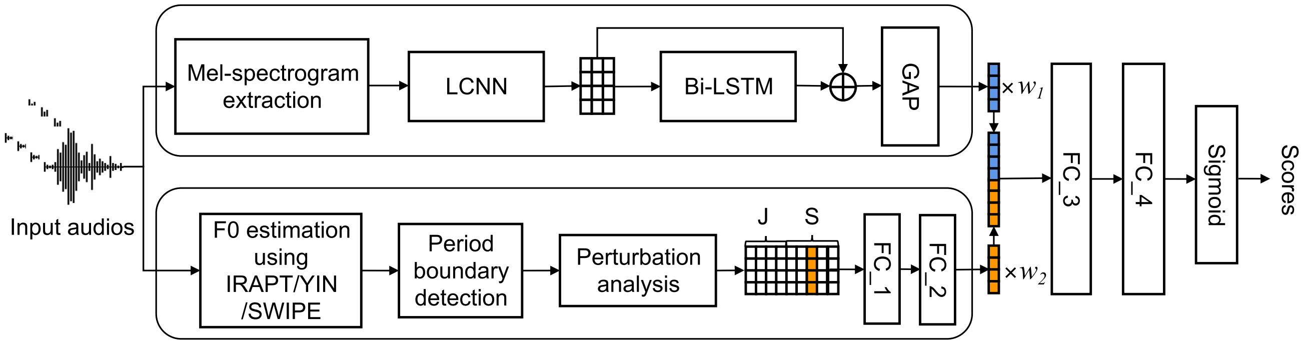


FIGURE 4. Proposed system using a combination of Mel-spectrogram and AFP features for FAD. The jitter features, consist of *CJ*1, *CJ*2, *CJ*3, and *CJ*4, are denoted by J. The shimmer features, encompassing *CS*1, *CS*2, *CS*3, *CS*4, and *CS*5, are denoted by S.

accurate index of underlying jitter in 55 consecutive periods.

Let $F(i)$ represent the frequency of the ith fundamental period in an utterance. The parameters L_p, with $p = 2, 3, 4$, represent the number of consecutive periods used in calculating $AJ2/CJ2$, $AJ3/CJ3$, and $AJ4/CJ4$, respectively. Specifically, L_2 is set to 3, L_3 to 5, and L_4 to 55. N is the total number of fundamental periods. With these definitions, we can calculate $AJ1/CJ1$, $AJ2/CJ2$, $AJ3/CJ3$, and $AJ4/CJ4$ as follows:

$$AJ1 = \frac{\frac{1}{N-1}\sum_{i=2}^{N}|F(i)-F(i-1)|}{\frac{1}{N}\sum_{i=1}^{N}F(i)} \times 100, \tag{1}$$

$$CJ1 = \frac{|F(i)-F(i-1)|}{\frac{1}{N}\sum_{i=1}^{N}F(i)} \times 100, \tag{2}$$

$$AJ_P = \frac{\frac{1}{N-L_p+1}\sum_{i=1+\frac{L_p-1}{2}}^{N-\frac{L_p-1}{2}}|F(i)-\widetilde{F(i)}|}{\frac{1}{N}\sum_{i=1}^{N}F(i)} \times 100, \tag{3}$$

$$CJ_P = \frac{|F(i)-\widetilde{F(i)}|}{\frac{1}{N}\sum_{i=1}^{N}F(i)} \times 100, \tag{4}$$

where

$$\widetilde{F(i)} = \frac{1}{L_p}\sum_{k=i-\frac{L_p-1}{2}}^{i+\frac{L_p-1}{2}} F(k). \tag{5}$$

TABLE 1. Architecture of LCNN-BLSTM-based deep classifier for FAD.

Type	Kernel Shape	Output Shape	Param
Feature_CS3	-	[64, 404]	-
Feature_Mel-spectrogram	-	[64, 80, 404]	-
Conv2d_0	[1, 64, 5, 5]	[64, 64, 404, 80]	1.66k
MaxFeatureMap2D_1	-	[64, 32, 404, 80]	-
MaxPool2d_2	-	[64, 32, 202, 40]	-
Conv2d_3	[32, 64, 1, 1]	[64, 64, 202, 40]	2.11k
MaxFeatureMap2D_4	-	[64, 32, 202, 40]	-
BatchNorm2d_5	-	[64, 32, 202, 40]	-
Conv2d_6	[32, 96, 3, 3]	[64, 96, 202, 40]	27.74k
MaxFeatureMap2D_7	-	[64, 48, 202, 40]	-
MaxPool2d_8	-	[64, 48, 101, 20]	-
BatchNorm2d_9	-	[64, 48, 101, 20]	-
Conv2d_10	[48, 96, 1, 1]	[64, 96, 101, 20]	4.704k
MaxFeatureMap2D_11	-	[64, 48, 101, 20]	-
BatchNorm2d_12	-	[64, 48, 101, 20]	-
Conv2d_13	[48, 128, 3, 3]	[64, 128, 101, 20]	55.42k
MaxFeatureMap2D_14	-	[64, 64, 101, 20]	-
MaxPool2d_15	-	[64, 64, 50, 10]	-
Conv2d_16	[64, 128, 1, 1]	[64, 128, 50, 10]	8.32k
MaxFeatureMap2D_17	-	[64, 64, 50, 10]	-
BatchNorm2d_18	-	[64, 64, 50, 10]	-
Conv2d_19	[64, 64, 3, 3]	[64, 64, 50, 10]	36.93k
MaxFeatureMap2D_20	-	[64, 32, 50, 10]	-
BatchNorm2d_21	-	[64, 32, 50, 10]	-
Conv2d_22	[32, 64, 1, 1]	[64, 64, 50, 10]	2.11k
MaxFeatureMap2D_23	-	[64, 32, 50, 10]	-
BatchNorm2d_24	-	[64, 32, 50, 10]	-
Conv2d_25	[32, 64, 3, 3]	[64, 64, 50, 10]	18.50k
MaxFeatureMap2D_26	-	[64, 32, 50, 10]	-
MaxPool2d_27	-	[64, 32, 25, 5]	-
Dropout_28	-	[64, 32, 25, 5]	-
LSTM_l_blstm	-	[25, 64, 160]	154.88k
LSTM_l_blstm	-	[25, 64, 160]	154.88k
GAP	-	[64,160]	-
FC_1	-	[64,256]	103.68k
FC_2	-	[64,160]	41.12k
Feature_Concat	-	[64, 564]	-
FC_3	-	[64, 128]	41.09k
FC_4	-	[64, 2]	258
Total	-	-	653.41k

C. CALCULATION OF SHIMMER FEATURES

Shimmer, a measure of variation in expiratory flow during articulation, has been successfully utilized in previous studies [28]. The ADD2022 and ADD2023 databases exhibit frequent amplitude variations in fake audio. Therefore, in this paper, we explore the potential usefulness of shimmer as a feature in FAD. Five shimmer features are considered for analysis. The first feature, $AS1/CS1$, represents the average and continuous basic shimmer measure. It is defined as the average absolute difference between the amplitudes of consecutive periods divided by the average amplitude [37]. To mitigate the influence of long-term changes in vocal intensity on $AS1/CS1$ and obtain a more effective representation of shimmer, we calculate four additional amplitude-perturbation quotients of shimmer. These are denoted as $AS2/CS2$, $AS3/CS3$, $AS4/CS4$, and $AS5/CS5$. The computation of these shimmer features follows a similar approach as that used for jitter. The calculations for the shimmer features are presented as follows:

$$AS1 = \frac{\frac{1}{N-1}\sum_{i=2}^{N}|A(i)-A(i-1)|}{\frac{1}{N}\sum_{i=1}^{N}A(i)} \times 100, \tag{6}$$

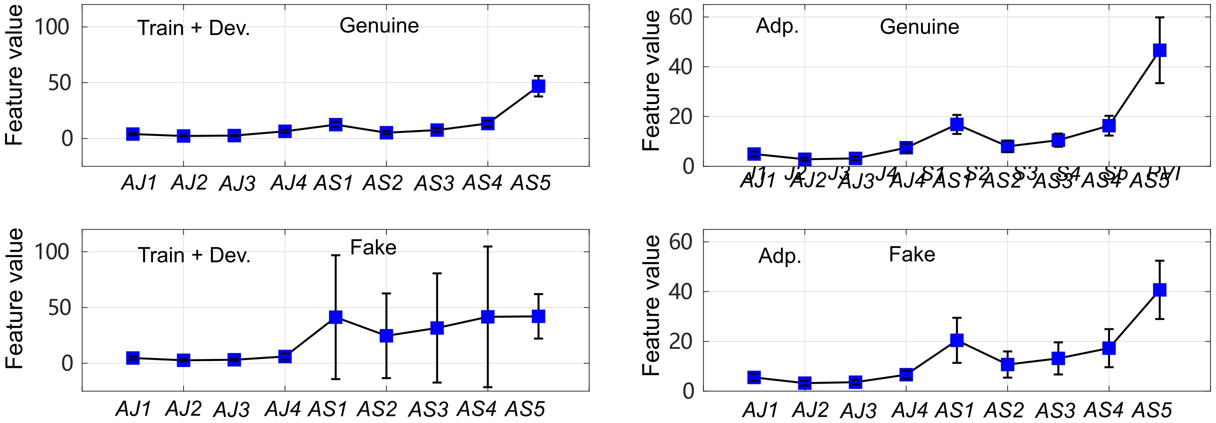


FIGURE 5. Statistical results using means and variances of averaged jitter and shimmer features in both Train + Dev. and Adp. datasets.

$$CS1 = \frac{|A(i) - A(i-1)|}{\frac{1}{N}\sum_{i=1}^{N} A(i)} \times 100, \tag{7}$$

$$AS_P = \frac{\frac{1}{N-L_P+1}\sum_{i=1+\frac{L_P-1}{2}}^{N-\frac{L_P-1}{2}} |A(i) - \widetilde{A(i)}|}{\frac{1}{N}\sum_{i=1}^{N} A(i)} \times 100, \tag{8}$$

$$CS_P = \frac{|A(i) - \widetilde{A(i)}|}{\frac{1}{N}\sum_{i=1}^{N} A(i)} \times 100, \tag{9}$$

where

$$\widetilde{A(i)} = \frac{1}{L_P}\sum_{k=i-\frac{L_P-1}{2}}^{i+\frac{L_P-1}{2}} A(k). \tag{10}$$

A(i) is the amplitude of the ith period, L_p, with $p = 2, 3, 4, 5$, represent the number of consecutive periods used in calculating $AS2/CS2$, $AS3/CS3$, $AS4/CS4$ and $AS5/CS5$, respectively. Specifically, L_2 is set to 3, L_3 to 5, L_4 to 11, and L_5 to 55.

III. FAD SYSTEM WITH JITTER AND SHIMMER FEATURES

Designing a FAD system that can combine the jitter and shimmer features with a conventional acoustic feature reasonably is also a challenging point. Fig. 4 illustrates the proposed FAD system, which combines jitter and shimmer features with a Mel-spectrogram. Previous studies [39], [40] have demonstrated that a shallow network is sufficient for downstream tasks, including anti-spoofing tasks. Therefore, this paper chooses a light convolutional neural network (LCNN) [40] based architecture as the baseline system. This LCNN is accompanied by two bi-directional recurrent layers utilizing LSTM units (BLSTM), a global-average pooling layer, and two fully connected output layers [41]. The dimensions of the BLSTM layers match the output dimensions of the LCNN. This specific architecture is commonly referred to as an LLGF network in the literature [7], [8].

We adopt a late-fusion approach to add jitter and shimmer features to the baseline system. Specifically, the jitter and shimmer features are first extracted using the method introduced in Section II. Next, these extracted features are utilized as input for two fully connected layers (FC_1 and FC_2). The resulting output from FC_2 is combined with the output of global average pooling (GAP) [42], employing distinct weights (w_1 and w_2). Different weights used here aim to regularize dynamic ranges of different features. This combined output is then passed into two additional fully connected layers (FC_3 and FC_4). Finally, to compute the score of each audio, a sigmoid function is employed. The detailed architecture of the LCNN-BLSTM model, including the kernel shape, the output shape of each layer, and the number of trainable parameters are listed in Table 1.

TABLE 2. Statistics information for the training, development, adaptation, and test datasets of the ADD2022 and ADD2023 challenges. The duration values are presented in a format indicating the minimum, mean, and maximum durations.

Dataset		Genuine	Fake	Total	Duration (sec.)
ADD 2022	Training	3,012	24,072	27,084	0.86/3.15/60.01
	Development	2,307	21,295	223,602	0.86/3.16/60.01
	Adaptation	300	700	1,000	1.13/3.63/60.01
	Test	-	-	109,199	0.35/5.51/217.46
ADD 2023	Training	3,012	24,072	27,084	0.86/3.15/60.01
	Development	2,307	26,017	28324	0.86/3.16/60.01
	Test	-	-	11,8477	0.35/5.51/217.46

An objective function named binary cross entropy (BCE) is used for optimizing the model parameters. BCE is defined as:

$$\mathcal{L}_{BCE} = -\sum_{i=1}^{N}[y_i \log P_\theta(\boldsymbol{x}_i) + (1-y_i)\log(1-P_\theta(\boldsymbol{x}_i))] \tag{11}$$

where N refers to the number of samples, θ denotes model parameters, y_i and $P_\theta(\boldsymbol{x}_i)$ are respectively the ground truth of the i-th training sample and its corresponding output probability from the model.

IV. EXPERIMENTS

A. DATA AND METRICS

The datasets from the ADD2022 [43] and ADD2023 [44] challenges were selected to assess the effectiveness of the proposed method. These challenges aim to shape the

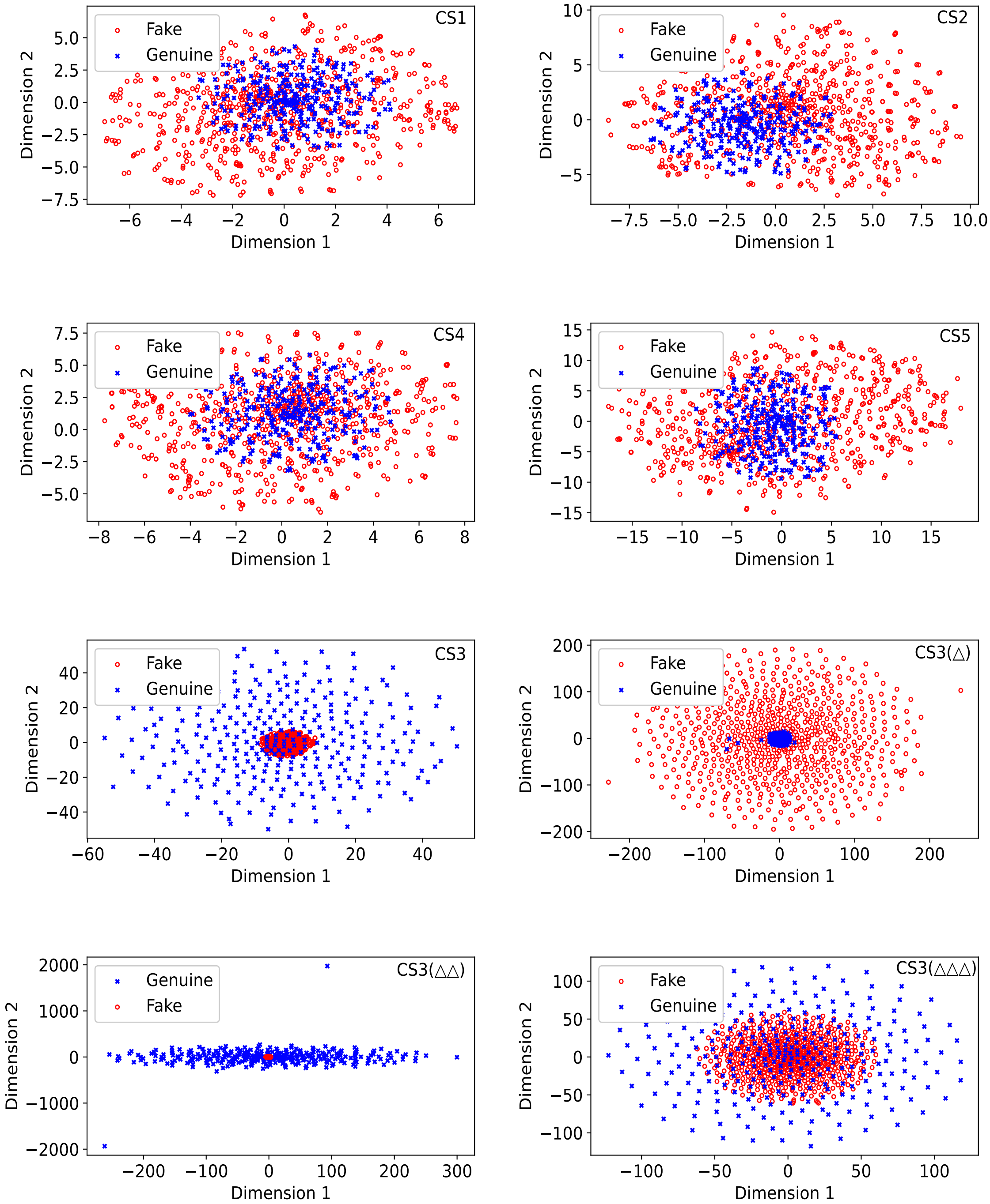


FIGURE 6. Comparison of discrimination of CS1, CS2, CS3, CS4, CS5, CS3 (△), CS3 (△△), and CS3 (△△△) for the ADD2022 adaptation set. The dimensions of these features were decreased to two and plotted by using the t-SNE toolkit [45].

future direction of detecting deep synthetic and manipulated audio in multimedia. In ADD2022, all tracks share the same training and development datasets, while an individual adaptation dataset is provided for fine-tuning and evaluation in each track. The ADD2023 comprises only the training and development datasets. For evaluation purposes, test datasets for ADD2022 and ADD2023 are available online. These datasets contain unseen audio samples obtained from various speech-synthesis systems. Notably, these samples present more real-life and challenging multimedia scenarios than those in the ASVspoof2021 challenge [46]. This paper uses the data from the track of low-quality FAD (LF) in ADD2022 and the track of audio fake game detection (FG-D) track in ADD2023. The difference between these two datasets is from the setting of the competition system, the FG-D track in the ADD2023 challenge includes two rounds of testing. The second round test is more difficult, so this paper considers the results from the second round only. Table 2 provides statistical information regarding these datasets.

TABLE 3. FAD results (EER) in the adaptation (Adp.) and test sets of ADD2022 Challenge. Data augmentation and VAD are applied in the extraction of the Mel-spectrogram only.

Front-end Features	Data Augmentation (Mel-spectrogram)	VAD (Mel-spectrogram)	Results (EER %) Adp. set	Results (EER %) Test set
Mel-spectrogram	✗	✗	17.40	38.63
CS3	✗	✗	31.00	45.46
CS3 (△)	✗	✗	32.45	43.99
CS3 (△△)	✗	✗	32.00	42.49
CS3 (△△△)	✗	✗	37.00	45.83
Mel-spectrogram	✓	✗	3.31	33.47
Mel-spectrogram + CS3	✓	✗	4.31	32.48
Mel-spectrogram + CS3 (△)	✓	✗	3.90	31.95
Mel-spectrogram + CS3 (△△)	✓	✗	3.90	**31.50**
Mel-spectrogram + CS3 (△△△)	✓	✗	3.62	32.60
Mel-spectrogram + CS3 + CS3 (△△)	✓	✗	4.81	32.28
Mel-spectrogram	✓	✓	4.55	30.41
Mel-spectrogram + CS3 (△△)	✓	✓	3.31	**29.90**

The performance of the proposed FAD system was assessed using the equal error rate (EER), following the same evaluation method used in the ADD2022 and ADD2023 challenges.

B. EXPERIMENTAL SETUP

The Mel-spectrogram was extracted by using the *MelSpectrogram* module in the *torchaudio.transforms* library [47]. The parameters used in the STFT were configured as follows: a fast Fourier transform size of 1024, a window length of 512, and a hop length of 256. In cases where the audio duration is shorter than 4 seconds, zero padding is applied. The resulting Mel-spectrogram has dimensions of [64 × 80 × 404], denoting the batch size, number of Mel filterbanks, and number of frames. The YIN and SWIPE algorithms, implemented through the *libf*0 toolbox [48], are used in this paper. The dimension of each continuous shimmer feature is [64 × 1 × 404].

Data augmentation techniques (including the introduction of reverberation, babble, and music noise) were used during the extraction of the Mel-spectrogram feature to enhance the diversity of the training dataset, hence enhancing the robustness of the back-end classifier. Additionally, voice activity detection (VAD) [49] was utilized in the pre-processing stage to minimize disturbances caused by silence clips. The training process consisted of 30 epochs, and the model that yielded the best results was compiled using the Adam optimizer, with a learning rate of 0.0001.

V. RESULTS AND DISCUSSION

This section is divided into two parts. The first part presents the results and discussion derived from ADD2022, demonstrating the efficacy of the jitter and shimmer features. The second part encompasses the results and discussion obtained from ADD2023, focusing on the utilization of various F_0 estimation methods.

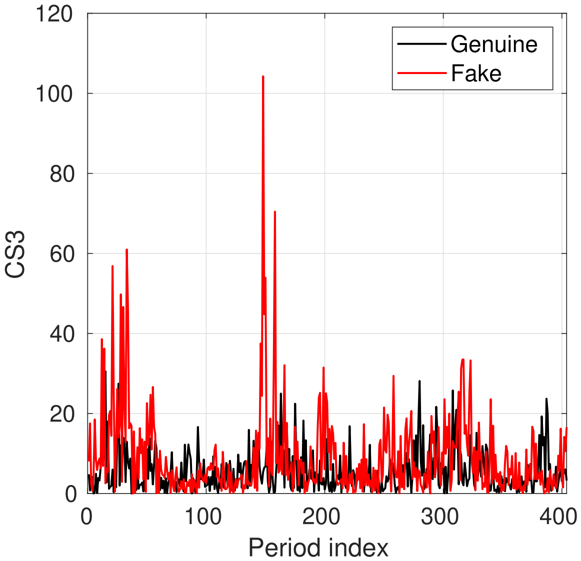


FIGURE 7. Comparison of CS3 features extracted from genuine and fake speech.

A. RESULTS AND DISCUSSION IN ADD2022

This paper examines the discrimination of various jitter and shimmer features by using statistical methods. The aim is to identify effective features for FAD. The most promising features are selected and combined with the Mel-spectrogram feature as the front-end input for an LCNN-BLSTM classifier.

The mean and variance of four averaged jitter features ($AJ1$, $AJ2$, $AJ3$, and $AJ4$) and five averaged shimmer features ($AS1$, $AS2$, $AS3$, $AS4$, and $AS5$) were analyzed in both the Train + Dev. (left column) and Adp. datasets (right column) as depicted in Fig. 5. The top graphs represent the results of genuine speech, while the bottom graphs illustrate the results of fake speech. Notably, the mean and variance of $AS1$, $AS2$, $AS3$, $AS4$, and $AS5$ exhibited an increase in the case of fake audio compared with genuine audio. This difference was more pronounced in the Train + Dev. dataset.

TABLE 4. **FAD results (EER) in the development (Dev.) and test set of ADD2023 Challenge. Different F_0 estimation methods, including IRAPT, YIN, and SWIPE, were utilized.**

Front-end features	Data augmentation	Loss			Dev. set (%)			Test set (%)		
		IRAPT	YIN	SWIPE	IRAPT	YIN	SWIPE	IRAPT	YIN	SWIPE
Mel-spectrogram	✗	0.34			1.09			61.21		
Mel-spectrogram	✓	0.39			2.99			41.29		
Mel-spectrogram+CS3	✓	0.36	0.37	0.38	3.03	3.08	1.81	38.14	**36.63**	37.32
Mel-spectrogram+CS3(△)	✓	0.39	0.41	0.37	4.25	3.38	2.64	37.31	37.05	**36.70**
Mel-spectrogram+CS3(△△)	✓	0.37	0.39	0.38	2.95	3.46	2.21	39.98	**36.18**	41.24

TABLE 5. **FAD results (EER) in the test set of ADD2023 Challenge using different combination weights between the Mel-spectrogram and CS3 △△ feature.**

Front-end features	Data augmentation	Weight	Test set (%)
Mel-spectrogram+CS3(△△)	✓	4:1	38.79
Mel-spectrogram+CS3(△△)	✓	3:2	**35.77**
Mel-spectrogram+CS3(△△)	✓	1:1	36.18
Mel-spectrogram+CS3(△△)	✓	2:3	40.70
Mel-spectrogram+CS3(△△)	✓	1:4	40.14

It is also clear that amplitude perturbation in the fake audio is much more unstable. This amplitude-perturbation instability of fake audio could provide discriminative information for accomplishing FAD.

The continuous-shimmer features ($CS1$, $CS2$, $CS3$, $CS4$, and $CS5$) were calculated to capture continuous variations in the speech waveform, providing more discriminative information than the averaged shimmer features. Principal Component Analysis (PCA) was used to visualize the discrimination capability of the continuous-shimmer features, reducing the dimensions to two and depicted in Fig. 6. Among these features, $CS3$ exhibited the fewest overlapping samples, indicating that it facilitated easier separation between genuine and fake speech. To enhance feature discrimination and incorporate dynamic variation characteristics, the first, second, and third derivatives of $CS3$ were considered ($CS3$ (△), $CS3$ (△△), and $CS3$ (△△△)), as depicted at the bottom of Fig. 6. It is evident from the figure that discrimination performance improved with the use of $CS3$ (△) and further with $CS3$ (△△), with the best performance achieved by $CS3$ (△△). However, the discrimination performance of $CS3$ (△△△) was lower than that of $CS3$ (△△). In general, $CS3$ and its dynamic features exhibit greater potential for successful FAD.

Fig. 7 illustrates the $CS3$ feature values for both genuine (blue) and fake (red) audio, providing an intuitive distinction between the two. The plot makes it evident that the CS3 feature exhibits more pronounced perturbation in fake audio than genuine audio.

The discrimination performance of $CS3$ and its dynamic features ($CS3$ (△), $CS3$ (△△), and $CS3$ (△△△)), measured by the EER, is presented in Table 3 and categorized into three parts on the basis of the utilization of data augmentation and VAD methods. We focus on the results obtained from the test dataset only, as they exhibit the same trend as the Adp. dataset. The baseline system, which utilizes an LCNN-BLSTM model with a Mel-spectrogram as input, achieved an EER of 38.63%. However, the static and dynamic CS3 features yielded higher EERs than the Mel-spectrogram. It is important to note that the dimension of the Mel-spectrogram is 80 times larger than that of the shimmer features. A specific-designed classifier could further improve the performance of shimmer features.

By incorporating additional reverberation, noise, and music during Mel-spectrogram extraction, the EER was decreased to 33.47%. Combining the Mel-spectrogram with $CS3$, $CS3$ (△), $CS3$ (△△), and $CS3$ (△△△) further decreased the EER, which is consistent with the results of statistical analysis. The best result (31.50%) was achieved when combining the Mel-spectrogram with $CS3$ (△△), resulting in a 5.89% improvement in EER compared with using only the Mel-spectrogram (33.47%). However, combining the Mel-spectrogram with $CS3$ and $CS3$ (△△), which have the same distribution state (as shown in Fig. 6), led to a slight increase in EER (from 31.50% to 32.28%). Applying VAD to filter out interference information from silent clips decreased EER to 29.90%.

B. RESULTS AND DISCUSSION IN ADD2023

Table 4 presents the experimental results conducted using three distinct methods for F_0 estimation. To ensure fairness, the losses of the selected epoch remain nearly unchanged (around 0.37). Implementing data augmentation significantly reduces the EER from 61.21% to 41.29%. Moreover, utilizing $CS3$, $CS3$ (△), and $CS3$ (△△) in conjunction with the Mel-spectrogram feature contributes to further reducing EER. These results validate the efficacy of utilizing pathological prosody information, specifically the shimmer features, for FAD.

Comparing the F_0 estimation algorithms, both the YIN and SWIPE methods improve the effectiveness of the shimmer features and exhibit lower EER values than the IRAPT algorithm. The reason for this may be that both YIN and SWIPE encompass a broader frequency search range and higher robustness for natural speech. The best result is achieved when extracting the $CS3$ (△△) feature with the YIN algorithm, resulting in an EER of 36.18%.

The exploration results of different combination weights between the Mel-spectrogram and CS3 (△△) features are presented in Table 5. The optimal result is achieved when the weight is set at a ratio of 3:2. This results in a significant improvement of 13.3% compared with using the Mel-spectrogram only, which yields a performance of 41.29%. This finding indicates that setting different combination weights can balance the effects of inconsistencies in the dynamic range of different features.

VI. CONCLUSION

This paper aimed to investigate the prosody information differences in the voice represented by using the jitter and shimmer features for the fake audio detection (FAD) task. In accordance with the statistical analysis results, the most promising features were selected and incorporated with a DNN-based FAD system. To further enhance the performance of the proposed FAD system, two additional F_0 estimation methods, namely YIN and IRAPT, were utilized in place of the IRAPT algorithm when extracting features. Different weights were tested to find out the optimal combination between the Mel-spectrogram and shimmer features.

Statistical analysis results indicate prosody differences captured by the shimmer features, especially the CS3, can provide important information to distinguish between fake and genuine speech. This finding can be further verified by combining the static and dynamic CS3 features with the Mel-spectrogram and integrating them into the LCNN-BLSTM-based FAD system. The results obtained from the ADD2023 dataset indicate that utilizing YIN and SWIPE algorithms can further improve the performance of the FAD system due to the accuracy of F_0 detection and broader frequency search range. During the online test of ADD2022, EER decreased from 33.47 % to 31.50 % in the absence of VAD, namely an improvement of 5.89 %. During the online test of ADD2023, a combined weight of 3:2 resulted in a significant improvement. The EER decreased from 41.29 % to 35.77 %, achieving an improvement of 13.37 %.

ACKNOWLEDGMENT

The authors would like to thank the assistance of Yao Wang and Prof. Minh Le Nguyen in the implementation of a portion of the experiments conducted during the early draft of this work.

REFERENCES

[1] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," 2019, *arXiv:1904.05441*.

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

[3] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 132–157, 2021.

[4] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4693–4702.

[5] 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. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 3171–3180.

[6] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," 2020, *arXiv:2006.04558*.

[7] X. Wang and J. Yamagishi, "Investigating self-supervised front ends for speech spoofing countermeasures," 2021, *arXiv:2111.07725*.

[8] X. Wang and J. Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," 2021, *arXiv:2103.11326*.

[9] Z. Wu, R. K. Das, J. Yang, and H. Li, "Light convolutional neural network with feature genuinization for detection of synthetic speech attacks," 2020, *arXiv:2009.09637*.

[10] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.

[11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[12] 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.

[13] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection," in *Proc. INTERSPEECH*, Sep. 2019, pp. 1068–1072.

[14] 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.

[15] 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*.

[16] 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.

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

[18] 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.

[19] C.-I. Lai, N. Chen, J. Villalba, and N. Dehak, "ASSERT: Anti-Spoofing with Squeeze-Excitation and Residual neTworks," 2019, *arXiv:1904.01120*.

[20] M. Todisco, H. Delgado, K. A. Lee, M. Sahidullah, N. Evans, T. Kinnunen, and J. Yamagishi, "Integrated presentation attack detection and automatic speaker verification: Common features and Gaussian back-end fusion," in *Proc. 19th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2018, pp. 77–81.

[21] 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. e2, 2020.

[22] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[23] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Comput. Speech Lang.*, vol. 45, pp. 516–535, Sep. 2017.

[24] Z. Lei, Y. Yang, C. Liu, and J. Ye, "Siamese convolutional neural network using Gaussian probability feature for spoofing speech detection," in *Proc. INTERSPEECH*, Oct. 2020, pp. 1116–1120.

[25] G. Pamisetty and K. S. R. Murty, "Prosody-TTS: An end-to-end speech synthesis system with prosody control," *Circuits, Syst., Signal Process.*, vol. 42, no. 1, pp. 361–384, Jan. 2023.

[26] S. Nooteboom, "The prosody of speech: Melody and rhythm," in *The Handbook of Phonetic Sciences*, vol. 5, 1997, pp. 640–673.

[27] I. R. Titze and H. Liang, "Comparison of F_o extraction methods for high-precision voice perturbation measurements," *J. Speech, Lang., Hearing Res.*, vol. 36, no. 6, pp. 1120–1133, Dec. 1993.

[28] 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.

[29] 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.

[30] 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.

[31] E. S. Kass, R. E. Hillman, and S. M. Zeitels, "Vocal fold submucosal infusion technique in phonomicrosurgery," *Ann. Otol., Rhinol. Laryngol.*, vol. 105, no. 5, pp. 341–347, May 1996.

[32] E. Azarov, M. Vashkevich, and A. Petrovsky, "Instantaneous pitch estimation based on RAPT framework," in *Proc. 20th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2012, pp. 2787–2791.

[33] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[34] M. Mauch and S. Dixon, "PYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 659–663.

[35] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 124, no. 3, pp. 1638–1652, Sep. 2008.

[36] 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.

[37] R. J. Baken and R. F. Orlikoff, *Clinical Measurement of Speech and Voice*, 2nd ed. San Diego, CA, USA: Singular Thomson Learning, 2000.

[38] J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, J. F. Vargas-Bonilla, K. Daqrouq, S. Skodda, J. Rusz, and E. Nöth, "Automatic detection of Parkinson's disease in running speech spoken in three different languages," *J. Acoust. Soc. Amer.*, vol. 139, no. 1, pp. 481–500, Jan. 2016.

[39] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12449–12460.

[40] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," in *Proc. INTERSPEECH*, Aug. 2017, pp. 82–86.

[41] K. Li, S. Li, X. Lu, M. Akagi, M. Liu, L. Zhang, L. Wang, J. Dang, and M. Unoki, "Data augmentation using mcadams-coefficient-based speaker anonymization for fake audio detection," in *Proc. INTERSPEECH*, 2022, pp. 664–668.

[42] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*.

[43] 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.

[44] J. Yi, J. Tao, R. Fu, X. Yan, C. Wang, T. Wang, C. Yuan Zhang, X. Zhang, Y. Zhao, Y. Ren, L. Xu, J. Zhou, H. Gu, Z. Wen, S. Liang, Z. Lian, S. Nie, and H. Li, "ADD 2023: The second audio deepfake detection challenge," 2023, *arXiv:2305.13774*.

[45] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.

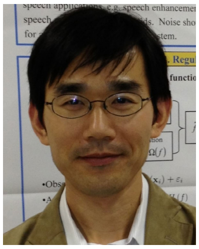
[46] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, "ASVspoof2021: Accelerating progress in spoofed and deep fake speech detection," 2021, *arXiv:2109.00537*.

[47] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8026–8037.

[48] S. Rosenzweig, S. Schwär, and M. Müller, "Libf0: A Python library for fundamental frequency estimation," in *Proc. Late Breaking Demos Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, Bengaluru, India, 2022.

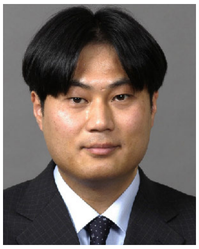
[49] J. Kim and M. Hahn, "Voice activity detection using an adaptive context attention model," *IEEE Signal Process. Lett.*, vol. 25, no. 8, pp. 1181–1185, Aug. 2018.

KAI LI received the double M.S. degree in computer science and technology from Tianjin University (TJU), Tianjin, China, in 2019, and in information science from the Japan Advanced Institute of Science and Technology (JAIST), Nomi, Japan, in 2020. He is currently pursuing the Ph.D. degree with the Acoustic Information Science (AIS) Laboratory, JAIST. His current research interests include modeling speech production, speaker anonymization, deep fake detection for audio, and anomalous sound detection for machine condition monitoring.



XUGANG LU (Member, IEEE) received the B.S. and M.S. degrees from the Harbin Institute of Technology, China, in 1994 and 1996, respectively, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, in 1999. In October 1999, he joined Nanyang Technological University, Singapore, as a Research Fellow. Since December 2001, he has been a Postdoctoral Fellow with McMaster University, Canada. From April 2003 to April 2008, he was an Assistant Professor with the Faculty of School of Information Science, Japan Advanced Institute of Science and Technology. In May 2008, he joined the Spoken Language Communication Research Laboratories, Advanced Telecommunications Research Institute (ATR), and then moved to the National Institute of Information and Communications Technology (NICT), Japan, as a Senior Researcher. His research interests include speech signal processing and recognition and machine learning.

MASATO AKAGI (Life Member, IEEE) received the B.E. degree from the Nagoya Institute of Technology, in 1979, and the M.E. and Ph.D. (Eng.) degrees from the Tokyo Institute of Technology, in 1981 and 1984, respectively. He joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Corporation (NTT), in 1984. From 1986 to 1990, he was with the ATR Auditory and Visual Perception Research Laboratories. Since 1992, he has been a Faculty Member with the School of Information Science, JAIST. He is currently a professor emeritus. His research interests include speech perception, modeling of speech perception mechanisms in humans, and the signal processing of speech. He was a recipient of the IEICE Excellent Paper Award from IEICE, in 1987, the Best Paper Award from the Research Institute of Signal Processing, in 2009, and the Sato Prize for Outstanding Papers from the Acoustical Society of Japan, in 1998, 2005, 2010, and 2011.



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