Annals of Emerging Technologies in Computing (AETiC)
Vol. 8, No. 2, 2024

Research Article

The Proposal of Countermeasures for DeepFake Voices on Social Media Considering Waveform and Text Embedding

Yuta Yanagi[1,*], Ryohei Orihara[1], Yasuyuki Tahara[1], Yuichi Sei[1], Tanel Alumäe[2] and Akihiko Ohsuga[1]

[1]The University of Electro-Communications, Japan
yanagi.yuta@ohsuga.lab.uec.ac.jp; orihara@acm.org; tahara@uec.ac.jp; seiuny@uec.ac.jp; ohsuga@uec.ac.jp
[2]Tallinn University of Technology, Estonia
tanel.alumae@taltech.ee
*Correspondence: yanagi.yuta@ohsuga.lab.uec.ac.jp

Received: 13th October 2023; Accepted: 17th February 2024; Published: 1st April 2024

Abstract: In recent times, advancements in text-to-speech technologies have yielded more natural-sounding voices. However, this has also made it easier to generate malicious fake voices and disseminate false narratives. ASVspoof stands out as a prominent benchmark in the ongoing effort to automatically detect fake voices, thereby playing a crucial role in countering illicit access to biometric systems. Consequently, there is a growing need to broaden our perspectives, particularly when it comes to detecting fake voices on social media platforms. Moreover, existing detection models commonly face challenges related to their generalization performance. This study sheds light on specific instances involving the latest speech generation models. Furthermore, we introduce a novel framework designed to address the nuances of detecting fake voices in the context of social media. This framework considers not only the voice waveform but also the speech content. Our experiments have demonstrated that the proposed framework considerably enhances classification performance, as evidenced by the reduction in equal error rate. This underscores the importance of considering the waveform and the content of the voice when tasked with identifying fake voices and disseminating false claims.

Keywords: *natural language processing; neural networks; speech synthesis; voice processing*

1. Introduction

1.1. Background

In recent years, speech synthesis technology has undergone substantial development [1–3]. On one hand, this progress has made it much easier for individuals to access more authentic and natural-sounding voices. On the other hand, there is growing concern regarding DeepFake voices, also known as fake voices or spoofing, which are used by malicious users to deceive others. We define DeepFake voices as synthesized speeches that convey false information. Previously, fake news, also known as disinformation or misinformation, has been a source of deception, leading to research efforts aimed at early detection [4]. DeepFakes can manifest as a multimodal form of fake news, which must be identified and addressed before widespread dissemination occurs.

Yuta Yanagi, Ryohei Orihara, Yasuyuki Tahara, Yuichi Sei, Tanel Alumäe and Akihiko Ohsuga, "The Proposal of Countermeasures for DeepFake Voices on Social Media Considering Waveform and Text Embedding", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 15-31, Vol. 8, No. 2, 1st April 2024, Published by International Association for Educators and Researchers (IAER), DOI: 10.33166/AETiC.2024.02.002, Available: http://aetic.theiaer.org/archive/v8/v8n2/p2.html.

Examples of fake voices utilizing voice synthesis are depicted in Figure 1(a). An occurrence of spoofing portrayed as a fictitious company's CEO[1] is illustrated in Figure 1(a). This example uses voice conversion (VC), transforming the attacker's voice to match the targets. Figure 1(b) reveals an instance of spoofing through Text-To-Speech (TTS). Notably, Bunn[2] advocates creating a novel TTS model utilizing merely three seconds of the target's voice. Identifying attackers becomes challenging as they input text rather than voices. Furthermore, voices converted by VC encapsulate pertinent speaking features (e.g., speed, pitch, etc.) to identify attackers. This is serious problem because social media is widely used in several situations[5,6]. We can verify the information by checking the spreader's details. However, given the security issues of social media platforms [7], there is also concern that DeepFakes could be sent through account hacking. In general, with TTS, it is easier to obtain the voice than with VC because it does not require the attacker's voice as input. Therefore, the number of instances of the DeepFakes generated by TTS will increase. Consequently, the focus of this study is directed toward attacks that adopt TTS. In the current year, additional instances of fake voices mimicking celebrities to propagate false accusations have surfaced on social media[3]. In conjunction with celebrities, other malicious endeavours include cloning the voices of the target's acquaintances or relatives for fraudulent purposes. Hence, considerable research has been directed toward thwarting fake voice attacks [8–11]. Moreover, extensive studies focus on detecting fraudulence through films [12] and media traceability [13]. In scenarios where users are unable to utilize films on social media or during online conversations, the methodology herein is dedicated to voice-based detection of false narratives about the given situation.

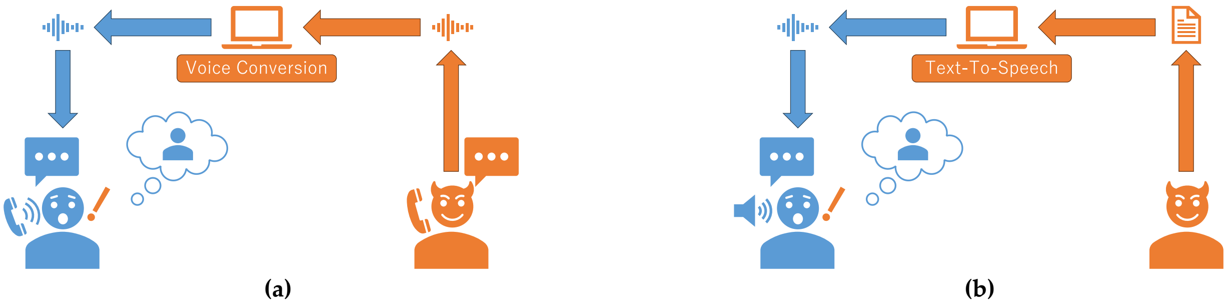


Figure 1. Examples of fraud utilized speech synthesis. (a) Voice conversion transforms the attacker's voice to the target. (b) Text-To-Speech generates voice from the input text. Identifying attackers from TTS voice is difficult due to the lack of voice origin.

The ASVspoof initiative stands out as the most widely recognized effort in the development of antispoofing methods [9,14–16]. This initiative has been ongoing every other year since its inception in 2015, attracting numerous participants who propose models aimed at distinguishing DeepFake voices (referred to as spoofing) from genuine voices. Originally, ASVspoof primarily focused on using DeepFake voices to challenge biometric systems. However, in 2021, the initiative expanded its scope to include tasks targeting users on social media [16]. The inaugural event in 2015 involved the generation of two types of speech, text-to-speech (TTS) and voice conversion, which were played back over a telephone line [17]. Subsequent iterations included the release of datasets designed to test the vulnerability of biometrics systems by playing back prerecorded real voices in 2017 [14]. In 2019, ASVspoof introduced the logical attack (LA) and the physical attack (PA) scenarios, incorporating new methods and environments [15]. In the 2021 edition, ASVspoof added a new DeepFake (DF) task to the LA, involving the classification of audio played in an online environment [16]. Notably, it was observed that the generalization performance of many methods participating in the DF task was lacking. Models trained on datasets up to 2019 struggled to accurately classify validation sets from 2020 onward [16].

[1] Catherine Stupp, 2019, Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case, Available: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

[2] Amy Bunn, Artificial imposters-cybercriminals turn to ai voice cloning for a new breed of scam, Available: https://mcafee.ly/3pNjfCE.

[3] Joseph Cox, 2023, Ai-generated voice firm clamps down after 4chan makes celebrity voices for abuse, Available: https://www.vice.com/en/article/dy7mww/ai-voice-firm-4chan-celebrity-voices-emma-watson-joe-rogan-elevenlabs.

1.2. Problem Statement

Furthermore, over the past couple of years, numerous new TTS methods have emerged, yielding voices that sound more natural than ever before. Consequently, there is a growing concern that DeepFake detection models may soon become obsolete as voice generation techniques continue to advance [16,18]. To address this evolving landscape, we recognize the importance of evaluating not only the quality of waveforms but also the content of the voice, specifically what the speaker is saying. Consequently, the present study conducts targeted experiments to assess the effectiveness of classification models trained on the ASVspoof 2021 DF dataset in identifying fake voices generated using recently developed TTS methods. Preliminary results reveal that the classification model failed to detect fake voices, aligning with previous findings [16], thus underlining the necessity for improved generalization performance.

1.3. Research Contribution

As a result, this study reports the failure of the classification model in detecting DeepFake voices in nearly all cases. Subsequently, we conducted an experiment aimed at classifying voices from tweets posted on Twitter, distinguishing between those conveying real news and fake news using tweet embeddings. Our findings indicate that considering both the waveform of the voices and the content provides a more effective method for assessing credibility compared to waveform analysis alone. In response to these results, this study delves into the prospects of detecting DeepFake voices disseminating false information on social media. Additionally, we propose a novel framework designed to identify DeepFake voices by simultaneously analysing the waveform and speech content. Moving forward, this study will outline its contents in the following sequence. Table 1 shows the differences between the previous models and proposed one. Our model employs a dual-processing approach, examining the waveform and speech content, as detailed in Table 1. In section 2, we provide an overview of related research in the field, focusing on the automatic detection of fake news and DeepFake voices. Next, in section 3, we introduce the models used for classifying DeepFake voices. We elaborate on the experimental procedures, including the generation of a DeepFake voice dataset and the setup of the classification models, in section 4. Our experimental results are detailed in section 5. Subsequently, we offer a comprehensive evaluation of these results from various perspectives in section 6. Finally, we summarize our findings in section 7. It is important to note that this study extends and builds upon the work presented in our conference paper at ICMECE 2022 [19].

Table 1. The difference between previous methods and our proposed framework.

Model Name	Input	Apply voice?	Evaluate content?
RawNet2	Voice Waveform	✓	✗
RawBoost	Voice Waveform	✗	✗
Classify text	Text embedding (emb.)	✗	✓
Proposed	**Waveform & Text emb.**	✓	✓

2. Literature Review

In this section, we introduce the research related to DeepFake voices, categorizing it into three sections to facilitate a comprehensive comparison. These sections include investigations into the automatic detection of DeepFake speech/texts and highlight the problem posed by advances in speech synthesis technologies.

2.1. Automatic Detection of DeepFake Voices

Numerous endeavours have been made in developing countermeasures against deceptive practices involving speech synthesis [8,20,21], voice conversion [22–25], and replay attacks [26], particularly before the year 2013. The need for standard datasets, protocols, and metrics was highlighted by a research team in 2013 [27,28]. Subsequently, in 2015, the ASVspoof challenge was introduced [9,17]. This initiative marked a significant step in creating a shared dataset, establishing evaluation protocols, and defining metrics. The ASVspoof challenge comprises various tasks. The first is the LA task, which centres on spoofing the telephony environment. Subsequently, ASVspoof 2017 focused on the PA task, which involves replay attacks using audio from the target [14]. These two tasks were further enhanced in ASVspoof 2019 [15], with the addition of new voice generation models and environments for replay attacks. The latest iteration in 2021 introduced the DF task, which involves the classification of voices in an online environment [16] with baseline models [29]. Notably, it was observed that the generalization performance in 2021 was insufficient [16]. This trend has emerged because models trained using methods developed up to 2019 struggled to accurately classify the DeepFake voices generated using newer techniques introduced in 2020. Additionally, an extended shared task that builds upon ASVspoof, known as the Audio DeepFake Detection, was proposed towards the end of 2021 [30]. The organizers have scheduled the latest iteration for 2023.

A common aspect shared by ASVspoof tasks is the classification of real and generated voices based on the same voice content, specifically what the speaker is saying. In other words, previous methods have lacked a perspective on assessing the credibility of the voice content. However, we believe that evaluating the speech content is crucial when attempting to detect generated voices disseminating false information on social media.

2.2. Developments of Text-To-Speech

In ASVspoof 2015, the organizers employed TTS methods for speech synthesis [17,31]. In ASVspoof 2019, they expanded their approach by incorporating numerous neural-network-based TTS models [32]. Historically, the prevailing method for speech generation from pre-processed sentences has involved a two-stage process. Initially, acoustic [2,33] and linguistic features [34] are extracted using one model, followed by the generation of output waveforms using another model [34,35]. A common characteristic of these approaches is the division of the voice generation process into two stages. However, this two-stage pipeline framework necessitates sequential training or fine-tuning [2,36]. During training, this means that users cannot simultaneously train in both stages but must first prepare for the feature extraction part. This can be a resource-intensive and time-consuming process. As a result, several research efforts have emerged to develop end-to-end models to reduce training costs and capture hidden representations for output generation [1,37,38]. The performance of these end-to-end models has been steadily improving [39]. Notably, the natural speech model achieved remarkable results, demonstrating no statistically significant difference between human recordings and the generated ones by incorporating a variational autoencoder at the waveform generation stage [40].

2.3. Problem of DeepFake Voice Detection

In summary, the rapid development of TTS technology has outpaced the capabilities of conventional detection methods. Competing with this advancement solely based on waveform analysis is nearly impossible. Consequently, we propose a novel detection framework that considers both waveform characteristics and speech content.

Table 2 summarizes the research we introduced in this article, with contributions, limitations, and concerns.

Table 2. Contributions and limitations of studies in the literature.

Ref.	Contributions	Limitations and concerns
[14,15,17]	ASVspoof is the largest project that focuses on identifying voice spoofing.	The situation is spoofing for biometrics, not social media, by 2019.
[16]	ASVspoof 2021 added a new scenario for DeepFakes (DFs) in social media	In all voice-read newspaper articles in Scotland, there is no misinformation
[29]	RawNet2 had the best score in the baselines of ASVspoof DF.	The test set's score decreased, including spoofing produced by newer models.
[39]	VITS produces a more natural voice with the conditional variational autoencoder.	There is a concern that attackers are making more natural DF spoofs with the model.

3. Framework of DeepFake Voice Detection

In this section, we present the framework for the automatic detection of DeepFake voices disseminating false claims. Our model incorporates two distinct processing components, one for analysing the waveform and the other for evaluating speech content.

Figure 2 shows the structure of our proposed model. We explain each part in detail.

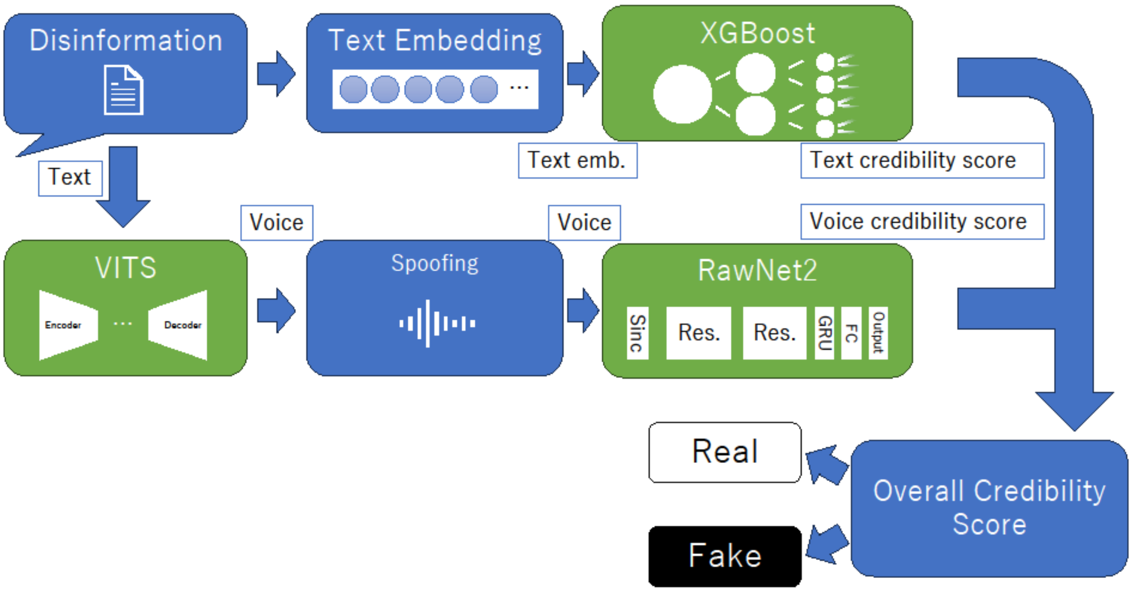


Figure 2. The structure of generating dataset and proposed model.

3.1. Speech Waveform

For classifying DeepFake voices, we employed RawNet2 [41] and RawBoost [42].

3.1.1. RawNet2

We employed the RawNet2 [41] model to consider waveform. This model serves as a baseline in the ASVspoof 2021 DF task [29]. We adopted the model because it is accessible online, and it produced the best scores among the baselines in the task.

RawNet2 is an enhanced version of the RawNet model [43]. Both models operate as end-to-end classifiers, integrating the extracting of utterance-level features and a feature enhancement phase [43]. For a detailed framework of RawNet2, please refer to Table 3. The first layer in the RawNet2 framework utilizes a sinc-convolution layer, as introduced in SincNet [44,45]. SincNet employs a convolutional neural network (CNN) architecture to filter raw waveforms using bank-pass filter shapes resembling sinc functions. The second layer consists of a residual block, which includes batch normalization (BN), LeakyReLU [46], a convolutional layer, max-pooling, and feature map scaling (FMS). FMS, proposed in [39], functions akin to an attention layer with a sigmoid activation [41]. The output layer is designed for binary classification, distinguishing between DeepFake voices and bona fide (real) voices. The model utilizes weighted categorical cross-entropy as its loss function, following the ASVspoof's baseline setup in the GitHub repository[4]. Given that this is a binary classification task for distinguishing between real and fake voices, the loss LRN is determined by the following equation:

$$L_{RN}(y,\hat{y}) = -0.1 * y\log\hat{y} - 0.9 * (1-y)\log(1-\hat{y}) \qquad (1)$$

In the equation, y represents the label values, where $y = 0$ corresponds to real voices, and $y = 1$ corresponds to fake voices. $\hat{y}$ represents the predicted values generated by RawNet2, which undergo processing through the softmax function followed by the log function. The constants 0.1 and 0.9 are derived from the label distribution observed in the ASVspoof training set [16]. These values are applied to appropriately weight the loss function.

[4] https://github.com/asvspoof-challenge/2021/tree/main/DF/Baseline-RawNet2

Table 3. The RawNet2 architecture which is applied for ASVspoof 2019 [29] and 2021 [16].

Layer	Input	Output shape

Fixed Sinc filters	Conv(129, 1, 128)
	Maxpooling(3)
	BN & LeakyReLU
		(21290, 128)

Res block	BN & LeakyReLU
	Conv(3, 1, 128)
	BN & LeakyReLU
	Conv(3, 1, 128)
	Maxpooling(3)
	FMS
		(2365, 128)

Res block	BN & LeakyReLU
	Conv(3, 1, 128)
	BN & LeakyReLU
	Conv(3, 1, 128)
	Maxpooling(3)
	FMS
		(29, 512)

GRU	GRU(1024)	(1024)
FC	1024	(1024)
Output	1024	2

3.1.2. RawBoost

Additionally, in our preliminary experiment, we employed the RawBoost model [42]. RawBoost is a data boosting and augmentation technique that operates at the raw waveform level [42]. We used this model because it scored better than the baselines, including RawNet2 in [42]. It incorporates noise addition as a data augmentation technique in three distinct forms: (1) linear and nonlinear convolutive noise, which replicates noise introduced during encoding, compression, and transmission processes, (2) impulsive signal-dependent additive noise, such as clipping and nonoptimal operation of devices (e.g., microphones and amplifiers), and (3) stationary signal-independent additive noise achieved by applying a single finite impulse response filter [42]. Furthermore, we employed the same weighted cross-entropy loss as the loss function during training, following the specifications provided in its GitHub repository[5].

3.2. Speech Content: Tweet Embedding

To consider the content of the voice, we leverage tweet embedding, which can be obtained from the MuMiN dataset [47]. We used this because it is ready to use as a text feature. This embedding is generated using the BART-large-CNN transformer [48]. We employed XGBoost [49] as the classifier for the tweet embedding. Our plans include incorporating a Deep Neural Network model. Additionally, we intend to integrate an automatic speech recognition model to provide a comprehensive approach for detecting DeepFake voices within the context of social media. The loss L_{emb} is computed as the mean of the weighted squared loss, as elaborated in Chen T *et al.* [49] and is defined as follows:

$$\begin{gathered} L_{emb} = \sum_{i} l(\hat{y}_i, y_i) + \sum_{k} \Omega(f_k) \\ \text{where } \Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \end{gathered} \tag{2}$$

In the equation, y_i denotes the label values, and Ω serves to penalize the model's complexity, which encompasses the number of leaf nodes (T) and the L2 norm of weights ($\|w\|^2$), as described in Chen and Guestrin [49]. The values of γ and λ are both set to 0 because we employed this model with default parameters, as detailed in the documentation[6]. Consequently, the loss is computed simply as the mean of the weighted squared loss. It is worth emphasizing that the proposed structure offers opportunities for enhancement through parameter tuning. This aspect is among the topics we plan to explore in the future.

[5] https://github.com/TakHemlata/RawBoost-antispoofing

[6] https://xgboost.readthedocs.io/en/stable/parameter.html

3.3. Integration of Waveform and Content

Figure 2 illustrates the structure of our proposed framework. After obtaining the credibility scores for waveforms (from RawNet2) and content (from XGBoost), we calculate the final output credibility score c_f by averaging these scores, as shown in the following equation:

$$c_f = \alpha\, c_w + (1-\alpha)\, c_{\mathrm{emb}} \tag{3}$$

Where the parameter α falls within the range [0,1] and serves as a weight that determines the balance between the contributions of the credibility score of waveforms c_w and text embedding c_{emb} to the final credibility score c_f. The c_w is normalized into [0,1] range. When we exclusively consider the waveform, α is set to 1. In our proposed framework, during the main experiment, we configured α to be 0.5 to treat credibility scores equally. This choice aligns with our primary objective, which is to present a novel approach for detecting fake voices in social media. There exists an alternative approach to integrating the mean scores of RawNet2 and text embedding through fully connected layers. However, this method demands substantial computational resources and time. Considering real-life scenarios in social media, we place a high priority on efficiency, which is why we calculate the loss solely based on the mean of the weighted squared loss.

4. Preliminary Experiment

In this study, we assessed the performance of anti-spoofing models to determine their ability to detect DeepFake voices generated by the latest voice generation (TTS) models. We conducted a preliminary experiment to evaluate the accuracy of these detection models in classifying voices. Additionally, we explored a new classification framework that considers both waveform and speech content in a main experiment. Here, we provide an overview of the preliminary experimental procedures, including data acquisition and classification. Table 4 shows environments of experiments in the article.

Table 4. Computing environments for experiments.

Term	Detail
CPUs	Intel® Xeon® CPU E5-2698 v4 @ 2.20GHz
GPUs	TeslaV100-PCIE-32GB x8
RAM	503GB
Storage	7TB
CUDA Version	11.4
Tools	Python 3.10.8, PyTorch 2.1.0

4.1. Dataset

We had to generate our DeepFake voices using the latest TTS model based on fake news content. We utilized news tweets from the MuMiN dataset [47], which comprises multilingual tweets from Twitter containing both fake and real news. The MuMiN dataset covers a wide range of news topics: politics, gossip, military, sports, and COVID-19. The dataset comes in three sizes: small, medium, and large. For the preliminary experiment, we employed the MuMiN small. Table 3 provides statistics on the dataset used in the preliminary experiment, which comprised 465 news tweets in English. Some of the tweets obtained were confirmed as fake news, while others appeared to be factual. This dataset may require a larger volume of tweets to effectively fine-tune the models. Therefore, we are actively considering the inclusion of additional fake and factual news tweets from other datasets. Additionally, we are exploring the possibility of incorporating not only the news articles themselves but also the tweets that propagate these articles. To maintain manageable audio durations, we have imposed a word limit of 480 words for longer tweets, ensuring that the resulting audio clips do not become excessively lengthy, refer to Table 5. The generated voice used in our study was produced using the VITS model, as referenced in [39]. We employed a pretrained model from the LJ Speech dataset, which is a publicly available speech dataset comprising 13,100 short audio clips featuring a single speaker reading passages from 7 nonfiction books, as documented in the providers page[7]. Each audio clip is accompanied by a corresponding transcription, with clip durations

[7] Ito K, Johnson L.: The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/.

ranging from 1 to 10 s, totalling approximately 24 hr of speech data. We obtained the synthesized voices based on official instructions[8].

Table 5. Statistics of our dataset for a preliminary experiment.

Statistics	Value
Number of news tweets	465
The upper limit of words	480
Average num. of words	342
Ave. duration of voice [s]	118.7

4.2. Classification Settings

In the ASVspoof 2021 DF dataset, the organizers opted to utilize speech production models, including TTS and voice conversion techniques, which were proposed before 2019, as noted in [16]. Notably, starting in 2020, these older models were deliberately excluded from the training datasets. The motivation behind this decision was to assess the generalization performance of the models. The report on the ASVspoof 2021 DF dataset highlights that the participants faced challenges in achieving satisfactory generalization performance with their models. However, it is worth mentioning that the report provided only a limited analysis of the classification results for each voice generation model. To further investigate the effectiveness of these models, we conducted a preliminary experiment in which we exclusively utilized waveform data. The primary objective of this experiment was to determine whether the models could accurately discern synthesized voices generated by the latest TTS models. We generated the voice dataset from MuMiN. Subsequently, we employed this dataset for evaluation with two detection models: RawNet2 and RawBoost. Both models underwent training using the ASVspoof 2021 DF dataset. [29,42] describe the detailed training processes. We obtained the pretrained model from GitHub [9,10]. To establish a threshold for these models, we leveraged the equal error rate (EER) metric, which represents the point at which the false rejection rate (FRR) equals the false acceptance rate (FAR). The EER, a pivotal evaluation criterion, is calculated using the following formula:

$$FRR = \frac{FN}{FN + TP} \tag{4}$$

$$FAR = \frac{FP}{FP + TN} \tag{5}$$

In this context, TP (True Positive) signifies instances where the models correctly classify synthesized (fake) voices as fake. FN (False Negative) represents situations where the model erroneously classifies fake voices as real or authentic. FP (False Positive) denotes occurrences where the model incorrectly categorizes authentic voices as fake. These definitions are essential for assessing the accuracy and effectiveness of the detection models in distinguishing between real and synthesized voices. This formula allows us to determine the threshold that optimally balances the FRR and FAR for our detection models.

Following that, we checked the remaining outputs of the models using our dataset. We assessed the number of outputs from our proposed dataset that surpassed the thresholds established based on the EER criterion as defined by the ASVspoof dataset. Given that our proposed dataset exclusively consists of synthesized voices, the proportion of instances that exceeded these thresholds represents the recall rate for synthesized voices. It is worth noting that this experiment has been described in the ICMECE 2022 article [19].

4.3. Results

Table 6, Table 7, and Figure 3 show the result of the preliminary experiment. Table 6 illustrates the EER values of the models and the thresholds that are used for detection. The output of the models plays a crucial role in determining the suspicion levels. When the model's output surpasses a certain threshold, it indicates that the model has identified the target voice as fake. Adjusting the threshold value has a significant impact on the model's performance. If we set the threshold to a low value, the model becomes more sensitive and

[8] https://github.com/jaywalnut310/vits/blob/main/inference.ipynb

[9] https://github.com/asvspoof-challenge/2021/tree/main/DF/Baseline-RawNet2

[10] https://github.com/TakHemlata/RawBoost-antispoofing

can detect more fake voices. However, this may lead to more false positives, where real voices are incorrectly classified as fake. Conversely, setting a high threshold makes the model less sensitive, reducing the chances of false positives. However, it can also make the model more susceptible to spoofing, where fake voices are not detected. To determine the optimal threshold values, we utilized the ASVspoof 2021 DF dataset. We selected the threshold that minimizes the EER within this dataset. It is important to note that the threshold for RawNet2 is a negative value. This is because the output of RawNet2 is always zero or negative, which is a result of the model's specification that applies the LogSoftmax function. In the case of RawNet2, its performance closely aligns with the results reported in the ASVspoof 2021 article by [16]. However, when it comes to RawBoost, its performance on the ASVspoof DF task was inferior to that of RawNet2. This trend might be attributed to the fact that RawBoost was not initially designed for this specific task. To standardize the application of the output values from our proposed dataset, we employed the threshold derived from the EER of the ASVspoof DF dataset. This approach allowed us to make consistent and meaningful comparisons between different models and datasets.

Table 6. The EER and thresholds of the classification models for the ASVspoof 2021 DF dataset

Model Name	RawNet2	RawBoost
Equal Error Rate (EER) [%]	25.5	81.0
Threshold in EER	-5.74	$2.77 * 10^{-6}$

* These are crucial aspects of our analysis. The output of RawNet2 is generated using the LogSoftmax function, resulting in output values within the range of $(-\infty, 0]$. A higher output value indicates a greater likelihood that the input speech is fake. To determine whether the input is fake, it is necessary to establish a threshold value.

Table 7. The statistics of output values, derived from two voice datasets

Dataset Name	Index	RawNet2	RawBoost
ASVspoof 2021 DF	Min.	-9.11	$2.77 * 10^{-7}$
	Max.	0.00	1.00
	Ave.	-6.31	0.06
	Med.	-7.57	$1.02 * 10^{-6}$
DeepFake voices from fake news in MuMiN	Min.	-8.67	$1.5 * 10^{-7}$
	Max.	0.00	1.00
	Ave.	-7.88	$2.3 * 10^{-3}$
	Med.	-8.05	$1.05 * 10^{-6}$

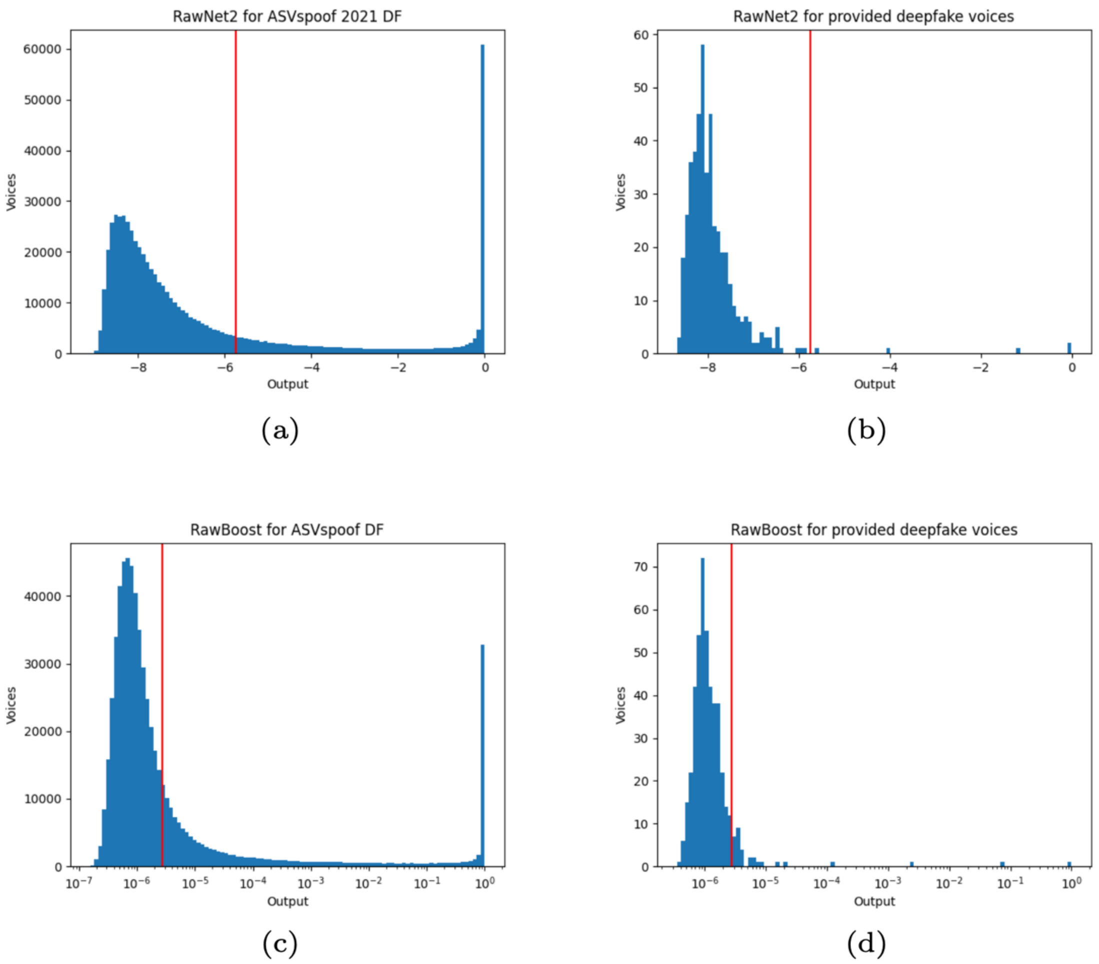
* The statistics of output values, also known as "suspiciousness", derived from two voice datasets: the DF scenario of ASVspoof 2021 and fake news from MuMiN with VITS.

Table 8. The number of DeepFake voices for which the output value exceeds the threshold value

Model Name	RawNet2	RawBoost
Greater than the threshold	5	32
Percentage of total [%]	1.08	6.88

* These threshold values correspond to those determined during the calculation of the EER.

Table 7 depicts the output values of the models. The common feature of both models is that the output values reflect the level of suspicion associated with the voice. Based on the findings from the ASVspoof dataset, we interpret values exceeding the threshold established within the ASVspoof dataset's EER as indicative of DeepFake voices. In comparison to the ASVspoof dataset, as illustrated in Figure 3, our provided voices exhibit a substantially lower rate of DeepFake voice detection. This outcome presents a significant concern, given that our dataset primarily consists of authentic voices, with only one authentic voice included. A similar trend is evident when examining the mean and median values in the Table 7, where the average and median values for our provided voices are notably smaller than those observed in the ASVspoof 2021 DF task. To provide further insight, Table 8 displays the count of DeepFake voices with outputs exceeding the threshold. The ASVspoof dataset sets the threshold, and if the output value surpasses this threshold, the voice is identified as a DeepFake. RawNet2 detects five voices as DeepFake, while RawBoost detects 32 voices as DeepFake. However, it is important to note that the reliability of RawBoost's results is questionable due to its low EER. This outcome indicates that both models identified fewer than 10% of the total samples in our provided dataset as DeepFakes. Furthermore, our thorough audio examination did not reveal any glaring artifacts that would classify any samples as obvious DeepFakes. In summary, while there are differences in the range of output values, both models exhibit a high degree of confidence in their classification of our dataset's voices as genuine, with little doubt regarding their authenticity as opposed to being DeepFakes.



(a) (b) (c) (d)

Figure 3. The histogram of output values by (a) RawNet2 for ASVspoof 2021 DF, (b) RawNet2 for our provided fake voices, (c) RawBoost for the ASVspoof, and (d) RawBoost for the provided dataset. The red lines represent the threshold value aligned with the EER situation within the ASVspoof dataset.

5. Main Experiment

We examined the newly proposed framework, taking into account waveform and speech content in the main experiment. The environment has the same conditions as the preliminary experiment, as Table 3 shows.

5.1. Dataset for Proposal Model

In the main experiment, we have curated a separate dataset that includes synthesized voices delivering both factual and fake claims in the proposed model. This new dataset employs binary labels, categorizing the synthesized voices into two classes: “factual and misinformation.” In this context, the “factual” label is assigned to synthesized voices conveying genuine news, while the “misinformation” label is attributed to synthesized voices disseminating fake news. One notable distinction between preliminary and main experiments is the ratio of factual to fake claims within the dataset. In the MuMiN-large dataset, comprising 8,013 English tweets. To address this imbalance, we have deliberately reduced the number of fakes to align more closely with the quantity of factual claims. The number of whole tweets is 723, 357 factual and 365 misinformation tweets. In addition, we removed URLs and emojis from every tweet for formatting for the voice generation. The synthesized voices used in this main experiment were obtained from the dataset under the same conditions as in the preliminary experiment, utilizing the VITS model. To facilitate experimentation and evaluation, we have divided this dataset into training, validation, and test sets in a ratio of 0.7/0.1/0.2 for testing.

5.2. Classification Settings

In this experiment, we assessed the credibility of the claims. We evaluated our novel framework, which considers both waveform and voice content. Furthermore, we introduced an additional dataset that includes synthesized voices associated with factual and fabricated claim.

In the waveform section of our proposed framework model, we employed RawNet2. We maintained the model's settings and input the voice waveform data without any fine-tuning. Subsequently, we normalized the waveform outputs to a range of zero to one, representing the suspiciousness of the waveform. For the content aspect, we utilized tweet embeddings sourced from the MuMiN dataset. We trained the XGBoost classifier using the training dataset. Finally, we computed the average of the outputs from both the waveform and content components of our proposed framework. It is important to note that the output scores indicate the degree of suspiciousness associated with the voice.

We compared the proposed framework model with single-unit models, including waveform-only and content-only models, based on EER and accuracy. We also checked the performance of RawNet2 on the ASVspoof 2019 LA set, intending to prove that the model could accurately detect DeepFake voices using previous methods. We used EER because this is one of the metrics in ASVspoof 2021 DF. We adopted accuracy to compare with the results of the MuMiN voice set and ASVspoof 2019 LA set because the ASVspoof 2019 LA set only contains only factual news text. It means we cannot get the EER of the ASVspoof dataset in content only due to a lack of the required values, True Positive and False Negative. Therefore, we set another metric, accuracy.

We also checked the performance of the SSL antispoofing model [50]. This model performs best among the participants of ASVspoof 2021 DF post-challenge [51] with the same data augmentation process as the RawBoost. We employed a pretrained model on GitHub[11].

5.3. Results

In the experiment, we examined whether the models could identify DeepFake voices associated with false claims. The results of this main experiment are presented in Table 9.

For the waveform-only model, the EER is 44%, and accuracy is close to 50%, akin to random chance. This outcome underscores the model's limited ability to generalize and detect DeepFake voices automatically. The most promising performance was observed in the proposed model that considered waveform and content. Intriguingly, our proposed framework model was demonstrated as the best. We also confirmed that the RawNet2 was classified accurately for the ASVspoof 2019 LA set. This result supports Tak *et al.*'s report [42]. The results of the SSL antispoofing show improvements from the other models in the ASVspoof results. This outcome supports the result of the challenge. However, the EER is 54% and still has room for improvement, akin to random chance. This result is worse than the result of RawNet2 only. This part also suggests that detecting DeepFakes by waveform only is not sufficient for the latest voice generation models.

Based on the average of those two datasets, the proposed method also showed the best performance in accuracy among the models, including SSL antispoofing. The other models showed better performance on only one hand, while the other has improvements. This outcome implies that our proposed framework can compete effectively with the latest voice generation models.

Table 9. The classification results for synthesized voices that make claims regarding both factual and fake content

	EER in datasets [%]		Accuracy [%]		
Model framework	MuMiN voices	ASVspoof 2019 LA	MuMiN	ASVspoof	Average
Waveform only w/ RawNet2	44.6	5.6	52.7	99.7	76.2
Waveform only w/ SSL antispoofing	54.1	2.9	46.3	99.8	73.1
Content (Embedding & XGBoost) only	32.4	N/A#	86.8	11.5	49.2
Proposed: waveform and content	**17.6**		81.7	81.6	**81.7**

* The # shows the value is uncalculatable because the dataset does not contain fake news article.

[11] https://github.com/TakHemlata/SSL_Anti-spoofing

6. Discussion

6.1. Competing the Speed of Advancing in Voice Generation

According to Table 6 pertaining to the preliminary experiment, the rate of progress in voice generation emerges as a critical issue. RawNet2, which ranks averagely in ASVspoof, fails to detect the most recent TTS fabricated voices, as does RawBoost. This outcome signifies that the model trained with the dataset incorporating TTS and VC models predating 2019 is unable to rival the models developed from 2020 onward. Notably, the generalization performance issue has been previously acknowledged in the ASVspoof 2021 report [16]. However, this study represents a significant step forward as it provides similar results specific to post-2020 methodologies. Importantly, we lack the means to evaluate the performance of the top-rated models from ASVspoof 2021. While the best-performing model may indeed excel in detecting the latest DeepFake voices, it remains uncertain whether it can effectively adapt to future developments when developers introduce entirely new voice-generation frameworks. In the main experiment, this study further examined the matter by employing synthesized voices generated by the latest TTS model. As outlined in Table 9, taking content into account can enhance the detection process, particularly if the model faces challenges in competing with the latest audio synthesis models. In the subsequent section, we delve into the enhancements introduced within the proposed framework, considering their applicability in real-life scenarios.

The result of the main experiment shows that the proposed method is the best in the performance ranking for each dataset in Table 9. This experiment aims to confirm the improvements by adopting part of the speech content consideration. Therefore, the results prove the positive effect of our concept because the proposed score was superior to the individual models in waveform and content.

6.2. Improvements of Proposed Framework

Our proposed framework model demonstrates remarkable improvements when applied to real-life scenarios. In particular, we are exploring the scenario where we deploy the target voices sourced from social media to our models. In this context, it is important to note that we face the challenge of not being able to use tweet embeddings directly. To overcome this limitation, we must employ automatic speech recognition (ASR) model to transcribe the voice content into text. Consequently, we need to ensure that our framework can effectively classify voices accurately, even when the input content consists solely of waveform data, as this may become a crucial consideration in the future.

Furthermore, there is a need to create an additional dataset that encompasses both synthesized voices and those recorded by human beings, with a focus on individuals reading fake news. It is worth mentioning that we did not incorporate the recorded voices in this article as our primary objective was to assess the proposed model's capability to detect fake claims. However, we are contemplating a separate experiment that involves 4-class classification, considering two perspectives: recorded versus synthesis voices and factual versus fake content.

7. Conclusion

The development of speech-processing technologies offers the advantage of acquiring natural-sounding voices. However, it also necessitates caution to distinguish between factual and DeepFake voices. Several efforts have been made to identify the DeepFake voices by analysing speech waveforms. Nevertheless, discerning DeepFake voices generated by the latest models remains a challenge for models trained on older voice-generation technologies. If this trend persists, detection within the current framework will continue to be fraught with risks and uncertainties. Hence, it is imperative to explore alternative detection scenarios. We proposed a model to build upon a fresh framework that considers waveform and speech content. In the experiment, we compared models that only use one perspective in voice waveform or speech content. Based on the results, our investigations prove that this model yields superior outcomes compared to waveform-only models with scores of equal error rate. This result showed the importance of incorporating speech content into the detection process. In the future, we have to assess the model's performance with a broader array of voice types than those included in the current experiment. Notably, our dataset exclusively comprises synthesized voices. Consequently, it becomes crucial to examine scenarios involving genuine speakers delivering false content and synthesized voices conveying factual claims.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Numbers JP21H03496, JP22K12157, and JP23H03688. This work was also supported by JST SPRING, Grant Number JPMJSP2131.

Contribution

Yuta Yanagi designed and planned this work, and Yuta Yanagi also accomplished experiments. Ryohei Orihara was mainly involved in the discussion of the whole project. Yasuyuki Tahara helped with the discussions. Yuichi Sei also joined the conversations and gave the inspiration for the process of the main experiment. Tanel Alumäe assisted in designing the work, especially the preliminary experiment, based on extensive research experience in voice recognition. Akihiko Ohsuga oversaw the work. We declare that all authors contribute productive critiques to advance our research, discussion, and this article.

References

[1] Wang Yuxuan, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss *et al.*, "Tacotron: Towards End-to-End Speech Synthesis", in *Proceedings of the Interspeech 2017,* 20-24 August 2017, Stockholm University, Stockholm, Sweden, Print ISBN: 978-1-5108-4876-4, pp. 4006–4010, DOI: 10.21437/Interspeech.2017-1452, Available: https://www.isca-archive.org/interspeech_2017/wang17n_interspeech.html.

[2] Shen Jonathan, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly *et al.*, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions", in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 15-20 April 2018, Calgary Telus Convention Center, Calgary, AB, Canada, Online ISBN: 978-1-5386-4658-8, E-ISBN: 978-1-5386-4657-1, pp. 4779–4783, Published by IEEE, DOI: 10.1109/ICASSP.2018.8461368, Available: https://ieeexplore.ieee.org/document/8461368.

[3] Wang Tao, Ruibo Fu, Jiangyan Yi, Jianhua Tao, Zhengqi Wen *et al.*, "Prosody and Voice Factorization for Few-Shot Speaker Adaptation in the Challenge M2voc 2021", in *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 06-11 June 2021, Metro Toronto Convention Centre, Toronto, ON, Canada, Online ISBN: 978-1-7281-7605-5, E-ISBN: 978-1-7281-7606-2, pp. 8603–8607, Published by IEEE, DOI: 10.1109/ICASSP39728.2021.9414427, Available: https://ieeexplore.ieee.org/document/9414427.

[4] Yuta Yanagi, Ryohei Orihara, Yuichi Sei, Yasuyuki Tahara and Akihiko Ohsuga, "Fake News Detection with Generated Comments for News Articles", in *Proceedings of the 2020 IEEE 24th International Conference on Intelligent Engineering Systems (INES)*, 08-10 July 2020, Reykjavík, Iceland, Online ISBN: 978-1-7281-1059-2, E-ISBN: 978-1-7281-1058-5, pp. 85–90, Published by IEEE, DOI: 10.1109/INES49302.2020.9147195, Available: https://ieeexplore.ieee.org/document/9147195.

[5] Syamsul H. Mahmud, Laromi Assan and Rashidul Islam, "Potentials of Internet of Things (IoT) in Malaysian Construction Industry", *Annals of Emerging Technologies in Computing (AETiC),* Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 44–52, Vol. 2, No. 4, 1st October 2018, DOI: 10.33166/AETiC.2018.04.004, Available: http://www.aetic.theiaer.org/archive/v2/v2n4/p4.html.

[6] Ahmad Firman and Aditya Halim Perdana Kusuma Putra, "The Effect of Social Media Utilization, Campus Environment and Entrepreneurship Knowledge on Student Entrepreneurial Interest", *Point of View Research Management*, Online ISSN: 2722-791X, pp. 131–143, Vol. 1, No. 4, 20th December 2020, Published by Ibnu Jaya Consultant, Available: http://www.journal.accountingpointofview.id/index.php/POVREMA/article/view/101.

[7] Jawaid A. Mangnejo, Arif R. Khuhawar, Muneer A. Kartio and Saima S. Soomro, "Inherent Flaws in Login Systems of Facebook and Twitter with Mobile Numbers", *Annals of Emerging Technologies in Computing (AETiC),* Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 53–61, Vol. 2, No. 4, 1st October 2018, Published by International Association of Educators and Researchers (IAER), DOI: 10.33166/AETiC.2018.04.005, Available: http://www.aetic.theiaer.org/archive/v2/v2n4/p5.html.

[8] Phillip L. De Leon, Michael Pucher, Junichi Yamagishi, Inma Hernaez and Ibon Saratxaga, "Evaluation of Speaker Verification Security and Detection of HMM-Based Synthetic Speech", *IEEE Trans Audio Speech Lang Process*, Print ISSN: 1558-7916, Online ISSN: 1558-7924, Vol. 20, No. 8, 25th May 2012, pp. 2280–2290, Published by IEEE, DOI: 10.1109/TASL.2012.2201472, Available: https://ieeexplore.ieee.org/document/6205335.

[9] Zhizheng Wu, Junichi Yamagishi, Tomi Kinnunen, Cemal Hanilçi, Mohammed Sahid *et al.*, "ASVspoof: The Automatic Speaker Verification Spoofing and Countermeasures Challenge", *IEEE Journal of Selected Topics in Signal Processing*, Print ISSN: 1932-4553, Online ISSN: 1941-0484, Vol. 11, No. 4, 17th February 2017, pp. 588–604, Published by IEEE, DOI: 10.1109/JSTSP.2017.2671435, Available: https://ieeexplore.ieee.org/document/7858696.

[10] Chen Tianxiang, Avrosh Kumar, Parav Nagarsheth, Ganesh Sivaraman and Elie Khoury, "Generalization of Audio Deepfake Detection", in *Proceedings of the Speaker and Language Recognition Workshop (Odyssey 2020)*, 1-5 November 2020, Tokyo Institute of Technology, Tokyo, Japan, pp. 132–137, DOI: 10.21437/Odyssey.2020-19, Available: https://www.isca-archive.org/odyssey_2020/chen20_odyssey.html.

[11] Ma Haoxin, Jiangyan Yi, Jianhua Tao, Ye Bai, Zhengkun Tian *et al.*, "Continual Learning for Fake Audio Detection", in *Proceedings of the Interspeech 2021*, 30th August – 3rd September 2021, Best Western Premier Hotel International Brno, Brno, Czechia, pp. 886–890, DOI: 10.21437/Interspeech.2021-794, Available: https://www.isca-archive.org/interspeech_2021/ma21b_interspeech.html.

[12] Xin Yang, Yuezun Li and Siwei Lyu "Exposing Deep Fakes Using Inconsistent Head Poses", in *Proceedings of the ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 12-17 May 2019, Brighton Conference Centre, Brighton, United Kingdom, pp. 8261–8265, Published by IEEE, DOI: 10.1109/ICASSP.2019.8683164, Available: https://ieeexplore.ieee.org/document/8683164.

[13] Haya R. Hasan and Khaled Salah, "Combating Deepfake Videos Using Blockchain and Smart Contracts", *IEEE Access*, Online ISSN: 2169-3536, pp. 41596–41606, Vol. 7, 17th March 2019, Published by IEEE, DOI: 10.1109/ACCESS.2019.2905689, Available: https://ieeexplore.ieee.org/document/8668407.

[14] Tomi Kinnunen, Md. Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans *et al.*, "The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection", in *Proceedings of the Interspeech 2017*, 20-24 August 2017, Stockholm University, Stockholm, Sweden, Print ISBN: 978-1-5108-4876-4, pp. 2–6, Published by International Speech Communication Association, DOI: 10.21437/Interspeech.2017-1111, Available: https://www.isca-archive.org/interspeech_2017/kinnunen17_interspeech.html.

[15] Massimiliano Todisco, Xin Wang, Ville Vestman, Md. Sahidullah, Héctor Delgado *et al.*, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection", in *Proceedings of the Interspeech 2019*, 15-19 September 2019, Messecongress Graz, Graz, Austria, pp. 1008–1012, DOI: 10.21437/Interspeech.2019-2249, Available: https://www.isca-archive.org/interspeech_2019/todisco19_interspeech.html.

[16] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino *et al.*, "ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection", in *Proceedings of the 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 16th September 2021, Online, pp. 47–54, Published by International Speech Communication Association, DOI: 10.21437/ASVSPOOF.2021-8, Available: https://www.isca-archive.org/asvspoof_2021/yamagishi21_asvspoof.html.

[17] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi *et al.*, "ASVspoof 2015: The First Automatic Speaker Verification Spoofing and Countermeasures Challenge", in *Proceedings of the Interspeech 2015*, 6-10 September 2015, Internationales Congress Center Dresden, Dresden, Germany, pp. 2037–2041, Published by International Speech Communication Association, DOI: 10.21437/Interspeech.2015-462, Available: https://www.isca-archive.org/interspeech_2015/wu15e_interspeech.html.

[18] Chengzhe Sun, Shan Jia, Shuwei Hou and Siwei Lyu, "AI-Synthesized Voice Detection Using Neural Vocoder Artifacts", in *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 17-24 June 2023, Vancouver Convention Center, Vancouver, BC, Canada, Online ISBN: 979-8-3503-0249-3, E-ISBN: 979-8-3503-0250-9, pp. 904–912, Published by IEEE, DOI: 10.1109/CVPRW59228.2023.00097, Available: https://ieeexplore.ieee.org/document/10208955.

[19] Yuta Yanagi, Ryohei Orihara, Yuichi Sei, Yasuyuki Tahara, Tanel Alumäe *et al.*, "Inspection of the Classifying Performance of the Deepfake Voices by the Latest Text-to-Speech Model", in *Proceedings of the II. Interdisciplinary Conference on Mechanics, Computers and Electrics (ICMECE 2022)*, 6-7 October 2022, Barcelona East School of Engineering, Polytechnic University of Catalonia, Barcelona, Spain, Online ISBN: 978-605-70842-1-7, pp. 330–335, Published by Erol Kurt, Available: http://www.icmece.org/2022/proceedings.pdf.

[20] Phillip L. De Leon, Inma Hernaez, Ibon Saratxaga, Michael Pucher and Junichi Yamagishi, "Detection of Synthetic Speech for the Problem of Imposture", in *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 22-27 May 2011, Prague Congress Center, Prague, Czech Republic, Online ISBN: 978-1-4577-0537-3, Print ISBN: 978-1-4577-0538-0, pp. 4844–4847, Published by IEEE, DOI: 10.1109/ICASSP.2011.5947440, Available: https://ieeexplore.ieee.org/document/5947440.

[21] Zhizheng Wu, Xiong Xiao, Eng Siong Chng and Haizhou Li, "Synthetic Speech Detection Using Temporal Modulation Feature", in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 26-31 May 2013, Vancouver Convention Center, Vancouver, BC, Canada, Online ISBN:978-1-4799-0356-6, pp. 7234–7238, Published by IEEE, DOI: 10.1109/ICASSP.2013.6639067, Available: https://ieeexplore.ieee.org/document/6639067.

[22] Zhizheng Wu, Eng Siong Chng and Haizhou Li, "Detecting Converted Speech and Natural Speech for Anti-Spoofing Attack in Speaker Recognition", in *Proceedings of the Interspeech 2012*, 9-13 September 2012, Portland Hilton, Portland, OR, USA, Online ISBN: 978-1622-76759-5, pp. 1700–1703, Published by International Speech Communication Association, DOI: 10.21437/Interspeech.2012-465, Available: https://www.isca-archive.org/interspeech_2012/wu12c_interspeech.html.

[23] Zhizheng Wu, Tomi Kinnunen, Eng Siong Chng, Haizhou Li and Eliathamby Ambikairajah, "A Study on Spoofing Attack in State-of-the-Art Speaker Verification: The Telephone Speech Case", in *Proceedings of the 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 3-6 December 2012, Loews Hollywood Hotel, Hollywood, CA, USA, Electronic ISBN: 978-0-6157-0050-2, Print ISBN: 978-1-4673-4863-8, pp. 1–5, Published by IEEE, Available: https://ieeexplore.ieee.org/abstract/document/6411897.

[24] Federico Alegre, Ravichander Vipperla, Asmaa Amehraye and Nicholas Evans, "A New Speaker Verification Spoofing Countermeasure Based on Local Binary Patterns", in *Proceedings of the Interspeech 2013*, 25-29 August 2013, Lyon Convention Center, Lyon, France, Online ISBN: 978-1629-93443-3, pp. 940–944, Published by International Speech Communication Association, DOI: 10.21437/Interspeech.2013-291, Available: https://www.isca-archive.org/interspeech_2013/alegre13_interspeech.html.

[25] Federico Alegre, Asmaa Amehraye and Nicholas Evans, "Spoofing Countermeasures to Protect Automatic Speaker Verification from Voice Conversion", in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 26-31 May 2013, Vancouver Convention Center, Vancouver, BC, Canada, Online ISBN:978-1-4799-0356-6, pp. 3068–3072, Published by IEEE, DOI: 10.1109/ICASSP.2013.6638222, Available: https://ieeexplore.ieee.org/document/6638222.

[26] Jesús Villalba and Eduardo Lleida, "Preventing Replay Attacks on Speaker Verification Systems", in Proceedings of the 2011 Carnahan Conference on Security Technology, 18-21 October 2011, Barcelona, Spain, Online ISBN:978-1-4577-0903-6, Print ISBN:978-1-4577-0902-9, pp. 1–8, Published by IEEE, DOI: 10.1109/CCST.2011.6095943, Available: https://ieeexplore.ieee.org/document/6095943.

[27] Nicholas Evans, Tomi Kinnunen and Junichi Yamagishi, "Spoofing and Countermeasures for Automatic Speaker Verification", in *Proceedings of the Interspeech 2013*, 25-29 August 2013, Lyon Convention Center, Lyon, France, Online ISBN: 978-1629-93443-3, pp. 925–929, Published by ISCA, DOI: 10.21437/Interspeech.2013-288, Available: https://www.isca-archive.org/interspeech_2013/evans13_interspeech.html.

[28] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre *et al.*, "Spoofing and Countermeasures for Speaker Verification", *Speech Communication*, Print ISSN: 0167-6393, Online ISSN: 1872-7182, pp. 130–153, Vol. 66, No. C, February 2015, Published by Elsevier, DOI: 10.1016/j.specom.2014.10.005, Available: https://www.sciencedirect.com/science/article/abs/pii/S0167639314000788.

[29] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans *et al.*, "End-to-End Anti-Spoofing with RawNet2", in *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 06-11 June 2021, Metro Toronto Convention Centre, Toronto, ON, Canada, Online ISBN: 978-1-7281-7605-5, E-ISBN: 978-1-7281-7606-2, pp. 6369–6373, DOI: 10.1109/ICASSP39728.2021.9414234, Published by IEEE, Available: https://ieeexplore.ieee.org/document/9414234.

[30] Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma *et al.*, "ADD 2022: The First Audio Deep Synthesis Detection Challenge", in *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 23-27 May 2022, Sands Expo and Convention Center, Singapore, Singapore, Online ISBN:978-1-6654-0540-9, Print ISBN: 978-1-6654-0541-6, pp. 9216–9220, Published by IEEE, DOI: 10.1109/ICASSP43922.2022.9746939, Available: https://ieeexplore.ieee.org/document/9746939.

[31] Junichi Yamagishi, Takao Kobayashi, Yuji Nakano, Katsumi Ogata and Juri Isogai. "Analysis of Speaker Adaptation Algorithms for HMM-Based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm", *IEEE Transactions on Audio, Speech, and Language Processing*, Print ISSN: 1558-7916, Online ISSN: 1558-7924, pp. 66–83, Vol. 17, No. 1, 6th January 2009, Published by IEEE, DOI: 10.1109/TASL.2008.2006647, Available: https://ieeexplore.ieee.org/document/4740153.

[32] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch *et al.*, "ASVspoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech", *Computer Speech and Language*, Print ISSN: 0885-2308, Online ISSN: 1095-8363, Print ISSN: 0885-2308, pp. 101114, Vol. 64, 2020, Published by Elsevier, DOI: 10.1016/j.csl.2020.101114, Available: https://www.sciencedirect.com/science/article/pii/S0885230820300474.

[33] Heiga Ze, Andrew Senior and Mike Schuster, "Statistical Parametric Speech Synthesis Using Deep Neural Networks", in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 26-31 May 2013, Vancouver Convention Center, Vancouver, BC, Canada, Online ISBN:978-1-4799-0356-6, pp. 7962–7966, Published by IEEE, DOI: 10.1109/ICASSP.2013.6639215, Available: https://ieeexplore.ieee.org/document/6639215.

[34] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals *et al.*, "WaveNet: A Generative Model for Raw Audio", in *Proceedings of the 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 13-15 September 2016, Plug and Play Tech Center, Sunnyvale, CA, USA, pp. 125, Published by International Speech Communication Association, Available: https://www.isca-archive.org/ssw_2016/vandenoord16_ssw.html.

[35] Kalchbrenner Nal, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande *et al.*, "Efficient Neural Audio Synthesis", in *Proceedings of the 35th International Conference on Machine Learning*, 10-15 July 2018, Stockholmsmässan, Stockholm Sweden, Online ISSN: 2640-3498, pp. 2410–2419, Published by MLResearchPress, Available: https://proceedings.mlr.press/v80/kalchbrenner18a.html.

[36] Ron J Weiss, RJ Skerry-Ryan, Eric Battenberg, Soroosh Mariooryad and Diederik P Kingma, "Wave-Tacotron: Spectrogram-Free End-to-End Text-to-Speech Synthesis", in *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 06-11 June 2021, Metro Toronto Convention Centre, Toronto, ON, Canada, Online ISBN: 978-1-7281-7605-5, E-ISBN: 978-1-7281-7606-2, pp. 5679–5683, Published by IEEE, DOI: 10.1109/ICASSP39728.2021.9413851, Available: https://ieeexplore.ieee.org/document/9413851.

[37] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao *et al.*, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech", in *Proceedings of the International Conference on Learning Representations (ICLR),* 4th May 2021, Vienna, Austria, pp. 1–15, Published by ICLR, Available: https://iclr.cc/virtual/2021/poster/2919.

[38] Jeff Donahue, Sander Dieleman, Mikolaj Binkowski, Erich Elsen and Karen Simonyan, "End-to-End Adversarial Text-to-Speech", in *Proceedings of the International Conference on Learning Representations (ICLR),* 4th May 2021, Vienna, Austria, pp. 1–23, Published by ICLR, Available: https://iclr.cc/virtual/2021/oral/3498.

[39] Jaehyeon Kim, Jungil Kong and Juhee Son. "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech", in *Proceedings of the 38th International Conference on Machine Learning*, 18th July 2021, Online, pp. 5530–5340, Published by MLResearhPress, Available: https://proceedings.mlr.press/v139/kim21f.html.

[40] Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang *et al.*, "NaturalSpeech: End-to-End Text-to-Speech Synthesis with Human-Level Quality", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Print ISSN: 0162-8828, Online ISSN: 1939-3539, pp. 1-12, 19th January 2024, Published by IEEE, DOI: 10.1109/TPAMI.2024.3356232, Available: https://ieeexplore.ieee.org/document/10409539.

[41] Jee weon Jung, Seung bin Kim, Hye jin Shim, Ju ho Kim and Ha Jin Yu, "Improved RawNet with Feature Map Scaling for Text-Independent Speaker Verification Using Raw Waveforms", in *Proceedings of the Interspeech 2020,* 25-29 October 2020, Shanghai International Convention Center, Shanghai, China, pp. 1496–1500, Published by International Speech Communication Association, DOI: 10.21437/Interspeech.2020-1011, Available: http://www.interspeech2020.org/index.php?m=content&c=index&a=show&catid=283&id=618.

[42] Hemlata Tak, Madhu Kamble, Jose Patino, Massimiliano Todisco and Nicholas Evans, "Rawboost: A Raw Data Boosting and Augmentation Method Applied to Automatic Speaker Verification Anti-Spoofing", in *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* , 23-27 May 2022, Sands Expo and Convention Center, Singapore, Singapore, Online ISBN:978-1-6654-0540-9, Print ISBN: 978-1-6654-0541-6, pp. 6382–6386, Published by IEEE, DOI: 10.1109/ICASSP43922.2022.9746213, Available: https://ieeexplore.ieee.org/document/9746213.

[43] Jee weon Jung, Hee Soo Heo, Ju ho Kim, Hye jin Shim and Ha Jin Yu, "RawNet: Advanced End-to-End Deep Neural Network Using Raw Waveforms for Text-Independent Speaker Verification", in *Proceedings of the Interspeech 2019,* 15-19 September 2019, Messecongress Graz, Graz, Austria, pp. 1268–1272, Published by International Speech Communication Association, DOI: 10.21437/Interspeech.2019-1982, Available: https://www.isca-archive.org/interspeech_2019/jung19b_interspeech.html.

[44] Mirco Ravanelli and Yoshua Bengio, "Speaker Recognition from Raw Waveform with SincNet", in *Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT)*, 18-21 December 2018, Royal Olympic Hotel, Athens, Greece, Online ISBN:978-1-5386-4334-1, Print ISBN:978-1-5386-4333-4, pp. 1021–1028, Published by IEEE, DOI: 10.1109/SLT.2018.8639585, Available: https://ieeexplore.ieee.org/document/8639585.

[45] Mirco Ravanelli and Yoshua Bengio, "Learning Speaker Representations with Mutual Information", in *Proceedings of the Interspeech 2019,* 15-19 September 2019, Messecongress Graz, Graz, Austria, pp. 1153–1157, Published by International Speech Communication Association, DOI: 10.21437/Interspeech.2019-2380, Available: https://www.isca-archive.org/interspeech_2019/ravanelli19_interspeech.html.

[46] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, "Rectifier Nonlinearities Improve Neural Network Acoustic Models", in *Proceedings of the International Conference on Machine Learning*, 17-19 June 2013, Atlanta, Georgia, USA, Available: http://robotics.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf.

[47] Dan S Nielsen and Ryan McConville, "MuMiN: A Large-Scale Multilingual Multimodal Fact-Checked Misinformation Social Network Dataset", in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 11-15 July 2022, Círculo de Bellas Artes, Madrid, Spain, Online ISBN: 978-1-4503-8732-3, pp. 3141–3153, Published by ACM, DOI: 10.1145/3477495.3531744, Available: https://dl.acm.org/doi/10.1145/3477495.3531744.

[48] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed *et al.*, "BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5-10 July 2020, Online, pp. 7871–7880, Published by ACM, DOI: 10.18653/v1/2020.acl-main.703, Available: https://aclanthology.org/2020.acl-main.703.

[49] Tianqi Chen and Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 13–17 August 2016, San Francisco California USA, Online ISBN: 978-1-4503-4232-2, pp. 785–794, Published by ACM, DOI: 10.1145/2939672.2939785, Available: https://doi.org/10.1145/2939672.2939785.

[50] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee weon Jung, Junichi Yamagishi *et al.*, "Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation", in *Proceedings of the Speaker and Language Recognition Workshop (Odyssey 2022)*, 28 June - 1 July 2022, 17th June 2022, Beijing, China, pp. 112–119, Published by International Speech Communication Association, DOI: 10.21437/ODYSSEY.2022-16, Available: https://www.isca-archive.org/odyssey_2022/tak22_odyssey.html.

[51] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Hector Delgado *et al.*, "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild", *IEEE/ACM Trans Audio Speech Lang Process,* Print ISSN: 2329-9290, Online ISSN: 2329-9304, pp. 2507–2522, Vol. 31, 19th June 2023, Published by IEEE, DOI: 10.1109/TASLP.2023.3285283, Available: https://ieeexplore.ieee.org/document/10155166.