SceneFake: An Initial Dataset and Benchmarks for Scene Fake Audio Detection

Jiangyan Yi[a,*], Chenglong Wang[a,b], Jianhua Tao[a,*], Zhengkun Tian[a], Cunhang Fan[c], Haoxin Ma[a], Ruibo Fu[a]

[a] *Institute of Automation, Chinese Academy of Sciences*
[b] *University of Science and Technology of China*
[c] *Anhui University*

Abstract

Previous databases have been designed to further the development of fake audio detection. However, fake utterances are mostly generated by altering timbre, prosody, linguistic content or channel noise of original audios. They ignore a fake situation, in which the attacker manipulates an acoustic scene of the original audio with another forgery one. It will pose a major threat to our society if some people misuse the manipulated audio with malicious purpose. Therefore, this motivates us to fill in the gap. This paper designs such a dataset for scene fake audio detection (SceneFake). A manipulated audio in the SceneFake dataset involves only tampering the acoustic scene of an utterance by using speech enhancement technologies. We can not only detect fake utterances on a seen test set but also evaluate the generalization of fake detection models to unseen manipulation attacks. Some benchmark results are described on the SceneFake dataset. Besides, an analysis of fake attacks with different speech enhancement technologies and signal-to-noise ratios are presented on the dataset. The results show that scene manipulated utterances can not be detected reliably by the existing baseline models of ASVspoof 2019. Furthermore, the detection of unseen scene manipulation audio is still challenging.

Keywords: Scene manipulated audio, fake audio detection, SceneFake dataset,

*Corresponding author
Email addresses: jiangyan.yi@nlpr.ia.ac.cn (Jiangyan Yi), jhtao@nlpr.ia.ac.cn (Jianhua Tao)

speech enhancement.

2010 MSC: 00-01, 99-00

---

1. Introduction

Speech signals contain rich information in real life scenarios. A spectrogram of an example utterance is shown in Figure 1. It involves not only timbre trait, prosody feature, linguistic content and channel noise but also acoustic scene and other information. Acoustic scene is a kind of acoustic environments [1]. It describes the location of happened events in the audio, such as bus, park, airport, metro station and so on [2]. The acoustic scene of the example utterance is *Airport*. If the scene of an original audio is manipulated with another scene, authenticity and integrity verification of the audio will be unreliable and even the semantic meaning of the original audio will be changed. The goal of speech enhancement is to remove noise signals and estimate a target clean speech from a noisy audio. During the past few years speech enhancement technologies has made significant progress with the development of deep learning [3, 4]. The acoustic scene of an audio is able to be effectively eliminated by speech

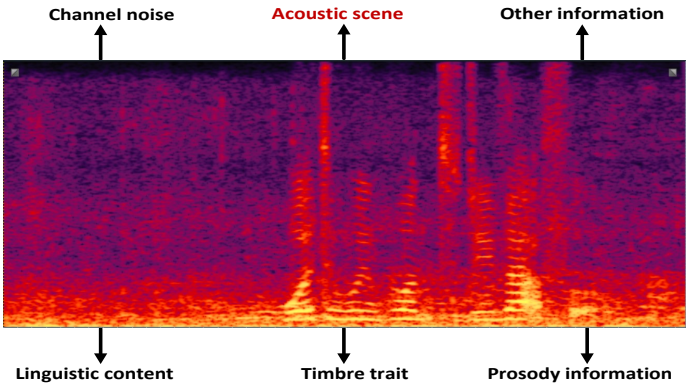

Figure 1: Spectrogram of an example utterance "*What do we want to do that for?*". It involves rich information, such as timbre trait, prosody feature, linguistic content, channel noise, acoustic scene and other information. The acoustic scene of the utterance is *Airport*.

enhancement models. The intelligibility and quality of the enhanced audio are very closed to the clean one [5, 6]. So it is easy and effective to manipulate the scene of an utterance with another scene using speech enhancement technologies.

It may bring great threats if someone misuses scene manipulated audios with the intent to cause harm. For one thing, it will mislead public cognition if scene manipulated audios are spread widely on social medium for unethical purposes. For another, the scene manipulated audio increases the risk of attacks for many applications [7] including intelligent wearable devices, context-aware services, robotics navigation systems, which understand the situations of user with the help of acoustic scene classification systems. Besides, the acoustic scene manipulation technology will make real-time crime localization systems unreliable [8]. For example, a victim being chased by an offender calls a police emergency call center. The actual location (i.e. car, street, living room etc.) of harassed persons may not be identified correctly if the localization system is attacked by the acoustical scene manipulation technology. In addition, it will bring challenges to audio forensics. Audio forensics is to evaluate and analyse audio recordings, which is commonly utilized for integrity verification and authenticity of the evidence in a court of law [9]. One scenario of audio forensics is to identify and rebuild crime or accident scenes. It may pose serious risks if real scenes of the recordings are tampered. So scene manipulation audio detection is of great significance. It is also nontrivial to design a scene manipulated audio dataset to help the development of this research.

In recent years a growing number of scholars [10, 11, 12, 13] have made attempts to detect fake audios. A variety of datasets have also been built to promote research on detecting fake utterances. Most of previous datasets are focused on detecting spoofed utterances for automatic speaker verification (ASV) systems. There are about four types of spoofing attacks [14]: impersonation, replay, speech synthesis and voice conversion. Few of datasets are designed for audio deepfake detection including in the ASVspoof 2021 [15] and the first

Audio Deep synthesis Detection (ADD 2022) challenge[1]. The ADD 2022 considers some ignored challenging fake scenarios in real life [16]. The deepfake utterances are roughly classified into three kinds of types: speech synthesis, voice conversion and speech manipulation. Impersonation [14] denotes human mimicking that an imitator mimics the voice timbre and prosody of a target speaker. Replay attack [17] is referred to as a form of replaying pre-recorded bona fide utterances of a target speaker to an ASV system. An example is the recording replayed using a smart device. Speech synthesis [14] is a technique for generating intelligible and natural speech for any arbitrary text using machine learning based models. Voice conversion [14] aims to changing from the timbre and prosody of the speech of a given speaker to that of another speaker via computer-aided technologies. Speech manipulation [18] is to generate partially fake utterances by manipulating the original genuine utterances with bona fide or synthesized audio segments.

The above mentioned datasets are crucial for accelerating research on detecting fake utterances. The datasets of ASVspoof [15] and ADD 2022 [16] challenges have especially played a significant role in promoting the development of this research. However, the fake utterances in previous datasets are mainly generated by changing timbre, prosody, linguistic content or channel noise of the original utterance. They have not covered a fake situation which involves manipulating the acoustic scene of the original audio with the another one. An example of scene manipulated audio is shown in Figure 2 (b). The scene "*Airport*" is tampered with another scene "*Public square*". Therefore, this paper is motivated to fill in the gap.

We report our progress in developing such a scene manipulated corpus involving changing the scene of an audio with another scene using speech enhancement technologies. The dataset is named scene manipulation audio detection (SceneFake) dataset. We describe a comparison of several scene manipulation detection methods to discriminate between the genuine and fake speech. A pre-

[1]http://addchallenge.cn/add2022

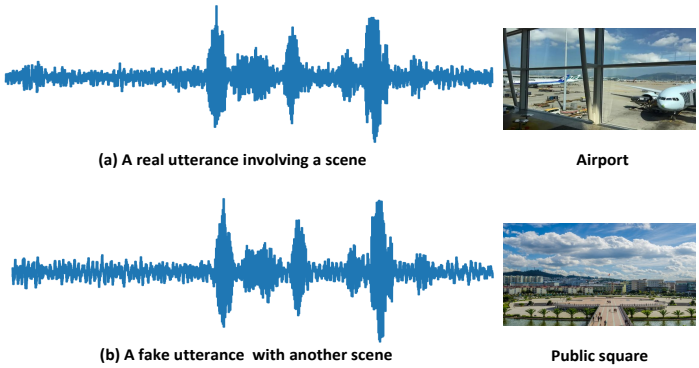

Figure 2: Waveforms of example utterances. (a) illustrates a real utterance involving a scene, such as "*Airport*". (b) shows a fake utterance: the scene of the real utterance is manipulated with another scene, such as "*Public square*".

liminary set of benchmark results for detecting fake utterances is presented in this paper. Furthermore, we conduct an analysis of fake attacks with various speech enhancement technologies on our designed SceneFake dataset.

The main contributions of this paper are as follows. To the best of our knowledge, this is the first attempt to pose such an audio fake attack using speech enhancement technologies and design a scene manipulated audio dataset for fake audio detection. The SceneFake dataset provides the speech enhancement technology information of the fake utterances. In addition, the dataset includes seen and unseen test sets. Researchers can evaluate the performance and generalization of a fake detection model. The SceneFake dataset will be publicly available soon [2].

The rest of this paper is organized as follows. Section 2 introduces the related work. Section 3 describes the design policy of the dataset. Evaluation metric is introduced in Section 4. Section 5 presents experiments and baselines. Section 6 discusses the results and future work. This paper is concluded in Section 7.

[2]If you want to obtain the dataset for non-commercial use before the manuscript is accepted, please contact the authors via the email address given in this manuscript.

2. Related Work

Previously there have existed many fake audio detection datasets and acoustic scene classification corpus.

2.1. Fake Audio Detection Datasets

Most of previous spoofed datasets are focused on developing countermeasures for automatic speaker verification systems, which mainly include four kinds of spoofing types [14]: impersonation, replay, speech synthesis and voice conversion. In 2004, an impersonation database is developed by Lau et al. [19], which is used for investigating the vulnerability of speaker verification. In 2013, a small Finnish impersonation dataset have been designed by Hautamaki et al. [20]. A few individual spoofing datasets are designed for speaker verification systems, in which the spoofing types only involve a kind of speech synthesis method[21, 22] or a sort of voice conversion approach [23, 24, 25, 26]. The spoofing types are not diverse. So some researchers develop spoofing databases include multiple sorts of spoofing attacks in 2013. Wu et al. [27] design a spoofing dataset including replay and a simple voice conversion attack. Alegre et al. [25] also develop a database consisting of artificial signal spoofing attacks, which involve a kind of voice conversion and one speech synthesis method. But the varieties of spoofing techniques are still limited. In 2015, a standard spoofing database SAS is designed by Wu et al. [28], which is comprised of various speech synthesis and voice conversion methods. The SAS database is used for supporting the first ASVspoof challenge [29], which has been organized for detecting the spoofed speech in 2015. Replay is a lowcost and challenging attack. Therefore, the dataset is developed only including replay attack in the ASVspoof 2017 challenge [17]. The ASVspoof 2019 database [22] is comprised of replay, speech synthesis and voice conversion attacks. Previous datasets in ASVspoof challenges aim to detect unforseen attack in microphone channel. Lavrentyeva et al. [30] design a PhoneSpoof dataset for speaker verification systems, in which the utterances are collected in telephone channels. A database [31] is designed

for speaker verification systems, where the spoofed audios are randomly generated via voice activity detection technologies.

Recently, a few attempts have been made to develop datasets mainly for fake audio detection systems. Reimao et al. [32] design a dataset for synthetic speech detection. The fake utterances generated by the open-sourced tools only using the latest speech synthesis technology. An English and Mandarin fake dataset are built with an open-sourced voice conversion and speech synthesis systems [11]. Frank et al. [33] develop a fake dataset named WaveFake, which contains fake utterances generated by the latest speech synthesis models. ASVspoof 2021 [15] incudes audio deepfake attacks except for replay, speech synthesis and voice conversion spoofing methods. However, these datasets have not covered many real-life challenging situations. The ADD 2022 challenge was motivated to fill the gap [16]. The ADD 2022 consists of various datasets including fully fake utterances contained various noises, partially fake utterances, and adversarial examples. Some partially fake utterances are selected from the HAD dataset designed by Yi et al. [18], which are generated by manipulated the original utterances with genuine or synthesized audio segments. Some adversarial examples are provided by the participants of generation task (Track 3.1) in the ADD 2022.

The above-mentioned datasets have played a key role in accelerating the development of anti-spoofing and audio deepfake detection. However, the fake utterances in these datasets mainly involve changing timbre, prosody, linguistic content or channel noise of the original audio. They do not consider the fake situation manipulating the acoustic scene of the original audio with the forgery one.

2.2. Acoustic Scene Classification Corpus

Acoustic scene classification has a wide range of applications ranging from audio recording integrity authentication to real-time crime identification [34]. It attempts to recognize acoustic scene labels of audio signals, such as an airport or a park environment. A diverse set of corpus are designed to identify acoustic

scenes. In 1997, a dataset consists of five scenes, which is developed by Sawhney and Maes et al. [35]. Sawhney et al. [35] design another database, in which the audio segments are recorded by wearing a microphone while riding the bike to a supermarket in 1998. There are a series of available datasets for acoustic scene classification, which are provided by the detection and classification of acoustic scenes and events (DCASE) challenges [36]. A development dataset is comprised of ten kinds of acoustic scenes each with 10 segments of 30 in DCASE 2013 [7]. In DCASE 2016, 15 scenes are used in the development set, each with 78 audio examples of 30 s [37]. The acoustic scene classification task of DCASE 2017 is held by providing 312 audio segments of 10 s per scene [38]. The acoustic scene classification dataset of DCASE 2022 contains recordings in 10 kinds of acoustic scenes using 4 different devices, which are collected from 12 European cities [39].

These datasets are critical for acoustic scene classification tasks. The output labels of the acoustic scene classification system will become no longer unreliable if the scene of the original audio is tampered by speech enhancement technologies. Different from these datasets, this paper aims to design a acoustic scene manipulated dataset named SceneFake for further promoting research on fake audio detection.

3. Dataset Design

Our scene manipulation audio detection (SceneFake) dataset is developed based on the logical access (LA) dataset of ASVspoof 2019 and the acoustic scene dataset from DCASE 2022 challenge. The LA dataset[3] consists of genuine and spoofed audio segments involving synthetic utterances and converted voices. It consists of three sets: training, development and test set. The acoustic scene dataset[4] contains 64 hours of 10-seconds audio segments from 10 acoustic scenes.

[3]https://datashare.ed.ac.uk/handle/10283/3336
[4]https://zenodo.org/record/6337421

Table 1: The description of acoustic scenes in the DCASE 2022 Challenge.

Scene	Description
Airport	Airport
Bus	Travelling by a bus
Park	Urban park
Public	Public square
Shopping	Indoor shopping mall
Station	Metro station
Metro	Travelling by an underground metro
Pedestrian	Pedestrian street
Street	Street with medium level of traffic
Tram	Travelling by a tram

Table 2: The statistics of LA dataset of ASVspoof 2019.

Set	#Speakers	#Genuine	#Spoofed	#Total
Training	20	2, 580	22, 800	25, 285
Development	20	2, 548	22, 296	24, 844
Test	67	7, 355	63, 882	71, 237

The description of 10 acoustic scenes in the DCASE 2022 Challenge is reported in the Table 1. The statistics of the LA dataset are listed in Table 2.

3.1. Design Policy

The SceneFake dataset is designed to evaluate and analyze the methods of detecting scene manipulated utterances. The dataset consists of genuine and fake utterances involving difference scenes. The core of the SceneFake dataset is the acoustic scene manipulated audio which is referred as to a fake utterance. An example of acoustic scene manipulation for a fake audio audio is illustrated in Figure 3. The generation procedure of manipulation is comprised of two steps:

1. Enhancing the real speech involving a scene

2. Adding another scene to the enhanced speech

The SNR of the real utterance and the fake utterance are denoted by SNR_real and SNR_fake, respectively. In order to simplify the problem, The SNR_real is identical to the corresponding SNR_fake for the manipulation in this work. For instance, the SNR_real and SNR_fake are both 5dB in Figure 3.

3.2. Real Utterance Collection

In the field of speech enhancement, researchers [40, 41, 42] usually simulate real noisy utterances by mixing clean utterances with noise signals at various signal-to-noise ratios (SNRs) to overcome the difficulty in obtaining clean utterances corresponding to noisy ones. This inspires us to generate a noisy speech by mixing a clean utterance with a acoustic scene. A simulated noisy speech is viewed as a real utterance involving a scene in the SceneFake dataset.

We add a variety of scenes to clean utterances to simulate genuine utterances. The clean voices come from bona fide utterances of the LA dataset in ASVspoof 2019. The bona fide utterances are from the voice cloning toolkit (VCTK) corpus [43]. It is a multi-speaker English speech database recorded in a hemi-anechoic chamber. The utterances of VCTK are under clean conditions, which are downsampled to 16 kHz at 16 bits-per-sample. The acoustic scenes come from the dataset in DCASE 2022 challenge as shown in Table 1.

The real utterances of our training, development and test sets are generated based upon the bona fide ones of training, development and test sets from the LA dataset, respectively. They are generated by randomly adding acoustic scenes to clean utterances at 6 different SNR_real each -5dB, 0dB, 5dB, 10dB, 15dB and 20dB.

3.3. Scene Manipulation for Fake Audio

Scene manipulation consists of two steps as shown in Figure 3. The first step is to remove the scene of the simulated real speech using speech enhancement technologies. In the last step the enhanced speech is added with another scene.

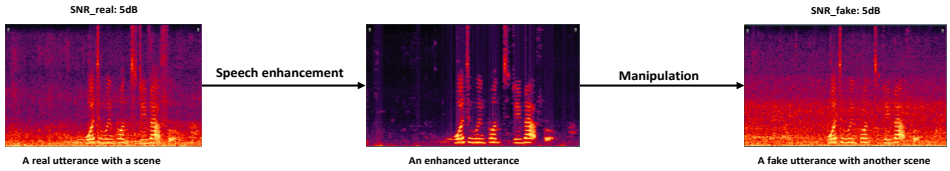

Figure 3: An example of acoustic scene manipulation for a fake utterance. The manipulation procedure consists of two steps: (1) Enhancing the real speech involving a scene, such as "*Airport*". (2) Adding another scene to the enhanced speech, such as "*Street*". The signal noise ratio (SNR) of the real utterance is denoted by SNR_real. The SNR of the fake utterance is referred as to SNR_fake. The SNR_real and SNR_fake are both 5dB in the example.

.

The scene manipulation for a fake speech is represented as:

$$\widehat{x}_{enhanced}(t) = \mathrm{SE}(x_{real}(t)) \tag{1}$$

$$y_{fake}(t) = \widehat{x}_{enhanced}(t) + n_{scene}(t) \tag{2}$$

where $\mathrm{SE}(\cdot)$ denotes the speech enhancement function, which aims to estimate the clean target speech from the real speech $x_{real}(t)$, $\widehat{x}_{enhanced}(t)$ is the enhanced speech, $n_{scene}(t)$ denotes target acoustic scene, $y_{fake}(t)$ is referred as to the fake audio whose scene is manipulated with another one.

We employ several recently state-of-the-art neural network based speech enhancement methods and some traditional open-source speech enhancement models to remove the scene of the real speech. A commonly used open source speech enhancement model is utilized to enhance the speech. The model trained using a full-band and sub-band fusion model, named as FullSubNet [5], which outperforms the top-ranked methods in the deep noise suppression (DNS) challenge [44]. A time-domain based speech enhancement method, named WaveU-Net [45], is employed to enhance the real audio. The WaveU-Net model is migrated and implemented based on an end-to-end source separation in the time-domain, which allows modelling phase information and avoids fixed spectral transformations. A gated convolutional recurrent network (GCRN) is employed to remove the scene [46]. The GCRN model obtains better performance

than an existing convolutional neural network in terms of both objective speech intelligibility and quality. Furthermore, we compare the performance of neural network based models with other traditional open-source speech enhancement methods [41], such as spectral subtraction (SSub), minimum mean square error (MMSE), and Wiener filtering (Wiener). The above-mentioned speech enhance models are public available. Thus, the reproducibility is ensured.

The fake utterances are generated by mixing another randomly sampled acoustic scenes with the enhanced utterances at 6 different SNR_fake each -5dB, 0dB, 5dB, 10dB, 15dB and 20dB. Fake utterances are also generated by using an open-source toolkit Augly.

3.4. Datasets Composition

There are five sets in the SceneFake dataset: training, development, seen test and unseen test. There are no overlaps among the speakers of training, development and seen test set. We design two unseen test sets to evaluate the generalization of the models. Data structure and detailed configurations of acoustic scene manipulation in our SceneFake dataset are listed in Figure 4.

The dataset consists of real and fake utterances with various scenes. The real and fake utterances of our training, development and seen test sets are with six kinds of acoustic scenes: Airport, Bus, Park, Public, Shopping, Station. The real and fake utterances of our unseen test sets are with four kinds of acoustic scenes: Metro, Pedestrian, Street, Tram. The acoustic scenes are randomly sampled to mix with the utterances at 6 different SNRs each -5dB, 0dB, 5dB, 10dB, 15dB and 20dB.

The statistics of the SceneFake dataset are shown in Table 3. The statistics of real and fake utterances in our SceneFake dataset at different SNRs are reported in Table 4 and 5.

4. Evaluation Metrics

The goal of audio scene manipulation detection is to develop a method or an algorithm to discriminate between the manipulated audio and the genuine one.

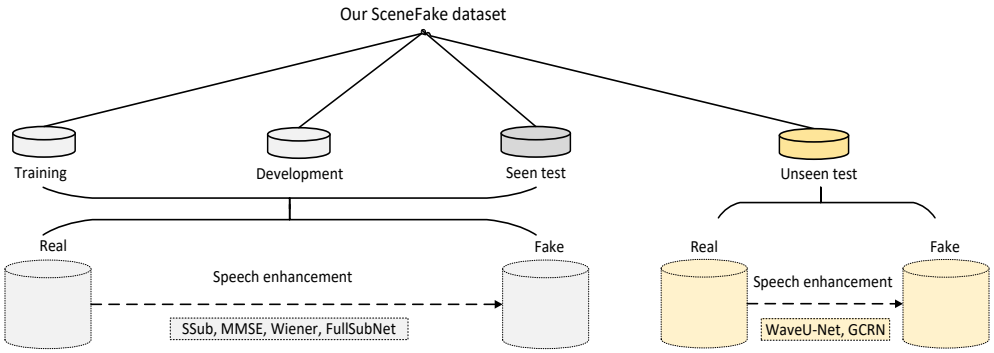

Figure 4: Data structure of the SceneFake dataset. It consists of five sets: training, development, seen test, unseen test 1 and unseen test 2 sets.

Table 3: The statistics of our SceneFake dataset.

Set	#Speakers	#SE	#Scenes	#Real	#Fake	#Total
Training	20	4	6	2, 580	10, 320	12, 900
Development	20	4	6	2, 548	10, 192	12, 740
Seen test	67	4	6	7, 355	29, 420	36, 775
Unseen test	67	2	4	7, 355	14, 710	22, 065

So equal error rate (EER) [29] is used as the evaluation metric for the detection tasks. Previously, EER is used in the ASVspoof challenges and ADD 2022 challenge. A real-valued, finite numerical value is assigned to each trial. It reflects the support for two competing hypotheses, namely that the trial is a bona fide audio or a manipulated one. But we do not optimize a decision threshold, and thus neither produce hard decisions. High detection score should indicate a genuine utterance and low score should indicate a manipulated utterance. The metric in this paper is the 'threshold-free' EER, defined as follows. Let $P_{fa}(\theta)$ and $P_{miss}(\theta)$ denote the false alarm and miss rates at threshold θ.

$$P_{fa}(\theta) = \frac{\#\{manipulated\ trials\ with\ score > \theta\}}{\#\{total\ manipulated\ trials\}} \tag{3}$$

$$P_{miss}(\theta) = \frac{\#\{genuine\ trials\ with\ score < \theta\}}{\#\{total\ genuine\ trials\}} \tag{4}$$

Table 4: The statistics of real utterances in our SceneFake dataset at 6 SNRs.

Set	Real						
	#-5dB	#0dB	#5dB	#10dB	#15dB	#20dB	#Total
Training	430	430	430	430	430	430	2, 580
Development	424	424	425	425	425	425	2, 548
Seen test	1, 226	1, 226	1, 226	1, 226	1, 226	1, 225	7, 355
Unseen test	1, 226	1, 226	1, 226	1, 226	1, 226	1, 225	7, 355

Table 5: The statistics of fake utterances in our SceneFake dataset at 6 SNRs.

Set	Fake						
	#-5dB	#0dB	#5dB	#10dB	#15dB	#20dB	#Total
Training	1, 720	1, 720	1, 720	1, 720	1, 720	1, 720	10, 320
Development	1, 696	1, 696	1, 700	1, 700	1, 700	1, 700	10, 192
Seen test	2, 452	2, 452	2, 452	2, 452	2, 452	2, 450	29, 420
Unseen test	2, 452	2, 452	2, 452	2, 452	2, 452	2, 450	14, 710

So $P_{fa}(\theta)$ and $P_{miss}(\theta)$ are, respectively, monotonically decreasing and increasing functions of θ. The EER corresponds to the threshold θ_{EER} at which the two detection error rates are equal, i.e. $EER = P_{fa}(\theta_{EER}) = P_{miss}(\theta_{EER})$. It is the lower value of EER, the better performance of the model.

5. Initial Benchmarking Experiments

A series of benchmark experiments are conducted on the SceneFake dataset.

5.1. Noisy LA Dataset

The bona fide and spoofed utterances of the LA dataset of ASVspoof 2019 are under clean conditions. However, the audio segments of our SceneFake dataset are under noisy conditions. So we also design a noisy LA dataset to evaluate whether the manipulated utterances can be detected reliably with existing audio fake detection models under the noisy conditions.

Table 6: The statistics of the generated noisy LA dataset of ASVspoof 2019.

Set	#Speakers	#Genuine	#Spoofed	#Total
Training	20	2, 580	22, 800	25, 285
Development	20	2, 548	22, 296	24, 844
Seen test	67	7, 355	63, 882	71, 237
Unseen test	67	7, 355	63, 882	71, 237

The noisy LA dataset is generated based upon the LA dataset of ASVspoof 2019. The bona fide and spoofed utterances from the LA dataset are randomly mixed with scenes from the acoustic scene dataset in DCASE 2022 challenge as shown in Table 1. The utterances of training, development and seen test set in the noisy LA dataset are generated based upon that of training, development and test set from the LA dataset, respectively. The utterances in these three sets are generated by using six scenes: Airport, Bus, Park, Public, Shopping, Station. The voices of unseen test set are simulated with four scenes: Metro, Pedestrian, Street, Tram. The acoustic scenes are randomly sampled to mix with the bona fide and spoofed utterances at 6 different SNRs each -5dB, 0dB, 5dB, 10dB, 15dB and 20dB.

We make the total number of utterances of each set be equal to that of the LA dataset so that the performance can be compared fairly. Table 6 illustrates the statistic distribution of the noisy LA dataset.

5.2. Experimental Settings

Motivated by the baselines of ASVspoof 2021 [15] and ADD 2022 challenges, we employ Gaussian mixture model (GMM), light convolutional neural network (LCNN) [47] and RawNet2 [48] to train fake audio detection models as our baselines. We use the officially released source code[5] with minor modification to build GMM, LCNN and RawNet2 models. Linear frequency cepstral coefficients (LFCCs) [49] are used as the input features of GMM and LCNN based models.

[5]http://github.com/asvspoof-challenge/2021

The input features of RawNet2 models are raw audio waveforms. The classifiers of the models are standard 2-class discriminators. The output labels are "*fake*" and "*real*". The source codes of three baseline models are public available[6].

GMM: the features are extracted using a 30-ms sliding window with a 15-ms shift, a 1024-point Fourier transform and 70 filters. LFCC features consist of 19 static cepstral coefficients appended by energy (C0 or 0^{th} cepstral coefficients) delta and delta-delta coefficients.

LCNN: LFCC features are extracted with a 20-ms sliding window with a 10-ms shift, a 1024-point Fourier transform and 70 filters. The features are comprised of 19 static cepstral coefficients plus energy, delta and delta-delta coefficients. The features are extracted with a 20-ms sliding window with a 10-ms shift. The architecture of the LCNN based models [12] incorporating average pooling and LSTM layers [15].

RawNet2: the input features are raw audio waveforms [48]. The RawNet2 is a fully end-to-end model, which operates directly upon raw audio waveforms. It comprises one fixed bank of sinc filters and six residual blocks. A gated recurrent units and fully connected layers prior followed by the output layer which has two labels: real or fake.

We only utilize the respective training data to train the models and use the respective development data to optimize the model. The development sets are employed to choose better models and hyper parameters. The training stops if there exist only a little improvement between two adjacent epochs. We also do not utilize any kind of data augmentation technology. The fake audio detection models are evaluated in terms of EER on test sets.

5.3. Performance of Speech Enhancement models

The scene manipulated audio is tampered using speech enhancement technologies in this work. We employ open source speech enhancement models to estimate an enhanced utterance by removing the original scene of a real utter-

[6]https://github.com/ADDchallenge/SceneFake

ance. A fake utterance is generated by mixing the enhanced utterance with another scene. We evaluate the performance of various speech enhancement models in this section.

The traditional speech enhancement models used in this work are SSub, MMSE and Wiener, source codes of which are public available[7]. The neural network based models are FullSubNet[8], WaveU-Net[9] and GCRN[10], which are trained using the LA and Noisy LA datasets. The results of all speech enhancement models on our SceneFake dataset are reported in terms of perceptual evaluation of speech quality (PESQ) [50, 40] and short-time objective intelligibility (STOI) [51]. PESQ is able to predict subjective quality of the enhanced speech in a very wide range of conditions. STOI is an objective intelligibility measure. The results of the models at different SNR_real on the SceneFake dataset are showed in Table 7 and Table 8. "Avg." denotes the average PESQ or STOI of different models on the according sets and "Total" is referred as to the performance of the models at all SNR_reals.

For the training, development and seen test sets, the results show that the SSub model obtains the worst performance and the FullSubNet achieves the best results. The MMSE model outperforms the SSub model but underperforms the Wiener model. The results also demonstrate that all the models obtain the worst PESQ and STOI at -5dB. The SSub, MMSE and Wiener models achieve the best PESQ at 15dB but obtain the best STOI at 20dB. The FullSubNet model obtains the best PESQ and STOI both at 20dB.

For the unseen test sets, the results demonstrate that the WaveU-Net model underperforms the GCRN model. The results also show that the WaveU-Net and GCRN model both obtain the lowest PESQ and STOI at -5dB. Besides, they both achieve the highest PESQ and STOI at 20dB.

[7] https://github.com/fchest/traditional-speech-enhancement
[8] https://github.com/haoxiangsnr/fullsubnet
[9] https://github.com/haoxiangsnr/Wave-WaveU-Net-for-Speech-Enhancement
[10] https://github.com/JupiterEthan/GCRN-complex

Table 7: The results of the speech enhancement models in terms of PESQ on our SceneFake dataset. "Avg." denotes the average PESQ of different models on the according sets. "Total" is referred as to the performance of the models at all SNR_reals.

Set	Models	PESQ						
		-5dB	0dB	5dB	10dB	15dB	20dB	Total
Training	SSub	1.23	1.88	2.08	2.35	3.15	2.86	2.26
	MMSE	1.28	1.89	2.10	2.36	3.18	2.88	2.28
	Wiener	1.37	1.98	2.26	2.44	3.28	3.06	2.41
	FullSubNet	1.73	2.27	2.68	2.84	3.25	3.38	2.75
	Avg.	1.40	2.01	2.28	2.50	3.22	3.05	2.43
Development	SSub	1.26	1.88	2.11	2.38	3.16	2.88	2.27
	MMSE	1.27	1.88	2.15	2.44	3.18	2.91	2.32
	Wiener	1.32	1.95	2.26	2.45	3.25	3.05	2.45
	FullSubNet	1.75	2.28	2.72	2.86	3.27	3.38	2.77
	Avg.	1.40	2.00	2.31	2.53	3.22	3.06	2.45
Seen test	SSub	1.26	1.87	2.05	2.37	3.17	2.81	2.24
	MMSE	1.30	1.88	2.09	2.37	3.20	2.84	2.28
	Wiener	1.38	1.96	2.27	2.49	3.26	3.06	2.46
	FullSubNet	1.78	2.27	2.67	2.85	3.28	3.39	2.77
	Avg.	1.43	2.00	2.27	2.52	3.23	3.03	2.44
Unseen test	WaveU-Net	1.56	2.05	2.39	2.65	3.31	3.17	2.44
	GCRN	1.63	2.09	2.47	2.61	3.19	3.26	2.51
	Avg.	1.60	2.07	2.43	2.63	3.25	3.22	2.48

Table 8: The results of the speech enhancement models in terms of STOI on our SceneFake dataset. "Avg." denotes the average STOI of different models on the according sets. "Total" is referred as to the performance of the models at all SNR_reals.

Set	Models	STOI						
		-5dB	0dB	5dB	10dB	15dB	20dB	Total
Training	SSub	0.58	0.62	0.64	0.83	0.86	0.87	0.74
	MMSE	0.60	0.65	0.65	0.82	0.84	0.89	0.75
	Wiener	0.62	0.66	0.67	0.82	0.86	0.89	0.75
	FullSubNet	0.68	0.72	0.78	0.91	0.94	0.95	0.84
	Avg.	0.62	0.66	0.69	0.85	0.88	0.90	0.77
Development	SSub	0.60	0.62	0.65	0.83	0.86	0.87	0.74
	MMSE	0.61	0.64	0.64	0.83	0.85	0.88	0.74
	Wiener	0.61	0.65	0.65	0.82	0.83	0.87	0.75
	FullSubNet	0.69	0.72	0.80	0.91	0.95	0.95	0.85
	Avg.	0.63	0.66	0.69	0.85	0.87	0.89	0.77
Seen test	SSub	0.61	0.61	0.65	0.81	0.84	0.88	0.75
	MMSE	0.62	0.65	0.65	0.83	0.85	0.89	0.76
	Wiener	0.62	0.66	0.67	0.82	0.86	0.88	0.76
	FullSubNet	0.68	0.70	0.79	0.92	0.94	0.96	0.84
	Avg.	0.63	0.66	0.69	0.85	0.87	0.90	0.78
Unseen test	WaveU-Net	0.65	0.68	0.72	0.83	0.89	0.91	0.77
	GCRN	0.66	0.68	0.74	0.86	0.89	0.92	0.79
	Avg.	0.66	0.68	0.73	0.85	0.89	0.92	0.78

5.4. Performance of Baseline Models of ASVspoof 2019

We conduct several groups of experiments to evaluate the performance of existing baseline models of LA task in ASVspoof 2019.

In the first group of experiments, we want to know whether the manipulated utterances can be detected effectively by the baseline models of ASVspoof. We report the results of the detection models trained using the training set of LA dataset in ASVspoof 2019. The results are reported on seen and unseen test sets of our SceneFake dataset. We also compare the results to that on test set of LA dataset and seen and unseen test sets of noisy LA dataset. The results are listed in Table 9. The results show that all the detection models obtain the best performance on the test set of LA dataset. They achieves worse performance on test sets of noisy LA dataset. Futhermore, their performance on three test sets of SceneFake dataset degrade significantly. The average EERs of the models on the three test sets of our SceneFake dataset are absolutely higher than that on the test sets of LA dataset by up to 51.71%. The results on the unseen test set of noisy LA are better than that on the seen test set. This is because that there are only four kinds of scenes contained in the unseen test set but six kinds of scenes involved in the seen test set. The results on the unseen test sets of SceneFake dataset are better than that on the seen test set.

The utterances of the LA dataset of ASVspoof 2019 are under clean conditions. However, the audio segments of our SceneFake dataset are under noisy conditions. In order to compare fairly, we assess the performance of the detection models under the noisy conditions in the second group of experiments. The detection models trained with the training set of noisy LA dataset. The performance are evaluated on seen and unseen test sets of our SceneFake dataset. Besides, the results are compared to that on seen and unseen test sets of noisy LA dataset. The results are reported in Table 10. The results demonstrate that all the detection models obtain better performance on the test sets of noisy LA dataset. The average EERs of the models on three test sets of our SceneFake dataset are absolutely higher than that on the test sets of noisy LA dataset by up to 34.67%.

Table 9: EERs (%) of the models trained with the training data of LA dataset in ASVspoof 2019 (under clean conditions).

Models	Test set				
	LA	Noisy LA seen	Noisy LA unseen	Our seen	Our unseen
GMM	8.72	15.82	13.69	65.74	51.88
LCNN	4.75	11.24	9.82	54.51	42.30
Rawnet2	5.46	13.75	10.44	53.82	44.16
Avg.	6.31	13.60	11.32	58.02	46.11

Table 10: EERs (%) of the models trained with the noisy training data of noisy LA dataset simulated upon the LA dataset.

Models	Test set			
	Noisy LA seen	Noisy LA unseen	Our seen	Our unseen
GMM	7.93	7.49	44.94	36.71
LCNN	2.86	2.36	35.85	26.94
Rawnet2	5.22	4.78	37.86	28.12
Avg.	5.34	4.88	39.55	30.59

Therefore, although the existing baseline systems of the ASVspoof 2019 obtain lowest EER, it is very difficult for them to correctly detect the scene fake utterances. Moreover, it is still hard for the existing baseline models to distinguish the real from manipulated audios, even when the models trained using the data under similar noisy environments.

5.5. Baselines on Our SceneFake dataset

We conduct experiments to evaluate the performance of baseline models on our SceneFake dataset in this section. The three baseline models are trained with the training set of our SceneFake dataset. "Avg." denotes the average EER of different baseline models on the according test sets.

EERs of our baselines are reported on seen and unseen test sets of our SceneFake dataset, which are listed in Table 11. The results show that the

Table 11: EERs (%) of the models trained with the training data of our SceneFake dataset. "Avg." denotes the average EER of different baseline models on the according test sets.

Training set	Models	Our seen test	Our unseen test
Our SceneFake	GMM	4.59	23.21
	LCNN	0.78	16.72
	Rawnet2	0.85	15.31
	Avg.	2.07	18.41

Table 12: EERs (%) of the models on our test set which manipulated with different speech enhancement models.

Models	Our seen test				Our unseen test	
	SSub	MMSE	Wiener	FullSubNet	WaveU-Net	GCRN
GMM	5.85	5.37	4.17	3.93	25.26	23.84
LCNN	1.49	0.52	0.11	0.07	17.61	16.23
Rawnet2	1.22	0.63	0.16	0.13	15.51	16.84
Avg.	2.85	2.17	1.48	1.38	19.46	18.97

models on the test set outperform that on the unseen test set. This is because the acoustic distribution of the seen test set is similar to that of the training set. But there exists mismatch between the training set and the unseen test sets.

EERs of our baselines using different speech enhancement models are reported on seen and two unseen test sets as shown in Table 12. On the test set, the results show that the better performance of the speech enhancement model the lower average EER of the detection model we obtain. The detection models with "*FullSubNet*" speech enhancement model obtain the best performance. But the detection models with "*SSub*" speech enhancement model achieve the worst results. On the unseen test set, the average EERs of the detection models with "*WaveU-Net*" speech enhancement model outperform that with "*GCRN*" speech enhancement model.

EERs of baseline models on seen and two unseen test sets at different SNRs are listed in Table 13. The results on the seen test set show that the baseline

Table 13: EERs (%) of the models on our test sets of the SceneFake dataset at different SNRs. "Avg." denotes the average EER of different baseline models on the according test sets. "Total" is referred as to the performance of the baseline models at all SNRs.

Test sets	Models	-5dB	0dB	5dB	10dB	15dB	20dB
Our seen test	GMM	4.12	4.66	5.78	4.92	5.32	5.98
	LCNN	0.00	0.05	0.42	0.05	0.41	0.79
	Rawnet2	0.23	0.17	0.72	0.14	0.77	0.93
	Avg.	1.45	1.63	2.31	1.70	2.17	2.57
Our unseen test	GMM	31.38	3.91	4.58	4.79	4.77	21.12
	LCNN	26.55	0.17	0.34	2.62	1.58	14.32
	Rawnet2	23.15	0.35	0.55	2.81	1.41	15.88
	Avg.	27.03	1.48	1.82	3.41	2.59	17.11

models achieve the best performance at -5dB but obtain the worst result at 20dB. However, the results on the two unseen test sets both show that the baseline models achieve the worst performance at -5dB but obtain the best results at 0dB. The results on the two unseen test sets also show that the models achieve obviously higher EERs at -5dB and 20dB than that at other SNRs. The possible reason may be that the baseline models try to distinguish the real utterance from the fake one mainly using acoustic scene information and distorted manipulation traces.

Acoustic scenes of real and fake utterance are the dominated signals at -5dB as shown in Figure 5, 6 and 7. In addition, manipulation traces of fake audios are the strongest at -5dB, which are brought by speech enhancement models. Signals of real and fake audios are both significantly distorted by adding scene noises at -5dB. However, speeches uttered by humans of real and fake utterance are the dominated signals at 20dB. Besides, manipulation traces of fake audios are the weakest at 20dB, which are brought by speech enhancement models as shown in Figure 5, 6 and 7. Also, signals of real and fake audios are insignificantly distorted by adding scene noises at 20dB. Therefore, the acoustic scenes are strong and the distorted manipulation traces are obvious at -5dB. But

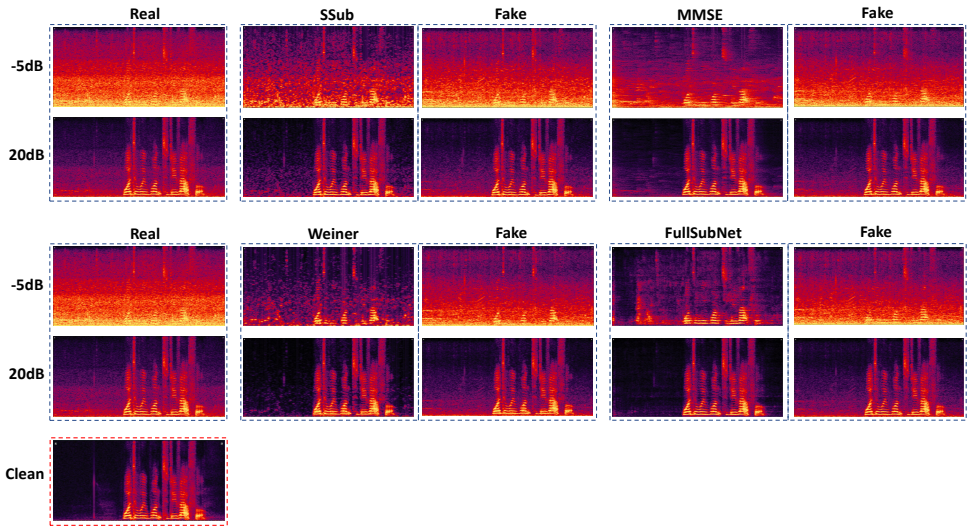

Figure 5: Spectrogram examples of utterances at -5dB and 20dB in the seen test set. "Clean" denotes the clean utterance. "Real" denotes the real utterances simulated by adding acoustic scene "*Airport*" to the clean utterance. "SSub", "MMSE", "Weiner", "FullSubNet" denote the respective enhanced utterance with the respective speech enhancement model. "Fake" denotes the respective fake utterance generated by adding scenes "*Public square*" to the respective enhanced speech.

the acoustic scenes are weak and the distorted manipulation traces are slight at 20dB.

When the data distribution of the seen test set matches the training set, it is very easy to distinguish the real one from the fake by using known strong acoustic scenes and known obvious distorted manipulation traces at -5dB. The baseline model achieves the best performance by 1.45% in terms of average EER at -5dB on the seen test set. But it is still hard to distinguish the real one from the fake by using known slight acoustic scenes and known weak distorted manipulation traces at 20dB even when the data distribution of the seen test set matches the training set. The baseline model achieves the worst result by 2.57% in terms of average EER at 20dB on the seen test set.

However, it exists serious mismatch between the unseen test set and the training set, which is brought by unseen acoustic scenes and unseen speech

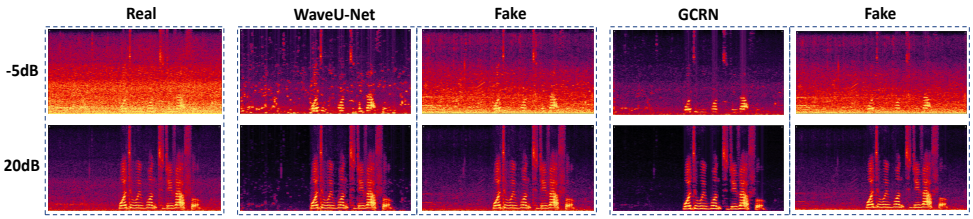

Figure 6: Spectrogram examples of utterances at -5dB and 20dB in the unseen test 1 set. "Real" denotes the real utterances simulated by adding acoustic scene "*Airport*" to the clean utterance that is the same one in Figure 5. "WaveU-Net" and "GCRN" denotes the respective enhanced utterance with the respective speech enhancement model. "Fake" denotes the respective fake utterance generated by adding scenes "*Public square*" to the respective enhanced speech.

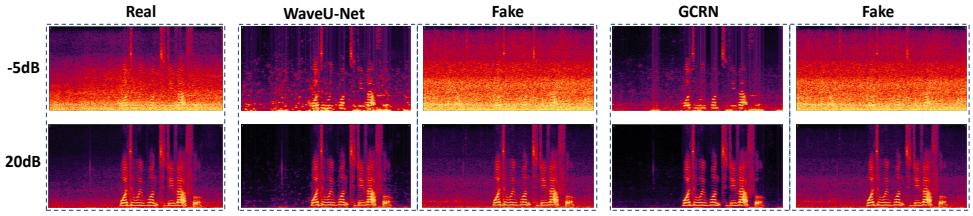

Figure 7: Spectrogram examples of utterances at -5dB and 20dB in the unseen test 2 set. "Real" denotes the real utterances simulated by adding acoustic scene "*Metro*" to the clean utterance that is the same one in Figure 5.. "WaveU-Net" and "GCRN" denote the respective enhanced utterance with the respective speech enhancement model. "Fake" denotes the respective fake utterance generated by adding scenes "*Street*" to the respective enhanced speech.

enhancement models. Thus it is difficult to distinguish the real one from the fake by using unknown strong acoustic scenes and unknown obvious distorted manipulation traces at -5dB. So the average EER of the baseline models is 27.03% at -5dB on the unseen test set. Besides, it is hard to distinguish the real one from the fake by using unknown slight acoustic scenes and unknown weak distorted manipulation traces at 20dB. So the average EER of the baseline models is 17.11% on the unseen test set at 20dB.

Although the models achieve a good performance on the seen test set, their performance are very poor on the unseen test set. Therefore, it is still challeng-

ing to effectively detect the scene manipulated fake audios.

6. Discussions

Some findings and observations of this work are summarized in this section. Furthermore, we discuss some limitations of our work. Both the findings and the limitations are suggested to be researched further.

6.1. Findings

The above benchmark results show some interesting observations and findings.

- Manipulated utterances can not be detected reliably with existing the LA baseline models of ASVSpoof 2019. Although the existing fake audio detection systems obtain low EER on the ASVspoof dataset, it is very difficult for them to detect the scene manipulated audios correctly. Moreover, it is still hard for the existing the LA baseline models of ASVSpoof 2019 to distinguish the real from manipulated audios, even when the models trained using the data under similar noisy environments.

- Scene manipulation audio detection task is still challenging. Although the models achieve a good performance on the seen test set, their performance are still very poor on the unseen test set. Therefore, it is challenging to detect reliably the scene fake fake utterances.

- The detection models achieve the best performance on seen test but obtain the worst results on unseen test sets at -5dB. When the data distribution of the seen test set matches the training set, it is very easy to discriminate the real from the fake by using known strong acoustic scenes and known obvious distorted manipulation traces at -5dB. However, it is hard to distinguish the real one from the fake at -5dB when mismatch exists between the unseen test set and the training set.

- The detection models obtain poor performance both on seen and unseen test sets at 20dB. It is difficult to tell the difference between the real and the fake utterance using unknown slight acoustic scenes or unknown weak distorted manipulation traces at 20dB. Besides, it is still hard to discriminate the real one from the fake at 20dB even when the data distribution of the seen test set matches the training set.

6.2. Limitations

Although we have designed an initial dataset and conducted benchmarking experiments for audio scene manipulation detection, it still exists some limitations which are suggested to be included in future work.

- Collecting utterances under realistic conditions: The utterances of the current SceneFake dataset are simulated data generated by mixing clean utterances with different scenes. Such emulations do not quite match with the real utterances recorded in real conditions. For instance, the linguistic content of the audio may not match up with the scene. The mismatch between the two will not happen in real recordings. Moreover, the real conditions of the utterances may be even worse and vary greatly than the simulated conditions. In order to asses manipulation detection methods in practical applications, the utterances with a variety of scenes are needed to be collected through realistic environment conditions, such as social media platforms.

- More diverse manipulation types: Our dataset here involves ten kinds of acoustic scenes and six sorts of speech enhancement methods. As a matter of fact, the acoustic scenes in the recordings are more diverse and complex in real-life scenarios. Besides, more kinds of speech enhancement technologies are utilized to manipulated the original audio. It is crucial to take more diverse fake types into consideration so that make the proposed dataset be more appropriate for real scenarios.

- Considering cross language scenarios: The current work only designs an English manipulated audio dataset and conduct benchmark experiments on such dataset. It may make the detection methods language dependent. But it is important to evaluate the performance of fake detection models in the cross language scenario and for code-switching between different languages. In order to make fake detection systems more suitable for other languages, it is necessary to design and develop multi-lingual datasets and language independent fake detection approaches countermeasures.

- Robustness and generalization of detection methods: The work here aims to provide benchmark results on the SceneFake dataset for future research. The GMM, LCNN and RawNet2 models employed in this work are the baseline models in ASVspoof and ADD 2022 challenges. Besides, although we provide the unseen test set to estimate the performance of the detection models, unseen acoustic conditions and unknown fake types can degrade the performance of the detection models. So More better features and stronger methods would be proposed to make the detection models obtain better performance. Better approaches also need to be studied to generalize well to unknown fake utterances and mismatch acoustic environments, such as continual learning and representation learning etc.

- Reasonable evaluation metrics: In this research, equal error rate (EER), which is employed in the previous ASVspoof and ADD challenges, is used as the evaluation metric in our work. However, we need to assess whether the EER is reasonable for the audio scene manipulation detection model or not in the future. We should consider human detection capabilities as well as the differences between humans and machines for detecting fake audios.

- Explainable fake utterances analysis: The aim of the current work is to distinguish the manipulated audio from the bona fide one. In addition, interpretability of detection results is needed to provide in real applications. It is nontrivial to know why the utterance is fake and find what

is the source scene of the original audio. Besides, it is also important to know what manipulation technologies are employed and even intention of the manipulation. It is particularly critical for audio forensics.

7. Conclusions

Existing literatures that we are aware of in the areas of audio fake detection, consider fake types mainly including: impersonation, speech synthesis, voice conversion, replay and audio manipulation. The fake utterances mostly generated by altering timbre, prosody, linguistic content or channel noise of the original audio. They not cover the fake type: the original scene of the utterance is manipulated by another scene. We design the first dataset that consider a fake audio that a scene of the audio is forged by another one using speech enhancement technologies. This paper presents design policy, manipulation of fake audio, evaluation metrics of the SceneFake dataset. The fake audio is enhanced by using several speech enhancement models. This paper also reported the baseline results on this dataset. The results show that it is more challenging to detect the unknown scene manipulated audio. We strongly believe that the SceneFake dataset will not only facilitate reproducible research but also further accelerate and foster research on fake audio detection and audio forensics. Future work includes the above-mentioned in the Section 6.

Acknowledgments

This work is supported by the National Key Research and Development Plan of China (No.2020AAA0140003), the National Natural Science Foundation of China (NSFC) (No.61901473, No.62101553, No.61831022).

References

References

[1] H. Zhao, H. Malik, Audio recording location identification using acoustic environment signature, IEEE Transactions on Information Forensics and Security 8 (11) (2013) 1746–1759.

[2] L. Ma, B. P. Milner, D. Smith, Acoustic environment classification, ACM Transactions on Speech and Language Processing (TSLP) 3 (2) (2006) 1–22.

[3] Y. Xu, J. Du, L.-R. Dai, C.-H. Lee, An experimental study on speech enhancement based on deep neural networks, IEEE Signal processing letters 21 (1) (2013) 65–68.

[4] A. Pandey, D. Wang, A new framework for cnn-based speech enhancement in the time domain, IEEE/ACM Transactions on Audio, Speech, and Language Processing 27 (7) (2019) 1179–1188.

[5] X. Hao, X. Su, R. Horaud, X. Li, Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 6633–6637.

[6] C. Fan, B. Liu, J. Tao, J. Yi, Z. Wen, L. Song, Deep time delay neural network for speech enhancement with full data learning, in: 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP), IEEE, 2021, pp. 1–5.

[7] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, M. Plumbley, Detection and classification of acoustic scenes and events, IEEE Transactions on Multimedia 17 (10) (2015) 1733–1746.

[8] H. Malik, Acoustic environment identification and its applications to audio forensics, IEEE Transactions on Information Forensics & Security 8 (11) (2013) 1827–1837.

[9] M. Zakariah, M. K. Khan, H. Malik, Digital multimedia audio forensics: past, present and future, Multimedia Tools & Applications.

[10] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, E. Khoury, Generalization of audio deepfake detection, in: Proc. of Odyssey: The Speaker and Language Recognition Workshop, 2020.

[11] R. Wang, F. Juefei-Xu, Y. Huang, Q. Guo, et al., Deepsonar: Towards effective and robust detection of ai-synthesized fake voices, in: Proc. of ACM MM, 2020.

[12] Z. Wu, R. K. Das1, J. Yang, H. Li, Light convolutional neural network with feature genuinization for detection of synthetic speech attacks, in: Annual Conference of the International Speech Communication Association (Interspeech), 2020.

[13] H. Ma, J. Yi, J. Tao, Y. Bai, C. Wang, Continual learning for fake audio detection, in: Annual Conference of the International Speech Communication Association (Interspeech), 2021.

[14] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, H. Li, Spoofing and countermeasures for speaker verification: A survey, Speech Communication 66 (2015) 130–153.

[15] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection.

[16] 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, H. Li, Add 2022: the first audio deep synthesis detection challenge, in: 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022.

[17] T. Kinnunen, M. Sahidullah, H. Delgado, N. E. M. Todisco, et al., The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack de-

tection, in: Annual Conference of the International Speech Communication Association (Interspeech), 2017.

[18] J. Yi, Y. Bai, J. Tao, Z. Tian, C. Wang, T. Wang, R. Fu, Half-truth: A partially fake audio detection dataset, in: Annual Conference of the International Speech Communication Association (Interspeech), 2021, pp. 1654–1658.

[19] Y. W. Lau, M. Wagner, D. Tran, Vulnerability of speaker verification to voice mimicking, in: Proc. of Int. Symposium on Intelligent Multimedia,Video and Speech Processing, 2004.

[20] R. G. Hautamaki, T. Kinnunen, V. Hautamaki, T. Leino, A. M. Laukkanen, I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry, in: Annual Conference of the International Speech Communication Association (Interspeech), 2013.

[21] P. L. D. Leon, M. Pucher, J. Yamagishi, Evaluation of the vulnerability of speaker verification to synthetic speech, in: Proc. of Odyssey: The Speaker and Language Recognition Workshop, 2010.

[22] X. Wang, J. Yamagishi1, M. Todisco1c, et al., Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech, Journal of Computer Speech and Language 64.

[23] J.-F. Bonastre, D. Matrouf, C. Fredouille, Artificial impostor voice transformation effects on false acceptance rates, in: Annual Conference of the International Speech Communication Association (Interspeech), 2007.

[24] T. Kinnune, Z.-Z. Wu, K. A. Lee, et al., Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech, in: IEEE International Conference on Acoustics, Speech and Signal, 2012.

[25] F. Alegre, A. Amehraye, N. Evans, A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary

patterns, in: Proc. of Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS), 2013.

[26] Z. Kons, H. Aronowitz, Voice transformation-based spoofing of text dependent speaker verification systems, in: Annual Conference of the International Speech Communication Association (Interspeech), 2013.

[27] Z. Wu, A. Larcher, K. A. Lee, et al., Vulnerability evaluation of speaker verification under voice conversion spoofing: the effect of text constraints, in: Annual Conference of the International Speech Communication Association (Interspeech), 2013.

[28] Z. Wu, A. Khodabakhsh, C. Demiroglu, et al., Sas : A speaker verification spoofing database containing diverse attacks, in: IEEE International Conference on Acoustics, Speech and Signal, 2015.

[29] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilc¸i, et al., Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge, in: Annual Conference of the International Speech Communication Association (Interspeech), 2015.

[30] G. Lavrentyeva1, S. Novoselov1, M. Volkova, et al., Phonespoof: A new dataset for spoofing attack detection in telephone channel, in: IEEE International Conference on Acoustics, Speech and Signal, 2019.

[31] L. Zhang, X. Wang, E. Cooper, J. Yamagishi, N. Evans, An initial investigation for detecting partially spoofed audio, in: Annual Conference of the International Speech Communication Association (Interspeech), 2021.

[32] R. Reimao, V. Tzerpos, For: A dataset for synthetic speech detection, in: 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), 2019, pp. 1–10. doi:10.1109/SPED.2019.8906599.

[33] J. Frank, L. Schnherr, Wavefake: A data set to facilitate audio deepfake detection, in: NeurIPS 2021 (Benchmark and Dataset Track), 2021.

[34] H. Malik, Acoustic environment identification and its applications to audio forensics, IEEE Transactions on Information Forensics & Security 8 (11) (2013) 1827–1837.

[35] N. Sawhney, P. Maes, Situational awareness from environmental sounds, Project Rep. for Pattie Maes (1997) 1–7.

[36] T. Heittola, A. Mesaros, T. Virtanen, Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions, in: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020), 2020, pp. 56–60.

[37] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, M. D. Plumbley, Detection and classification of acoustic scenes and events: Outcome of the dcase 2016 challenge, IEEE/ACM Transactions on Audio Speech & Language Processing 26 (2) (2017) 379–393.

[38] A. Mesaros, T. Heittola, T. Virtanen, Acoustic scene classification: An overview of dcase 2017 challenge entries, in: 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), 2018.

[39] T. Heittola, A. Mesaros, T. Virtanen, TAU Urban Acoustic Scenes 2022 Mobile, Development dataset (Mar. 2022). doi:10.5281/zenodo.6337421. URL https://doi.org/10.5281/zenodo.6337421

[40] J.-H. Changa, S. Gazorb, N. S. Kimc, S. K. Mitra, Multiple statistical models for soft decision in noisy speech enhancement, Pattern Recognition 40 (1) (2007) 1123 – 1134.

[41] P. C. Loizou, Speech enhancement: Theory and practice, CRC Press, Inc.

[42] Kolbaek, Morten, Tan, Zheng-Hua, Jensen, Jesper, Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems, IEEE/ACM Transactions on Audio Speech & Language Processing.

[43] C. Veaux, J. Yamagishi, K. Macdonald, Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit.

[44] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, et al., The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results, arXiv preprint arXiv:2005.13981.

[45] D. Stoller, S. Ewert, S. Dixon, Wave-u-net: A multi-scale neural network for end-to-end audio source separation, CoRR abs/1806.03185. arXiv:1806.03185.
URL http://arxiv.org/abs/1806.03185

[46] K. Tan, D. L. Wang, Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (1) (2019) 380–390.

[47] X. Wu, R. He, Z. Sun, T. Tan, A light cnn for deep face representation with noisy labels, IEEE Transactions on Information Forensics and Security 13 (11) (2018) 2884–2896. doi:10.1109/TIFS.2018.2833032.

[48] J. W. Jung, S. B. Kim, H. J. Shim, J. H. Kim, H. J. Yu, Improved rawnet with filter-wise rescaling for text-independent speaker verification using raw waveforms, in: Annual Conference of the International Speech Communication Association (Interspeech), 2020.

[49] M. Sahidullah, T. Kinnunen, C. Hanilçi, A comparison of features for synthetic speech detection, in: Annual Conference of the International Speech Communication Association (Interspeech), 2015.

[50] A. W. Rix, M. P. Hollier, A. P. Hekstra, J. G. Beerends, Perceptual evaluation of speech quality (pesq) the new itu standard for end-to-end speech quality assessment: Part i: Time-delay compensation, Journal of the Audio Engineering Society 50 (10) (2002) 755–764.

[51] C. H. Taal, R. C. Hendriks, R. Heusdens, J. R. Jensen, A short-time objective intelligibility measure for time-frequency weighted noisy speech, in: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2010, pp. 4214–4217.