PARTIALLY FAKE AUDIO DETECTION BY SELF-ATTENTION-BASED FAKE SPAN DISCOVERY

Haibin Wu14, Heng-Cheng Kuo2, Naijun Zheng5, Kuo-Hsuan Hung2, Hung-yi Lee1, Yu Tsao2, Hsin-Min Wang3, Helen Meng15

1 Graduate Institute of Communication Engineering, National Taiwan University, Taiwan
2 Research Center for Information Technology Innovation, Academia Sinica, Taiwan
3 Institute of Information Science, Academia Sinica, Taiwan
4 Centre for Perceptual and Interactive Intelligence, The Chinese University of Hong Kong
5 Human-Computer Communications Laboratory, The Chinese University of Hong Kong

ABSTRACT

The past few years have witnessed significant advances in speech synthesis and voice conversion technologies. However, such technologies can undermine the robustness of broadly implemented biometric identification models and can be harnessed by in-the-wild attackers for illegal uses. The ASVspoof challenge mainly focuses on synthesized audios by advanced speech synthesis and voice conversion models, and replay attacks. Recently, the first Audio Deep Synthesis Detection challenge (ADD 2022) extends the attack scenarios into more aspects. Also, ADD 2022 is the first challenge to propose the partially fake audio detection task. Such brand new attacks are dangerous and how to tackle such attacks remains an open question. Thus, we propose a novel framework by introducing the question-answering (fake span discovery) strategy with the self-attention mechanism to detect partially fake audios. The proposed fake span detection module tasks the anti-spoofing model to predict the start and end positions of the fake clip within the partially fake audio, address the model’s attention into discovering the fake spans rather than other shortcuts with less generalization, and finally equips the model with the discrimination capacity between real and partially fake audios. Our submission ranked second in the partially fake audio detection track of ADD 2022.

Index Terms— Anti-spoofing, partially fake audio detection, audio deep synthesis detection challenge

1. INTRODUCTION

The past few years have witnessed significant advances in speech synthesis and voice conversion technologies, and recently emerged adversarial attacks, such that even humans may not be capable to distinguish the real users’ speech from the synthesised speech [1–23]. Such technologies can undermine the robustness of broadly implemented biometric identification models, e.g. automatic speaker verification (ASV) models, and can be harnessed by in-the-wild attackers for criminal usage. For instance, an attacker can generate fake audios to manipulate the voiceprint-based security entrance system to accept the attacker falsely, and get access to normally protected information and valuables. Additionally, an impostor can call the bank center, fool the biometric identification system to accept him/her as a registered user, and transfer money to the impostor’s account. Considering the severe harm caused by synthesized fake audio, it is critical to devise methods to tackle such threats.

The ASVspoof challenge [1–4], a community-led challenge, arouses the attention from both the industry and the academia to tackle the spoofing audios in both physical access and logical access. In logical access, attacks are mainly from synthesized audios by advanced speech synthesis and voice conversion models, while in physical access, replayed audios are adopted as attacks. The challenge attracts various international teams, and various high-performance anti-spoofing models have been proposed to address the two kinds of attacks mentioned above. The adversarial attacks for ASV and anti-spoofing models have been well investigated [6–9, 14, 15]. To solve further challenging attack situations in realistic applications, the first Audio Deep Synthesis Detection challenge (ADD 2022) [5] extends the attack scenarios to fake audio detection. They consider the fake audios perturbed by diverse background noise, and attacks from the latest speech synthesis and voice conversion models. Additionally, the organizers propose partially fake audio detection track, where the attacks are composed of hiding small fake clips into real speech. Partially fake audios are dangerous, and ADD 2022 is the first challenge attempting to tackle this type of brand new attacks, which is an open question, and is the focus of this paper.

During generation of partially fake audio, only small clips of synthetic speech are inserted, and thus the fake audio contains a large proportion of genuine user’s audio. Through experimentation, we find it is hard to distinguish the fake and real audios by directly implementing the previous state-of-the-art spoofing countermeasure models, such as Light Convolutional Neural Network (LCNN) [24] and Squeeze-and-Excitation Network (SENet) [25]. To allow the model discover the small anomalous clip in real speech, we design a proxy task to make the model answer “where are the start and end points” of such anomalous clips. During training, the anti-spoofing model not only has to predict the fake or real label for each utterance, but also to find the start and end positions of the fake clips within the utterance. Identifying the time segments of the fake clips is similar to extraction-based question-answering, which determines the answer span in a document. Also, to further improve the capacity of the anti-spoofing model to tackle the “question-answering” task, we introduce the self-attention [26] strategy. The experimental results illustrate the effective discrimination capacity of our proposed method between real and partially fake audios.

Our main contributions are two-fold:

• We proposed a brand new framework inspired by the extraction question-answering strategy for locating the fake regions
in the fake, overall input audio, in order to improve the performance for partially fake audio detection.

- We further equipped the fake span discovery strategy with the self-attention mechanism to get a better detection capacity.

Also, our submission ranked second in the partially fake audio detection track of ADD 2022.

This paper is organized as follows: Section 2 introduces the proposed method, namely self-attention-based question-answering framework for partially fake audio detection. Section 3 presents experimental setups, followed by section 4 reporting on experimental results and analysis. Section 5 presents the conclusion.

2. METHODOLOGY

In this section, we will introduce the anti-spoofing method equipped with the proposed question-answering strategy and self-attention mechanism. We firstly present the details of the proposed framework. And then we will clarify the rationale of the proposed method.

2.1. Proposed anti-spoofing model

We adopt the base model SENet [25], which is a variant of ResNet [27] equipped with squeeze-and-excitation networks [28], and we perform some modifications to that model. The modified model architecture is shown in Table 1. Let $X = [x_1, x_2, ..., x_T]$ denote the $T$ frames of input acoustic features. The extracted hidden features by the SENet feature extractor are denoted as $f(X) = Z = [z_1, z_2, ..., z_T]$, where $f$ is the (1)-(8) layers in Table 1. The bottleneck features are denoted as $g(Z) = H = [h_1, h_2, ..., h_T]$, where $g$ is the self-attention layer, the layer (9) in Table 1, and $h_t \in \mathbb{R}^n$. The self-attention layer is one layer of transformer [26]. The question-answering layer (a) is one fully-connected layer with the input dimension as $n$, the dimension for $h_t$, and with output dimension as 2. The 2 dimensions represent how likely will $h_t$ be the start or end position of the fake clip. Given $H$ as the input, the question-answering layer will output $A = [a_1, a_2, ..., a_T]$, where $a_t \in \mathbb{R}^2$. The question-answering loss $L_{qa}$ is denoted as:

$$L_{qa} = -(log \frac{exp(a_1)}{\sum_{i=1}^{T} exp(a_i^1)} + log \frac{exp(a_2)}{\sum_{i=1}^{T} exp(a_i^2)}),$$  \hspace{1cm} (1)

where $s$ and $e$ are the start and end positions for the fake clip, $a_1^1$ and $a_2^1$ are the values for first and second dimensions of $a_t$ at the $i^{th}$ frame. We will not incorporate the $L_{qa}$ for training with real utterances. For the pooling layer (b), there are three pooling strategies in this paper, average pooling, self-attentive pooling (SAP) [29] and attentive statistics pooling (ASP) [30]. Based on the bottleneck features $H$, the pooling layer (b) followed by the prediction layer (c) will output $S = [s_0, s_1]$, indicating whether the utterance is fake or real. The anti-spoofing loss $L_{af}$ is denoted as:

$$L_{af} = -(log \frac{exp(s_1)}{\sum_{j=0}^{T} exp(s_j)}).$$  \hspace{1cm} (2)

where $l \in \{0, 1\}$ is the target label. The final loss is

$$L = L_{qa} + L_{af}.$$  \hspace{1cm} (3)

2.2. Rationale

In the partially fake audio detection track, there is only a small proportion of fake audio frames in the overall piece of input speech. Previous state-of-the-art anti-spoofing models [24, 25] tackle the problem of identifying whether a whole audio utterance is real or fake. Hence, previous strategies are not designed to identify anomalous regions within one utterance. Thus the previous models intuitively attain the ability to distinguish between utterances but there is no guarantee that such models can discover the abnormal regions within
3. EXPERIMENTAL SETUP

3.1. Data preparation

3.1.1. Dataset construction

The training set and dev set, which are based on Mandarin publicly available corpus AISHELL-3 [31], provided by the organizers of ADD 2022, cannot be directly adopted to tackle the problem of partially fake audio detection track, as the whole utterance sample in them is either real or fake. During the training phase, for constructing fake audios, we generate the partially fake audio by inserting a clip of audio into the real audios. The inserted clips are derived from three sources: 1) fake audios in the training and dev set provided by ADD 2022. 2) Real audios other than the victim audio in the training and dev set. 3) audios re-synthesised by the traditional vocoders, including Griffin-Lim [32] and WORLD [33], based on the real audios in the training and dev set. It is hard to train text-to-speech (TTS) or voice conversion (VC) models based on the limited real data provided by the organizer, so we choose the traditional vocoders, namely Griffin-Lim and WORLD, to increase the diversity of fake audios. As for the validation set, we adopt the adaptation set consisting of partially fake audios synthesised by ADD 2022 for selecting the models. We report the equal error rate (EER) for the testing set released by the organizer, as EER is the main evaluation metric for the partially fake audio detection track.

### Table 1. Proposed anti-spoofing model.

<table>
<thead>
<tr>
<th>layer</th>
<th>Type</th>
<th>Filter / Stride</th>
<th>Output shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>Conv</td>
<td>7 × 7/1 × 2</td>
<td>16 × 501 × 40</td>
</tr>
<tr>
<td>(2)</td>
<td>BatchNorm</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>(3)</td>
<td>ReLU</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>(4)</td>
<td>MaxPool</td>
<td>3 × 3/1 × 2</td>
<td>16 × 501 × 20</td>
</tr>
<tr>
<td>(5)</td>
<td>SEResNet Module × 3</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>(6)</td>
<td>SEResNet Module × 4</td>
<td>—</td>
<td>32 × 501 × 10</td>
</tr>
<tr>
<td>(7)</td>
<td>SEResNet Module × 6</td>
<td>—</td>
<td>64 × 501 × 5</td>
</tr>
<tr>
<td>(8)</td>
<td>SEResNet Module × 3</td>
<td>—</td>
<td>128 × 501 × 3</td>
</tr>
<tr>
<td>(9)</td>
<td>Self-attention</td>
<td>—</td>
<td>501 × 384</td>
</tr>
<tr>
<td>(a)</td>
<td>Question-answering</td>
<td>—</td>
<td>501 × 2</td>
</tr>
<tr>
<td>(b)</td>
<td>Pooling</td>
<td>—</td>
<td>384</td>
</tr>
<tr>
<td>(c)</td>
<td>Prediction</td>
<td>—</td>
<td>2</td>
</tr>
</tbody>
</table>

### Table 2. The EERs with (w/l) or without (w/o) self-attention.

<table>
<thead>
<tr>
<th>FFT window size</th>
<th>w/o attention</th>
<th>w/ attention</th>
</tr>
</thead>
<tbody>
<tr>
<td>384</td>
<td>23.6%</td>
<td>14.3%</td>
</tr>
<tr>
<td>768</td>
<td>22.0%</td>
<td>17.9%</td>
</tr>
</tbody>
</table>

3.1.2. Input representations

Mel-spectrograms, which are based on short-time Fourier transform (STFT) where the window size of fast Fourier transform (FFT) is varied from 384 to 768, the hop size is 128, and the number of output bins is 80, are used as input features for most of our experiments and are denoted by MSTFT in following sections. Besides spectral features, some extra experiments are operated on cepstral and NN-based features to increase diversity for achieving a better performance in the stage of fusion. The FFT window size, hop size, and number of output bins are fixed to 384, 128, and 80 respectively for Mel-frequency cepstral coefficients (MFCC), linear frequency cepstral coefficients (LFCC), and SincNet [34], as we find the FFT window size of 384 performs well as shown in Table 3.

3.1.3. Data augmentation

We perform on-the-fly data augmentation by adding noise from MUSAN dataset [35], performing room impulse response (RIR) simulation [36] and applying codec algorithms (a-law and μ-law) [37].

3.2. Implementation details

The backbone model is shown in Table 1. Three kinds of attention, average pooling (Avg), attentive statistics pooling (ASP) and self-attentive pooling (SAP) are adopted for experiments. All the models are optimized by Adam with the learning rate of 0.001 and weight decay as $1e^{-4}$.

4. EXPERIMENTAL RESULTS AND ANALYSIS

We first illustrate that the question-answering (QA) strategy drastically decreases the EERs. We conduct experiments with and without the QA strategy. The experimental results show that the trained models without the QA strategy attain the EERs of around
Table 3. The EERs using MSTFT features. w/o or w/ mean with or without. w/ or w/o re-synthesis correspond to using the re-synthesised audios by Griffin-Lim and WORLD or not.

<table>
<thead>
<tr>
<th>feature</th>
<th>FFT window size</th>
<th>pooling method</th>
<th>w/o re-synthesis</th>
<th>w/ re-synthesis</th>
<th>w/o re-synthesis</th>
<th>w/ re-synthesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSTFT</td>
<td>384</td>
<td>Avg</td>
<td>14.3%</td>
<td>19.9%</td>
<td>11.9%</td>
<td>14.2%</td>
</tr>
<tr>
<td></td>
<td>512</td>
<td>Avg</td>
<td>13.2%</td>
<td>20.5%</td>
<td>13.0%</td>
<td>14.8%</td>
</tr>
<tr>
<td></td>
<td>640</td>
<td>Avg</td>
<td>18.5%</td>
<td>19.9%</td>
<td>18.9%</td>
<td>13.3%</td>
</tr>
<tr>
<td></td>
<td>768</td>
<td>Avg</td>
<td>17.9%</td>
<td>16.8%</td>
<td>14.8%</td>
<td>12.6%</td>
</tr>
<tr>
<td>MSTFT</td>
<td>384</td>
<td>SAP</td>
<td>16.9%</td>
<td>17.5%</td>
<td>15.6%</td>
<td>12.6%</td>
</tr>
<tr>
<td></td>
<td>512</td>
<td>SAP</td>
<td>17.0%</td>
<td>18.0%</td>
<td>13.9%</td>
<td>12.5%</td>
</tr>
<tr>
<td></td>
<td>640</td>
<td>SAP</td>
<td>12.1%</td>
<td>15.3%</td>
<td>15.3%</td>
<td>11.1%</td>
</tr>
<tr>
<td></td>
<td>768</td>
<td>SAP</td>
<td>15.2%</td>
<td>17.8%</td>
<td>11.7%</td>
<td>14.8%</td>
</tr>
<tr>
<td>MSTFT</td>
<td>384</td>
<td>ASP</td>
<td>17.3%</td>
<td>15.9%</td>
<td>14.9%</td>
<td>11.9%</td>
</tr>
<tr>
<td></td>
<td>512</td>
<td>ASP</td>
<td>14.9%</td>
<td>15.8%</td>
<td>12.9%</td>
<td>11.1%</td>
</tr>
<tr>
<td></td>
<td>640</td>
<td>ASP</td>
<td>17.5%</td>
<td>15.9%</td>
<td>15.8%</td>
<td>11.2%</td>
</tr>
<tr>
<td></td>
<td>768</td>
<td>ASP</td>
<td>14.8%</td>
<td>17.9%</td>
<td>14.5%</td>
<td>22.1%</td>
</tr>
</tbody>
</table>

Table 4. The EERs for three different features

<table>
<thead>
<tr>
<th>feature</th>
<th>MFCC</th>
<th>LFCC</th>
<th>SincNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>EER</td>
<td>12.5%</td>
<td>11.1%</td>
<td>16.1%</td>
</tr>
</tbody>
</table>

In order to increase diversity of models for achieving a better performance in the stage of model fusion, we further take MFCC, LFCC and SincNet as input features to train the models. We cannot exhaust all the settings due to limited computing resources, thus we refer to Table 3 to select the setting to conduct the experiments. We fix the FFT window size as 384, apply only ASP pooling, adopt data augmentation and the re-synthesised data. We observe from Table 4 that the LFCC feature gets EER as 11.1%, reaching the best single model performance in our experimental settings. For the further work, we plan to explore the potential of different front-end features to get better performance.

For the fusion method, we tried average fusion, weighted average fusion, min fusion and max fusion. The best submission, which is fused by the average scores of the top 5 models, achieves the best 7.9% EER and ranks second in partially fake audio detection track.

5. CONCLUSION

Inspired by extraction-based question answering, this paper proposes a self-attention-based, fake span discovery strategy. The proposed strategy tasks the anti-spoofing model to predict the start and end position of the fake clip within the partially fake audio, address the model’s attention into discovering the fake spans rather than other patterns with less generalization, and finally equips the model with the discriminate capacity between real and partially fake audios. Our final submitted model gave 7.9% EER, and ranked 2nd in the partially fake audio detection track of ADD 2022. Such a strategy can be model-agnostic and feature-agnostic. Our future work will explore the potential of the proposed strategy by adopting other backbone anti-spoofing models and front-end features.

6. ACKNOWLEDGEMENTS

This research is funded by the Centre for Perceptual and Interactive Intelligence, an InnoCentre of The Chinese University of Hong Kong. This work was done while Haibin Wu was a visiting student at The Chinese University of Hong Kong. Haibin Wu is supported by Google PHD Fellowship Scholarship.
7. REFERENCES


