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

Paper #: 9333

Haibin Wu[14], Heng-Cheng Kuo[2], Naijun Zheng[5], Kuo-Hsuan Hung[2]
Hung-yi Lee[1], Yu Tsao[2], Hsin-Min Wang[3], Helen Meng[45]

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

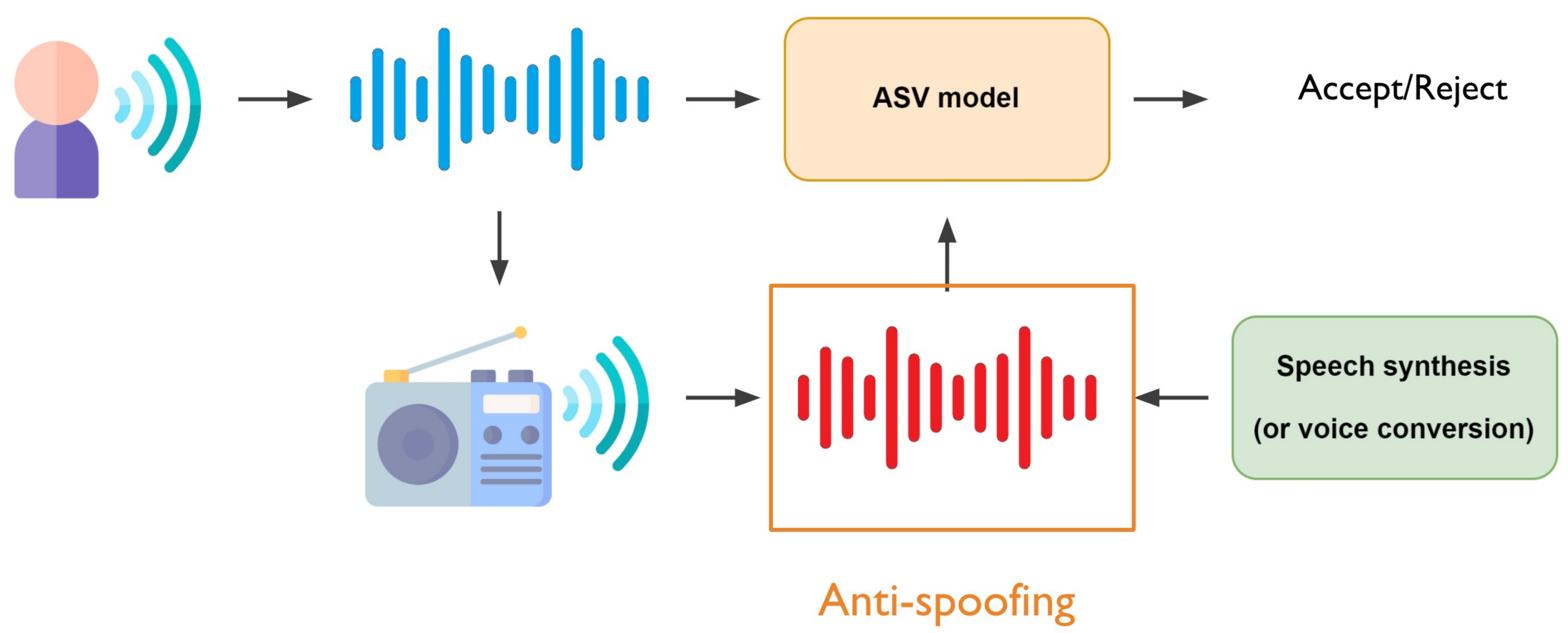
中央研究院 ACADEMIA SINICA

icassp 2022 Singapore

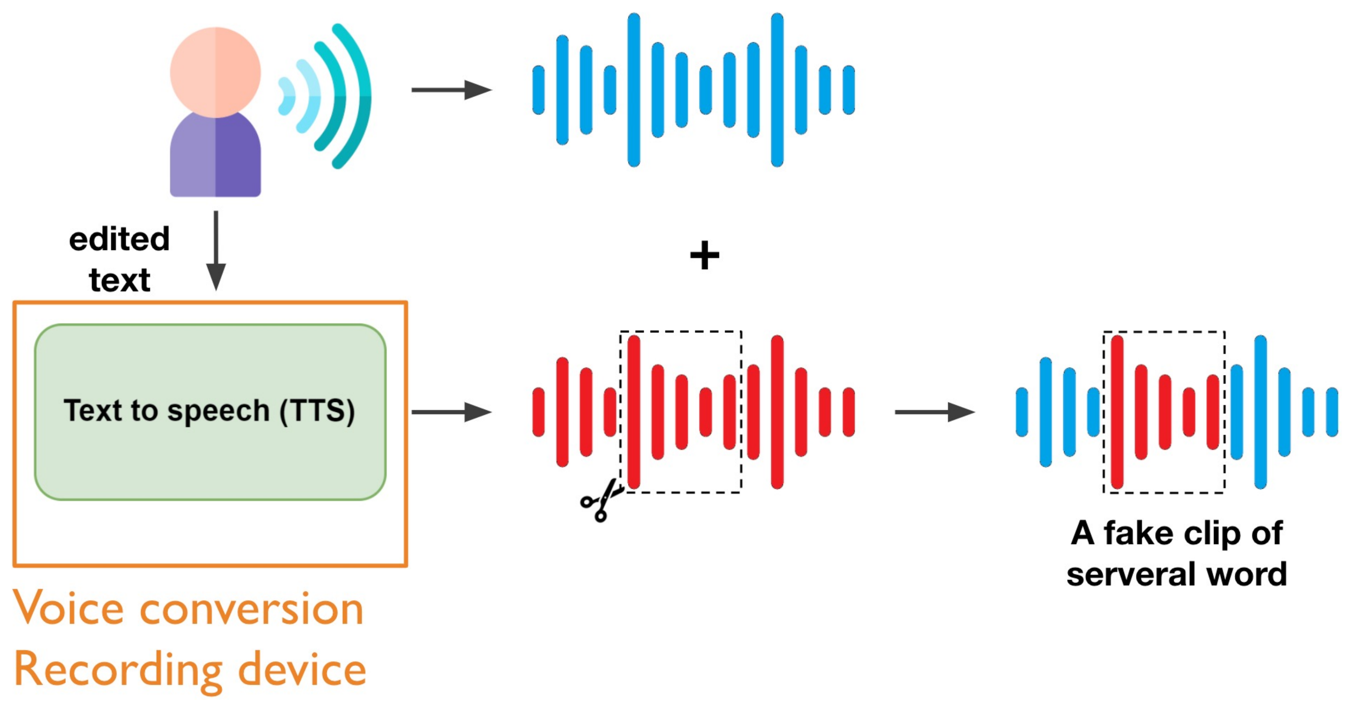
Motivation

- The significant advances in speech synthesis and voice conversion technologies can undermine the robustness of speaker verification models.
- The ASVspoof challenge arouses the attention of fostering spoofing speech detection research. However, they didn't consider partially fake audio into consideration.
- The first Audio Deep Synthesis Detection challenge (ADD2022) extends the attack scenarios to the partially fake audio detection task, which is a brand new scenario.
- However, such brand new attacks have not been well addressed. So we propose a novel method to tackle it.

Anti-Spoofing for ASV



Partially Fake Audio



Overall Architecture

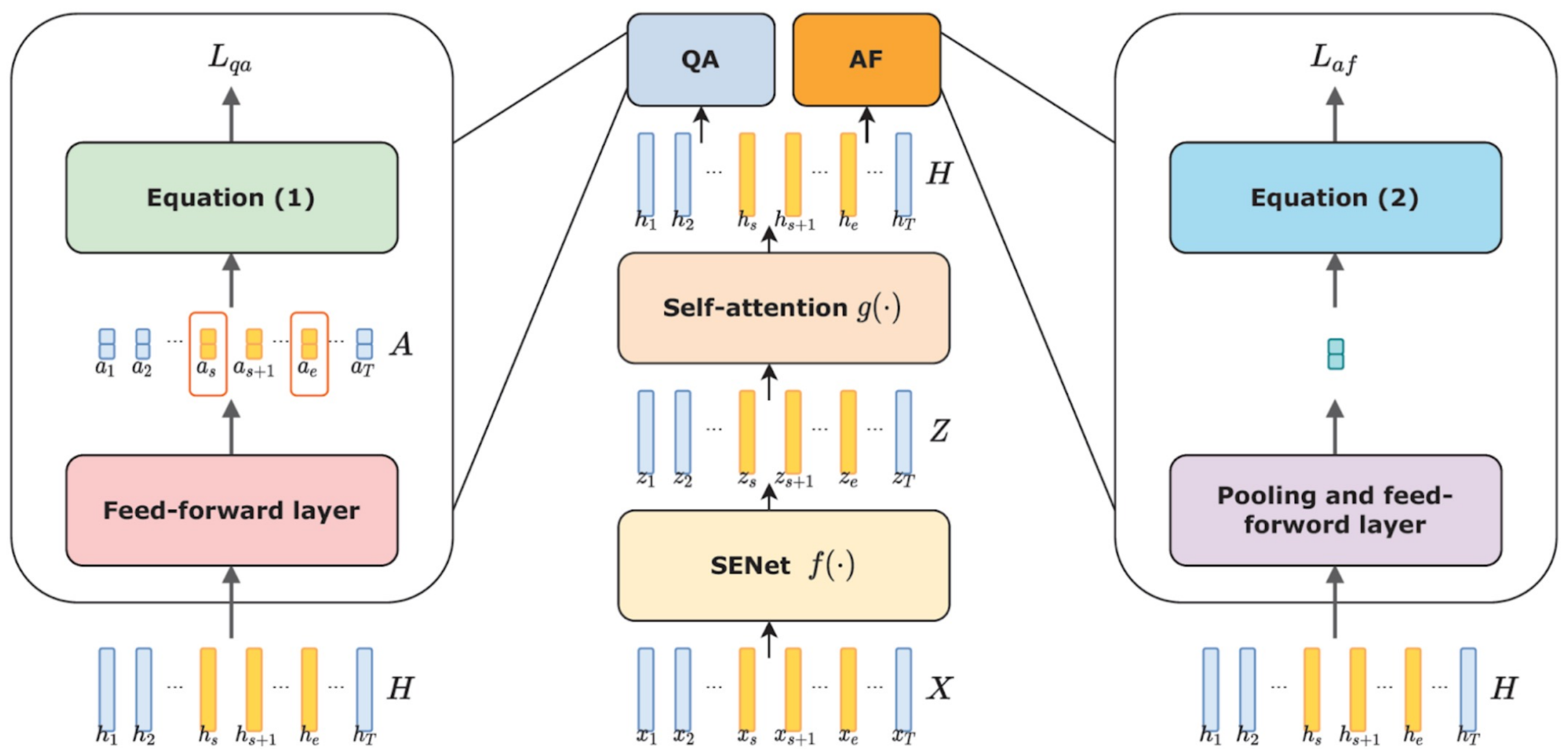


Table 1. Proposed anti-spoofing model.

layer	Type	Filter / Stride	Output shape
(1)	Conv	$7 \times 7/1 \times 2$	$16 \times 501 \times 40$
(2)	BatchNorm	–	–
(3)	ReLU	–	–
(4)	MaxPool	$3 \times 3/1 \times 2$	$16 \times 501 \times 20$
(5)	SEResNet Module×3	–	$16 \times 501 \times 20$
(6)	SEResNet Module×4	–	$32 \times 501 \times 10$
(7)	SEResNet Module×6	–	$64 \times 501 \times 5$
(8)	SEResNet Module×3	–	$128 \times 501 \times 3$
(9)	Self-attention	–	501×384
(a)	Question-answering	–	501×2
(b)	Pooling	–	384
(c)	Prediction	–	2

$$L_{qa} = -(log\frac{exp(a_s^1)}{\sum_{t=1}^{T} exp(a_t^1)} + log\frac{exp(a_e^2)}{\sum_{t=1}^{T} exp(a_t^2)})$$

$$L_{af} = -log\frac{exp(s_l)}{\sum_{j=0}^{1} exp(s_j)}$$

Rationales

- We introduce a proxy task named question-answering, or fake span discovery proxy task, in which the model has to answer "where is the fake clip" in a piece of partially fake audio.
- As a result, the proposed anti-spoofing model has to predict not only whether the input utterance is real or fake, but also output the start and end of each anomalous region.

Experimental Setup

Data Preparation

- 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:
  - Fake audios in the training and dev set provided by ADD 2022
  - Real audios other than the victim audio
  - Audios re-synthesised by the traditional vocoders, including Griffin-Lim and WORLD
- As for the validation set, we adopt the adaptation set consisting of partially fake audios synthesised by ADD 2022 for model selection.

Data Preprocessing

- Most input representations in this paper are Mel-spectrograms (MSTFTs) with hop size of 128 and output bins as 80. On the other hand, the FFT window sizes range from 384 to 768.
- We perform on-the-fly data augmentation by adding noise from MUSAN dataset, adopting room impulse response (RIR) simulation, and applying codec algorithms.

Experimental Results

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

FFT window size	w/o attention	w/ attention
384	23.6%	14.3%
768	22.0%	17.9%

- First, we verify the effectiveness of the self-attention layer (one layer of Transformer encoder).
- In two settings with FFT window sizes of 384 and 768, the improvements after adding self-attention are significant. The other settings are with the same trend.
- Therefore, the models with self-attention will be adopted for the following experiments.

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.

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

- Firstly, the EERs are improved with the help of data augmentation in most of the setups.
- Secondly, enlarging the training set by the re-synthesised data usually benefits the EERs when data augmentation is conducted.
- Lastly, the SAP and ASP poling significantly improve the EERs when both data re-synthesis and augmentation are applied.

Table 4. The EERs for three different features

feature	MFCC	LFCC	SincNet
EER	12.5%	11.1%	16.1%

- To increase the diversity of models in the fusion stage, **MFCC, LFCC, and SincNet** are further taken as input features.
- We fix the FFT window size as 384, apply only ASP pooling, adopt data augmentation, and the re-synthesised data due to limited computing resources.

The average fusion of the top 5 models achieves the best 7.9% EER and ranks second in the partially fake audio detection track.

Conclusion

- Inspired by extraction-based question-answering, this paper proposes a self-attention-based, fake span discovery strategy for partially fake audio detection.
- The proposed strategy tasks the model to predict the start and end position of the fake clip and address the attention of the model into discovering the fake span.
- The final submission achieves 7.9% EER, and ranked 2nd in the partially fake audio detection track of ADD2022.
- Our future work will explore the proposed strategy by adopting other backbone models and front-end features.

Acknowledgement

- 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 CUHK
- Google PHD Fellowship Scholarship