A Comparative Study on Proactive and Passive Detection of Deepfake Speech

GitHub

Chia-Hua Wu[1,2], Wanying Ge[1], Xin Wang[1], Junichi Yamagishi[1], Yu Tsao[2], Hsin-Min Wang[2]
[1]National Institute of Informatics, Japan
[2]Academia Sinica, Taiwan

Motivation & Introduction

Two main approaches for detecting real vs. deepfake speech

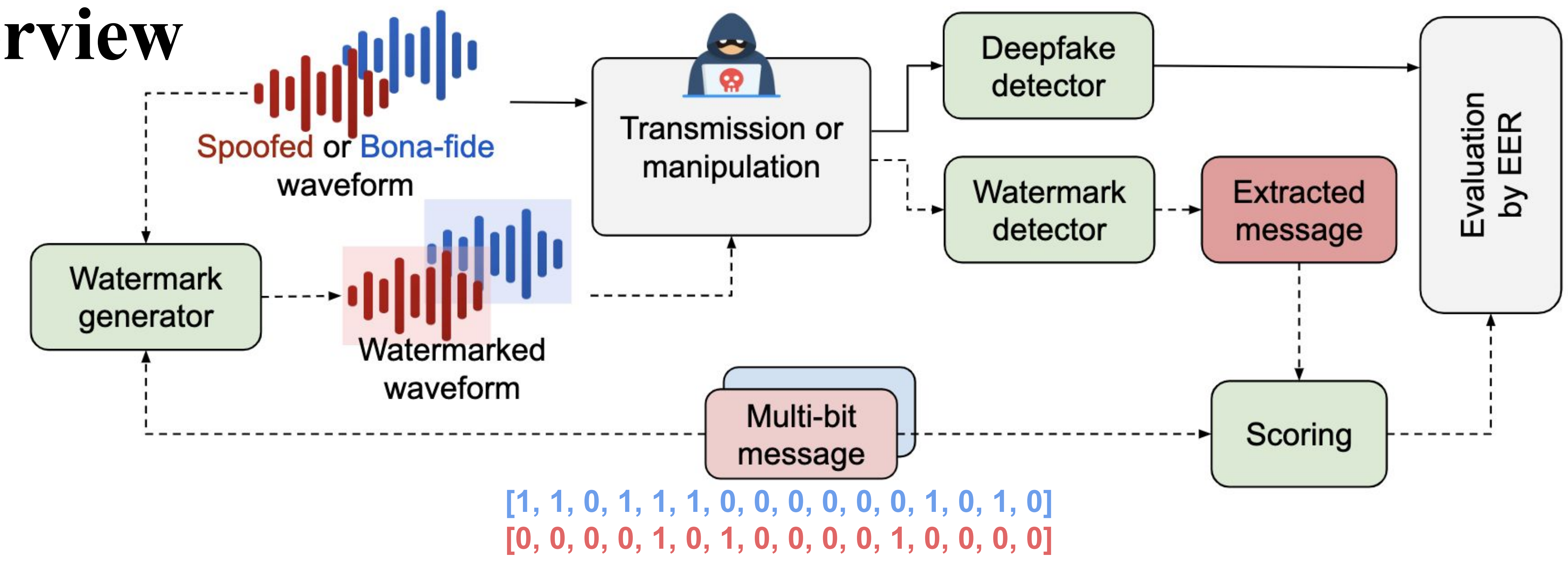
- Passive models: Directly analyze the input waveform for detection
- Proactive models: Embed a watermark into the signal to assist detection

Fair comparison is missing — no prior work has systematically compared the two approaches under identical conditions, which is essential for guiding practical adoption

Our Contributions

- First side-by-side evaluation of proactive and passive defense models **using the same training set, test set, and evaluation metrics**
- Analyze the feasibility and limitations of both models in practical **transmission** and **manipulation** scenarios

Quick Overview



Passive model (e.g., Anti-Spoofing):

- Popular models: AASIST, SSL-AASIST
- Input: real/fake speech
- Goal: detect whether audio is spoofed

Proactive model (e.g., Audio Watermarking):

- Popular models: Timbre, AudioSeal
- Input: real/fake speech with n-bit watermark message
- Goal: detect whether audio is spoofed via embedded message

Metrics (shared): Equal Error Rate (EER)

All models evaluated under identical conditions (transmission, training set, test set, metrics)

Results

Experimental setup

- Dataset: Train on ASVspoof 2019 LA training set; test on LA test set
- Models: Passive (models trained by others) vs. Proactive (retrained)
- Partially seen: Similar augmentation methods, but not used in training

Key Observations

- **Clean condition** → All models perform excellently
- **Codecs**: Opus, DAC, WavTokenizer significantly impact both model types
- **Temporal & spectral modifications**: Time stretch, Pitch shift, Random trimming significantly affect model performance

	Transmission Manipulation	EER (%)↓ of ASVspoof 2019 LA			
		Passive Models		Proactive Models	
		AASIST	SSL-AASIST	Timbre	AudioSeal
	None from § 3.3	0.83	0.23	0.00	0.00
Partially seen	Gaussian noise	18.06	1.95 ∗	17.60	15.83 ∗
	DAC	1.66	0.27	0.01	97.40 ∗
	WavTokenizer	17.84	15.92	50.12	60.95 ∗
	Random trimming	19.56 ∗	8.15	0.00	37.50
	Time stretch	66.53	44.42	0.00	0.03 ∗
	Pitch shift	66.12	48.36	52.62	47.30 ∗
Unseen	MUSAN	17.84	1.73	1.31	2.91
	RIR	35.49	4.41	0.00	57.08
	Quantization	26.15	3.31	8.66	19.59
	Compressor	9.30	1.02	0.00	0.00
	Opus	36.27	27.55	17.35	47.38
	Clipping	1.22	0.23	0.00	0.00
	Overdrive	15.30	6.19	0.11	0.00
	Equalizer	1.75	0.23	0.00	0.03
	Frequency masking	43.32	33.11	2.94	24.40
	Noise gate	10.56	2.56	0.13	2.56
	Noise reduction	17.18	11.61	0.00	0.05
Average w/o None		23.77	12.41	8.87	24.29