Audio Codec Augmentation for Robust Collaborative Watermarking of Speech Synthesis

Aalto University
School of Electrical Engineering

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

Lauri Juvela
Department of Information and Communications Engineering, Aalto University, Finland
Xin Wang
National Institute of Informatics, Tokyo, Japan

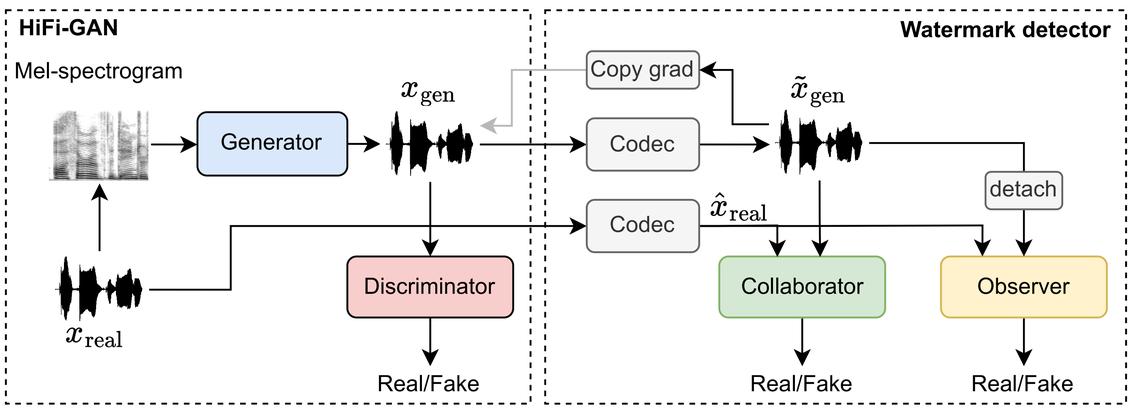
Introduction

- Text-to-speech tools with voice-cloning capability are widely available as open-source distributions or commercial services.
- Research on audio deepfake detection focuses mostly on passive protection using machine learning methods, which is also referred to as **speech anti-spoofing**.
- EU AI Act (article 50) will require marking AI-generated content
- Can speech synthesis take **active measures** to make synthetic speech easier to detect?

Watermarking for generative models

- Watermark requirements include **perceptual transparency**, **capacity**, **computational cost**, and **robustness**.
- Typical watermarking methods are separate from generative methods, and the watermark is applied as post-processing.
- In open-source settings, separate watermarks are often trivial to disable.
- **Detection can be seen as a zero-bit watermark**

Collaborative watermarking



A Generator model takes a mel-spectrogram as input and outputs a corresponding synthetic speech waveform. Detector models all try to classify between real and generated speech, and the training dynamics change based on which role the classifier takes:

1. **Discriminator** is adversarial to the Generator. The Generator attempts to fool the discriminator into classifying generated samples as real.
2. **Observer** acts as a passive detector. Gradient flow from the Observer to Generator is detached. This corresponds to traditional ASVspoof countermeasure training.
3. **Collaborator** shares its objective with the Generator. The pair attempt to discover a *watermark* that is embedded into the generated signal to aid in the binary classification task, while not hindering Generator's other objectives.

Experimental setting

- Generator and Discriminator are from HiFi-GAN, fine-tune a pre-trained model
- Detector is AASIST, fine-tune a pre-trained model
- Use straight-through estimator to get gradients for classic codecs (MP3 and OPUS) and DAC as a neural audio codec

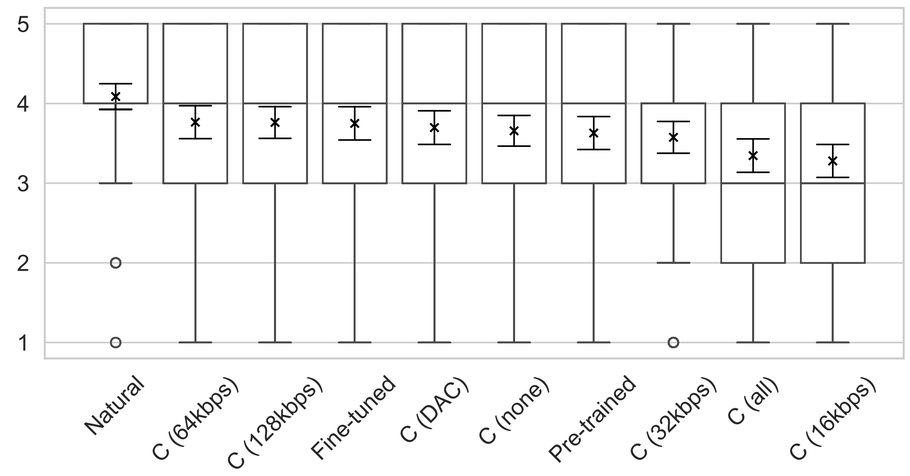
Results

How and what to read in the results table

- Detection Equal Error Rate (EER), don't look at the numbers, **look at the color**
- Columns are training conditions, Rows are test conditions
- More compression is challenging, look at rows with same codec at different rates
- Collaborative training helps, look at neighboring columns marked *o* for Observer and *c* for Collaborator
- Augmentation helps, compare the None column to various codec augmentation columns

Listening test

- Evaluate quality of clean vocoded speech using a mean opinion score (MOS) test
- Observer models correspond to fine-tuning the generator, augmentation has no effect by definition
- Collaborative detection through codecs has no significant quality impact at high bitrates and DAC
- Quality degrades when the model attempts to be robust at lower rate codecs
- With multiple bitrates quality is determined by the worst case (lowest bitrate)



Conclusions

- Actively making generated speech easier to detect can be viewed as zero-bit watermarking
- Augmentation with audio codecs improves detection robustness
- Collaborative training improves detection performance over passive Observer training
- Augmentation at high bitrates does no cause significant quality drop
- DAC augmentation generalizes surprisingly well to other codecs

Audio samples and source code

Source code and models
`https://github.com/ljuvela/collaborative-watermarking-with-codecs`

Sound samples
`https://ljuvela.github.io/collaborative-watermarking-with-codecs-demo/`

Augmentation tool
`https://github.com/ljuvela/DAREA`

codec		None		DAC		MP3/Opus									
	bitrate	-		8		16		32		64		128		all	
		o	c	o	c	o	c	o	c	o	c	o	c	o	c
None	-	1.53±0.3	0.29±0.1	2.75±0.4	0.00±0.0	33.16±0.9	3.66±0.5	9.39±0.6	14.95±0.9	3.43±0.4	0.17±0.1	1.28±0.2	0.85±0.3	3.91±0.4	4.44±0.5
DAC	8	35.85±1.0	24.29±0.9	5.09±0.5	0.00±0.0	32.81±1.0	3.97±0.5	27.54±0.9	24.68±0.9	34.46±0.9	75.54±1.2	34.03±0.9	10.03±0.6	20.86±0.8	5.62±0.5
OGG-Opus	16	53.42±1.0	41.53±1.0	34.67±0.9	31.73±0.9	37.09±1.0	5.17±0.6	37.46±1.0	36.49±1.3	47.07±1.0	89.91±0.9	44.88±1.0	41.39±0.9	40.65±1.0	10.50±0.7
OGG-Opus	32	21.75±0.8	9.18±0.6	14.55±0.7	0.72±0.2	34.30±1.0	4.01±0.5	19.29±0.9	14.76±0.9	17.12±0.8	4.09±0.4	20.55±0.8	6.74±0.5	13.64±0.7	6.59±0.6
OGG-Opus	64	4.55±0.4	0.45±0.1	5.31±0.5	0.00±0.0	33.12±1.0	3.62±0.4	10.90±0.7	13.93±0.8	5.79±0.6	0.29±0.1	3.45±0.4	0.70±0.2	5.40±0.5	4.78±0.5
OGG-Opus	128	2.27±0.3	0.21±0.1	3.12±0.4	0.00±0.0	33.24±1.0	3.62±0.5	9.32±0.6	14.58±0.8	3.85±0.5	0.14±0.1	1.61±0.3	0.79±0.2	4.07±0.4	4.44±0.5
MP3	16	49.70±1.0	48.42±1.0	30.72±0.9	6.82±0.6	34.46±1.0	4.82±0.6	36.28±1.0	78.89±1.1	48.56±1.0	79.18±1.2	48.34±1.0	52.82±1.0	37.25±0.9	8.39±0.7
MP3	32	31.96±1.0	11.49±0.7	12.16±0.6	0.52±0.2	33.64±1.0	4.07±0.5	21.73±0.9	16.13±0.8	26.32±0.9	28.76±1.2	26.24±0.9	6.59±0.5	19.14±0.8	6.00±0.6
MP3	64	6.26±0.5	0.79±0.2	4.80±0.4	0.04±0.1	33.33±0.9	3.91±0.5	12.43±0.7	14.55±0.8	7.38±0.6	0.33±0.1	5.15±0.4	0.93±0.3	6.76±0.5	4.96±0.5
MP3	128	1.92±0.3	0.29±0.1	2.85±0.3	0.00±0.0	33.18±0.9	3.89±0.4	9.53±0.6	14.86±0.8	3.56±0.5	0.17±0.1	1.34±0.3	0.89±0.3	4.13±0.4	4.78±0.5
OGG-Vorbis	q=1	46.50±1.0	27.08±0.9	35.21±1.0	30.23±1.0	34.75±0.9	5.81±0.6	37.03±1.0	21.73±0.9	47.53±1.0	55.20±1.3	46.85±1.1	32.25±0.9	35.97±1.0	12.07±0.8
OGG-Vorbis	q=2	42.17±1.0	17.39±0.8	21.94±0.9	14.39±0.8	33.18±0.9	4.20±0.5	25.49±0.9	14.99±0.8	41.10±1.0	57.70±1.4	41.70±1.0	15.09±0.7	29.07±0.9	6.08±0.6
OGG-Vorbis	q=3	38.10±0.9	9.59±0.7	15.79±0.7	9.24±0.7	33.41±1.0	3.76±0.5	21.07±0.8	13.33±0.8	35.52±1.0	24.85±1.2	34.32±0.9	5.75±0.4	23.57±0.9	4.98±0.5
All	-	28.58±0.2	16.38±0.2	19.06±0.2	9.12±0.2	33.87±0.3	4.20±0.1	21.52±0.2	17.82±0.2	24.77±0.2	31.68±0.4	26.37±0.2	13.81±0.2	20.74±0.2	6.43±0.2

Department of Information and Communications Engineering
School of Electrical Engineering
Aalto University, Finland

ICASSP 2025
Lauri Juvela, Xin Wang
Email: lauri.juvela@aalto.fi; wangxin@nii.ac.jp