Collaborative Watermarking for Adversarial Speech Synthesis

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

Aalto University
School of Electrical Engineering

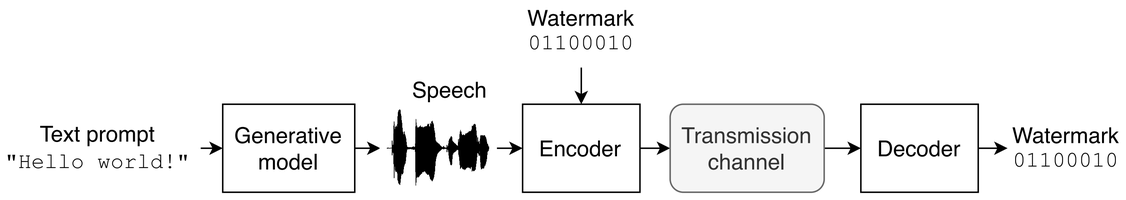
大学共同利用機関法人 情報・システム研究機構 国立情報学研究所
National Institute of Informatics

Introduction

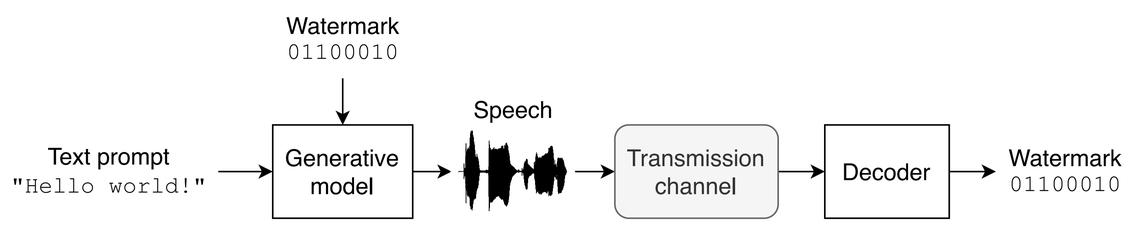
- Text-to-speech tools with voice-cloning capability are widely available as open-source distributions or commercial services.
- Research on audio deepfake detection focuses mostly on passive protection using machine learning methods, which is also referred to as **speech anti-spoofing**.
- EU AI Act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

Watermarking for generative models

- Watermark requirements include **perceptual transparency**, **payload capacity**, **computational cost**, and **robustness**.
- Typical watermarking methods are separate from generative methods, and the watermark is applied as post-processing.
- While this works well in a hosted commercial setting, in an open-source scenario separate watermarks are often trivial to disable.



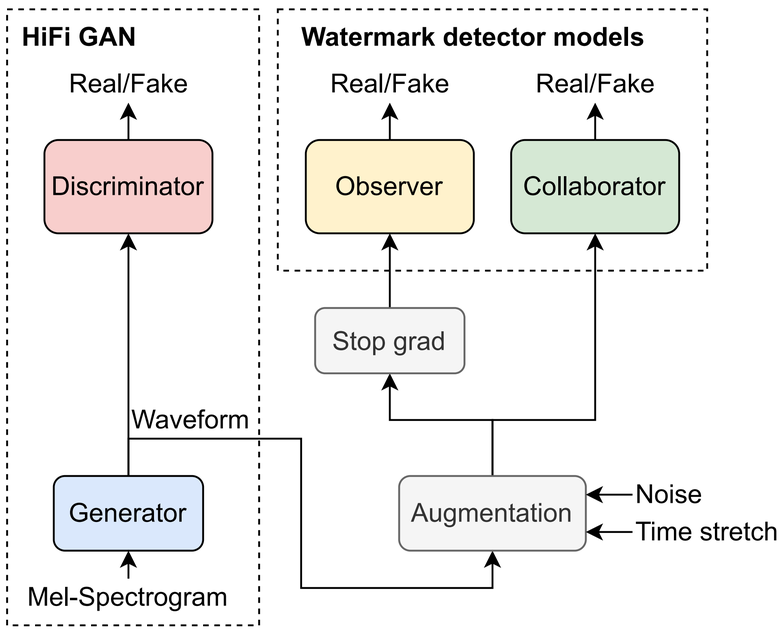
- When watermarking is integrated with the generative model, removing the watermark will take considerable effort, expertise and model retraining.



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.



Evaluation

- Metric: detection Equal Error Rate (EER)
- Detector models are baselines from ASVspoof2021 challenge
- Generator treats Detector either as Observer or Collaborator
- Detector is either trained jointly with the Generator or is a pre-trained.
- Optional channel augmentation with time-stretching and additive noise.

			Clean	Stretch	Noise	S+N				
Spectral patchwork watermarking			0.21	32.65	85.29	98.25				
WM configuration			LFCC-LCNN				RawNet			
Training	Augmentation	Role	Clean	Stretch	Noise	S + N	Clean	Stretch	Noise	S + N
Joint	None	collaborator	1.61	2.95	34.66	42.90	0.12	43.25	44.33	48.62
		observer	3.46	5.13	37.87	46.20	0.34	17.86	35.35	49.65
	Stretch + noise	collaborator	1.05	1.38	15.25	34.94	1.36	2.33	4.03	54.10
		observer	3.73	4.24	27.72	41.35	3.72	4.87	13.31	46.48
Pre-trained	None	collaborator	17.73	20.87	42.13	45.19	10.64	21.36	28.32	51.48
		observer	49.47	49.50	49.05	49.56	47.76	48.35	48.55	50.15
	Stretch + noise	collaborator	32.83	32.79	40.79	44.37	45.91	46.75	47.18	48.40

Listening test

- Evaluate quality of clean vocoded speech using a mean opinion score (MOS) test
- Each configuration was trained for 100k iterations
- Baseline HiFi-GAN was trained using the standard recipe without feedback from the detector
- Collaborative HiFi-GAN vocoders attempt to help the watermark detector model
- Adding a collaborative detection task does not have a significant impact on

Ref.			Natural recording	4.13 ± 0.11
			Spectral patchwork	3.49 ± 0.13
			Baseline HiFi-GAN	3.54 ± 0.13
Proposed	Training	Augmentation	LFCC-LCNN	RawNet
	Joint	None	3.38 ± 0.17	3.59 ± 0.13
		Stretch + noise	3.46 ± 0.13	3.60 ± 0.13
	Pre-trained	None	3.58 ± 0.13	3.51 ± 0.13
		Stretch + noise	3.46 ± 0.14	3.67 ± 0.12

Audio samples and source code

Demo page with audio samples
https://ljuvela.github.io/CollaborativeWatermarkingDemo

Source code available at
https://github.com/ljuvela/CollaborativeWatermarking

Conclusions

- This work proposes a collaborative training scheme to actively assist in **detection of synthetic speech** using a HiFi-GAN vocoder and ASVspoof 2021 baseline detector models.
- Collaborative training consistently improves detection performance over different detector model configurations and transmission channel conditions.
- Integrated watermarks are difficult to remove in an open-source setting and **add no computational cost to synthesis**

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

Contact information: Lauri Juvela
Email: lauri.juvela@aalto.fi