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

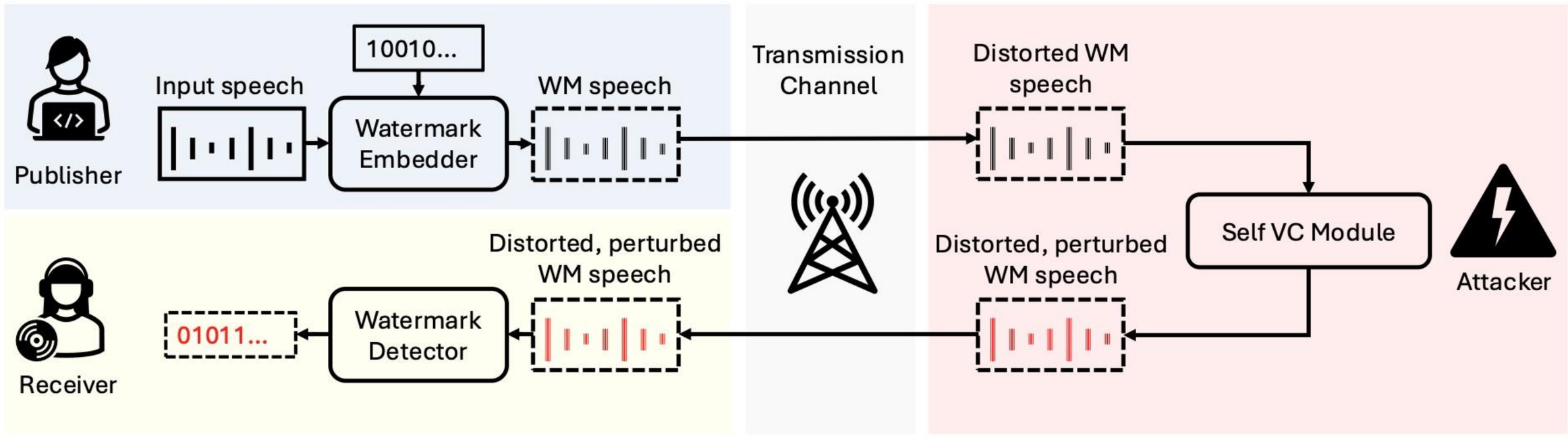
3月2日(月) 午後 ポスタ
SP/SLP1

一般社団法人
電子情報通信学会

Self Voice Conversion as an Attack against Neural Audio Watermarking

Yigitcan Özer, Wanying Ge, Zhe Zhang, Xin Wang, Junichi Yamagishi | National Institute of Informatics, Tokyo

{yiitozer, gewanying, zhe, wangxin, jyamagis}@nii.ac.jp



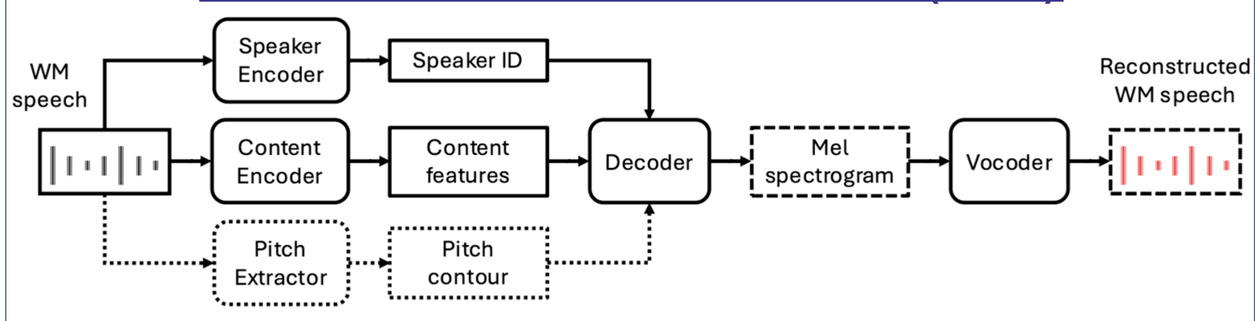
Summary

- Audio watermarking embeds auxiliary bits into speech while preserving speaker identity, linguistic content, and perceptual quality.
- Most robustness evaluations largely focus on channel distortions such as compression, noise, and resampling, which may lead to an overestimation of robustness when facing motivated attackers.
- Modern, intentional attacker scenarios remain underexplored.

Attacker goal: invalidate the watermark while keeping speech natural and attributable to the same speaker.
- Attacker uses a modern voice conversion (VC) toolkit
- Operates without access to the watermark internal details (a.k.a. *black box* setting)

Findings:
- Watermark cues are discarded while speech stays natural.
- Self VC alters fine acoustic details that watermarking often relies on.
- Defenses must survive representation-level reconstruction.

Self Voice Conversion (VC)



Key idea: Self Voice Conversion (Self VC) as an attack
- Regenerate speech while keeping speaker identity and content.
- Alter acoustic realizations via representation-level processing, suppressing watermark cues without audible degradation.

kNN-VC (VC with nearest neighbors):
- Matches frames in a self-supervised space.
- Replaces each source frame by mean of k nearest reference frames.

RVC (Retrieval-based VC):
- Involves a content encoder & optional pitch conditioning.
- Adapted for self VC via same-speaker conditioning.

Evaluation

- **Dataset:** LibriTTS test split
- Watermarking models in use:
  - **DCT:** conventional, DSP-based
  - **AudioSeal:** waveform-based
  - **Timbre:** spectral-based
  - **WMCodec:** latent-based, against neural codecs
  - **VoiceMark:** latent-based, against voice cloning attacks

Condition	Speaker Similarity ↑	WER ↓	UTMOS ↑
GT	1.000	0.114	4.152
kNN-VC	0.857	0.115	3.941
RVC	0.748	0.120	4.190

Attack Type	Attack Name	Channel	DCT	AudioSeal	Timbre	WMCodec	VoiceMark
No attack	-	×	0.148	0.000	0.000	0.000	0.015
No attack	-	✓	0.405	0.032	0.045	0.034	0.216
Baseline Vocoder	HiFiGAN	×	0.499	0.482	0.006	0.527	0.017
Baseline Vocoder	HNSincNSF	×	0.497	0.490	0.194	0.521	0.030
Baseline Vocoder	HiFiGAN	✓	0.499	0.495	0.186	0.500	0.333
Baseline Vocoder	HNSincNSF	✓	0.499	0.497	0.344	0.506	0.323
Proposed Self VC	kNN-VC	×	0.496	0.498	0.502	0.539	0.496
Proposed Self VC	RVC	×	0.490	0.499	0.490	0.501	0.498
Proposed Self VC	kNN-VC	✓	0.500	0.496	0.498	0.525	0.504
Proposed Self VC	RVC	✓	0.500	0.493	0.498	0.507	0.494

Attacker performance values in bitwise error rate. Values near 0.5 mean near-random guessing (effective watermark failure).