



中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

《 Traceability and Copyright Protection in Neural Speech Process 》

Institute of Automation, Chinese Academic of Science
Junzuo Zhou

Audio Watermarking: Concept



Embed encrypted information into speech;
Detect and decode via dedicated models

Copyright protection

Traceability of synthesized speech sources

Regulatory registration and supervision of synthesized speech content

.....

Key property of audio watermarking:

Imperceptible to human auditory perception

Image watermarking: primarily emphasizes *robustness*,
enabling reliable detection and verification

Image steganography: primarily emphasizes *concealment*
and *information confidentiality*

Trade-off Properties

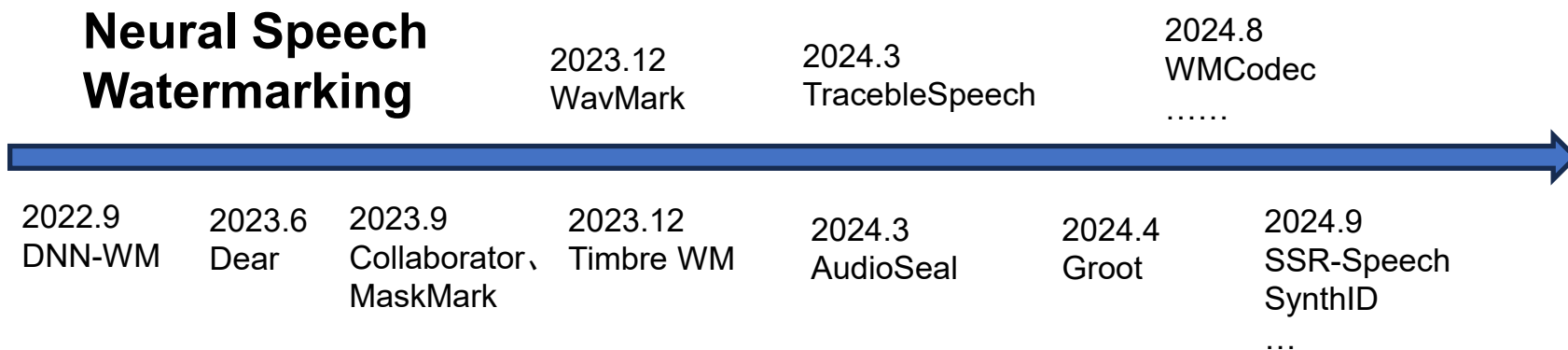
- Imperceptibility: Signal-to-Noise Ratio (SNR) and Speech Quality Metrics such as PESQ
- Capacity: Average Embeddable Bits per Second (BPS)
- Robustness: BER between decoded and original bits, average bitwise accuracy, AUC, or $\text{TPR@FPR} = 0.01$

Once imperceptibility is ensured, the trade-off between robustness and embedding capacity is determined by the target application requirements

Audio Watermarking: Development



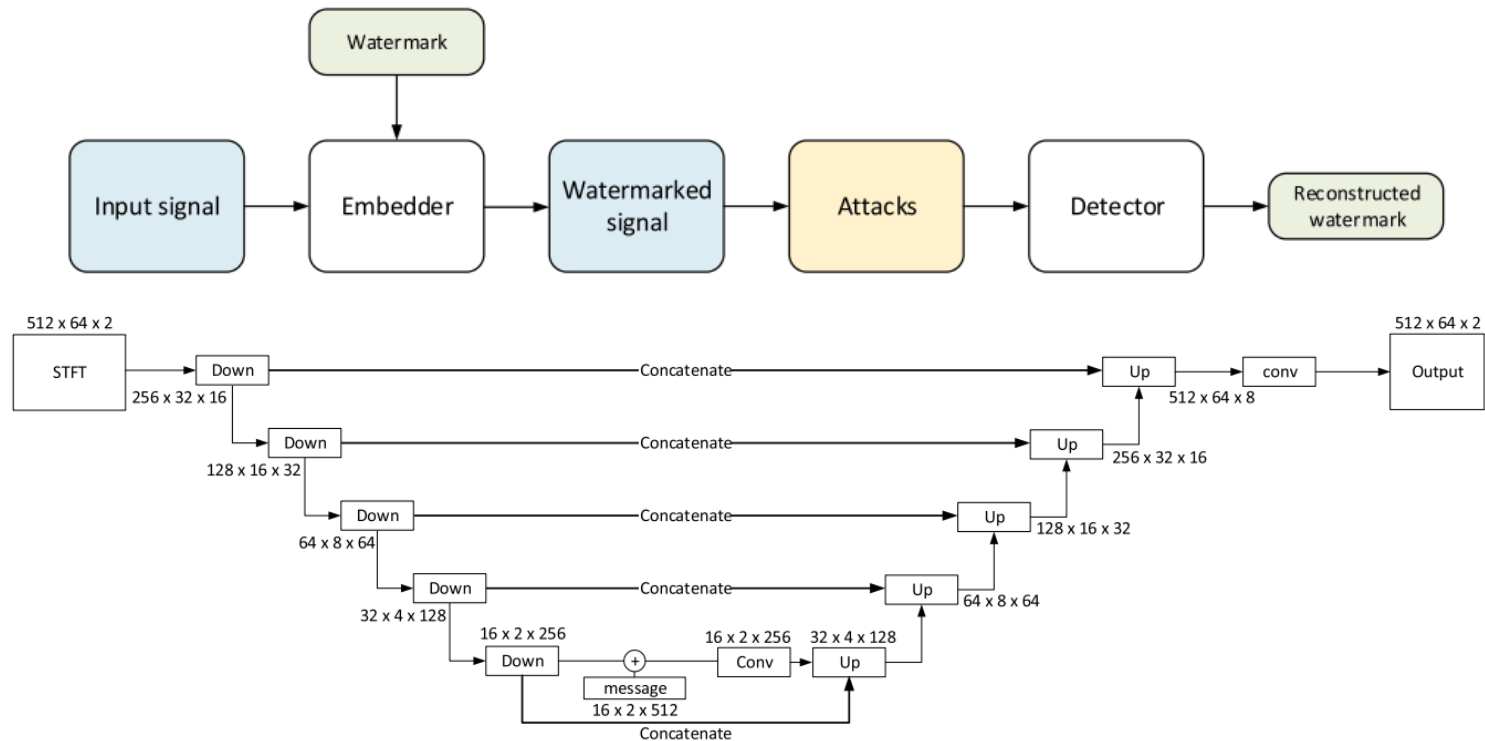
Conventional speech watermarking methods are based on expert knowledge and heuristic design, exhibiting limited generalization and robustness.



- 1: General Post-hoc Audio Watermarking Methods
- 2: Task-driven / Task-integrated Audio Watermarking
- 3: Audio Watermarking for Open-source Models

General Post-hoc Watermarking

DNN-WM

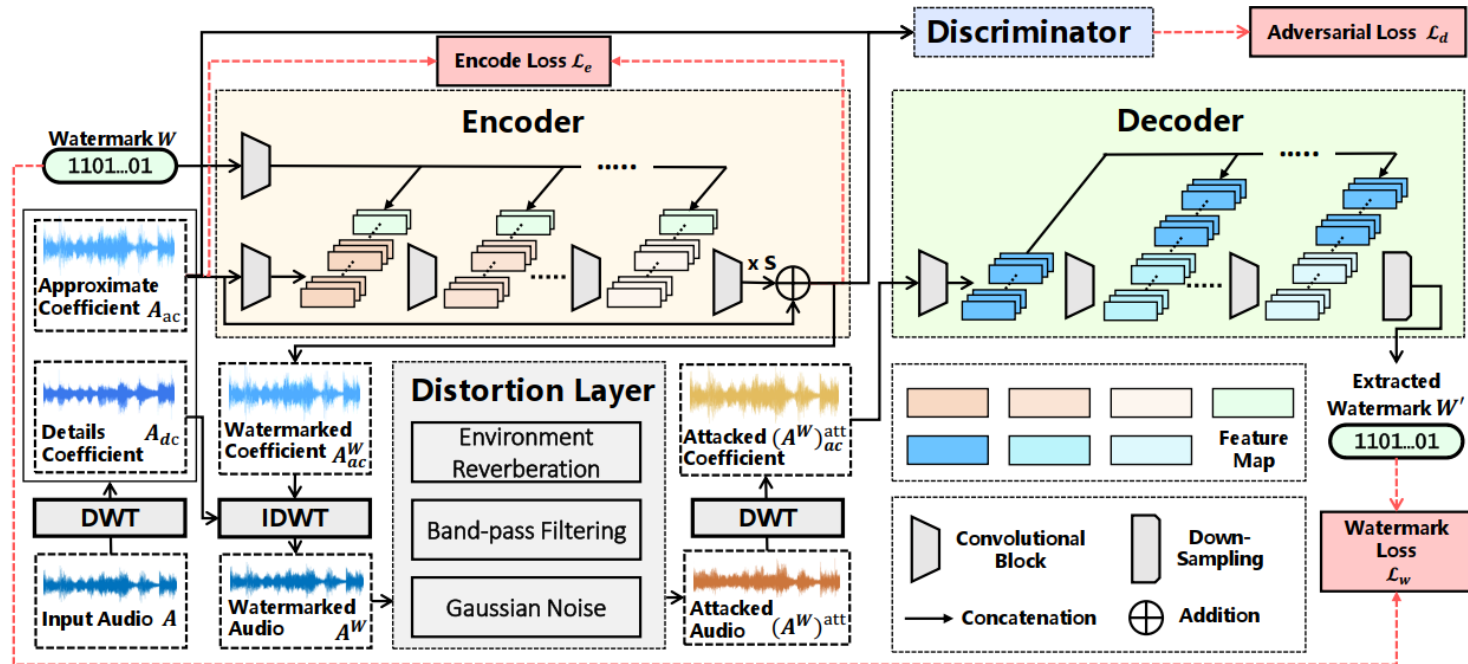


- 1: Embedding performed in the STFT frequency domain
- 2: Robust against three types of attacks (dropout, random noise, high-pass filtering)
- 3: Low embedding capacity (2.5 bit / 2 s):

Pavlović K, Kovačević S, Djurović I, et al. Robust speech watermarking by a jointly trained embedder and detector using a DNN[J]. Digital Signal Processing, 2022, 122: 103381.

General Post-hoc Watermarking

DeAR

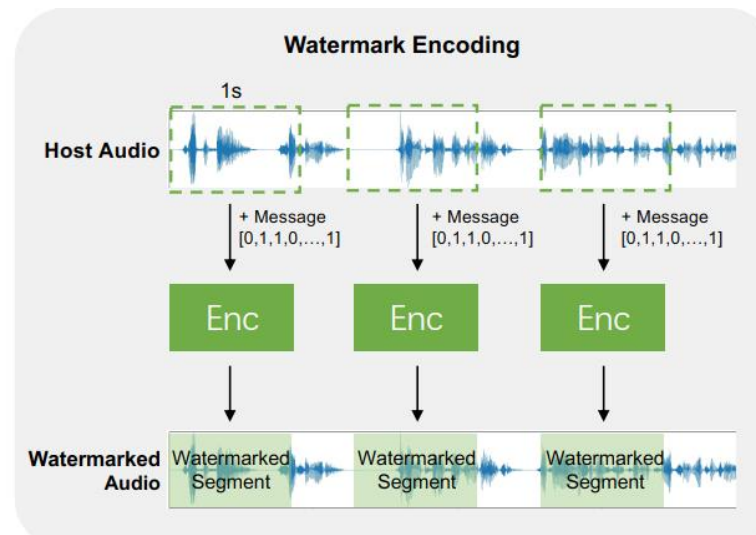
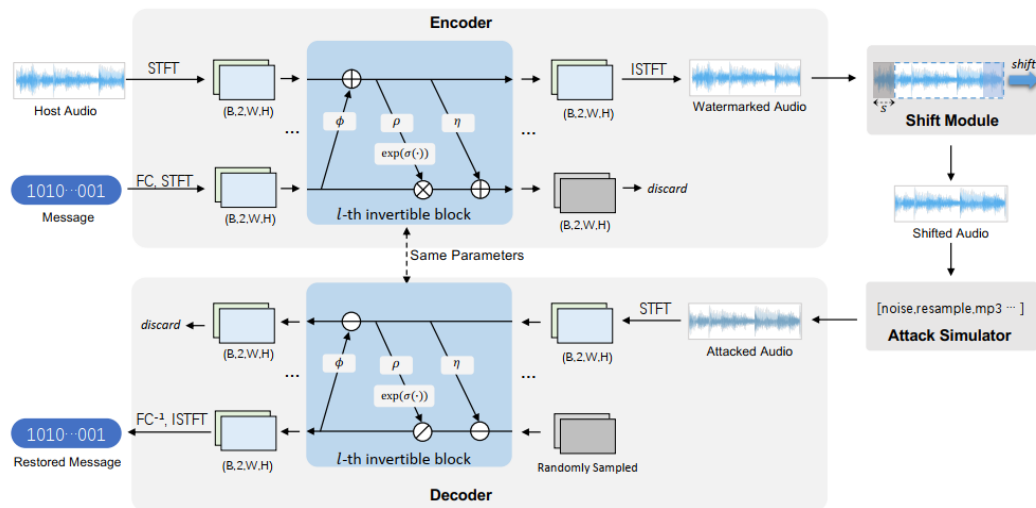


- 1: Embedding performed in the DWT frequency domain
- 2: Watermark integrated via an encoder with residual design to adjust the watermark–speech ratio
- 3: Considers audio transcription environments as simulated attacks
- 4: Further improved embedding capacity (100 bit / 11 s)

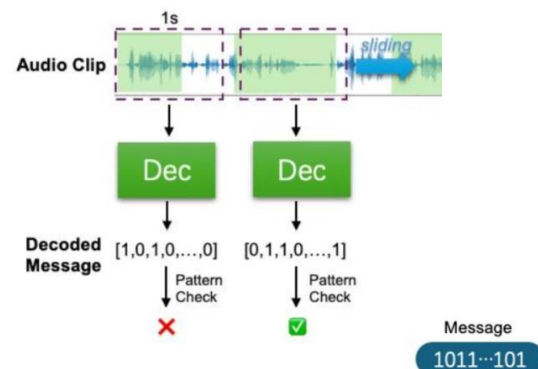
Liu C, Zhang J, Fang H, et al. Dear: A deep-learning-based audio re-recording resilient watermarking[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(11): 13201-13209.

General Post-hoc Watermarking

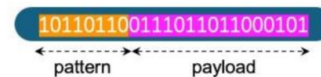
WavMark



- 1: Encoding and decoding with a reversible network design:
 - $y = f(x), x = f^{-1}(x)$
- 2: Evaluated under nine simulated attacks:
- 3: Further improved embedding capacity (32 bit / 1 s)
- 4: Watermark segment localization in long speech
 - detection window with force matching
 - pattern(16bit) + payload(16bit)



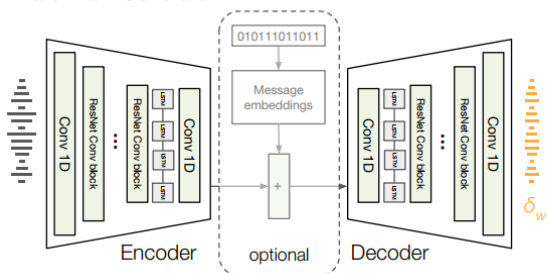
Chen G, Wu Y, Liu S, et al. Wavmark: Watermarking for audio generation[J]. arXiv preprint arXiv:2308.12770, 2023.



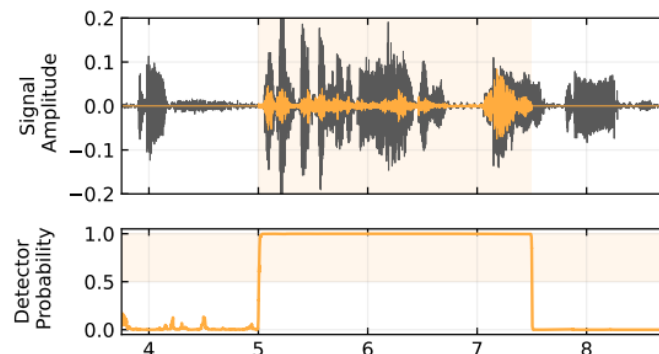
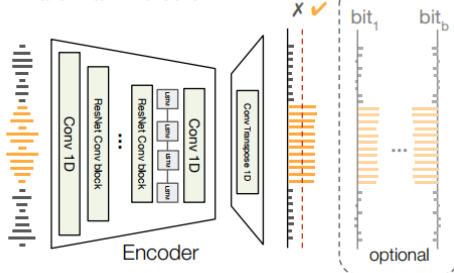
General Post-hoc Watermarking

AudioSeal

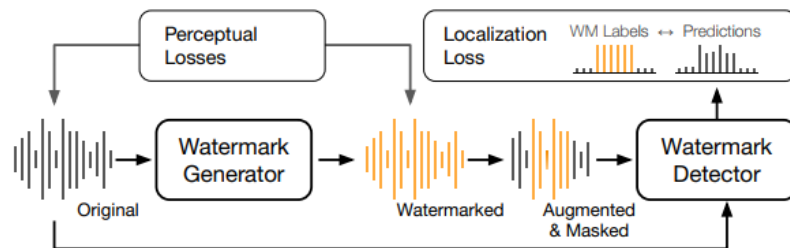
Watermark Generator



Watermark Detector



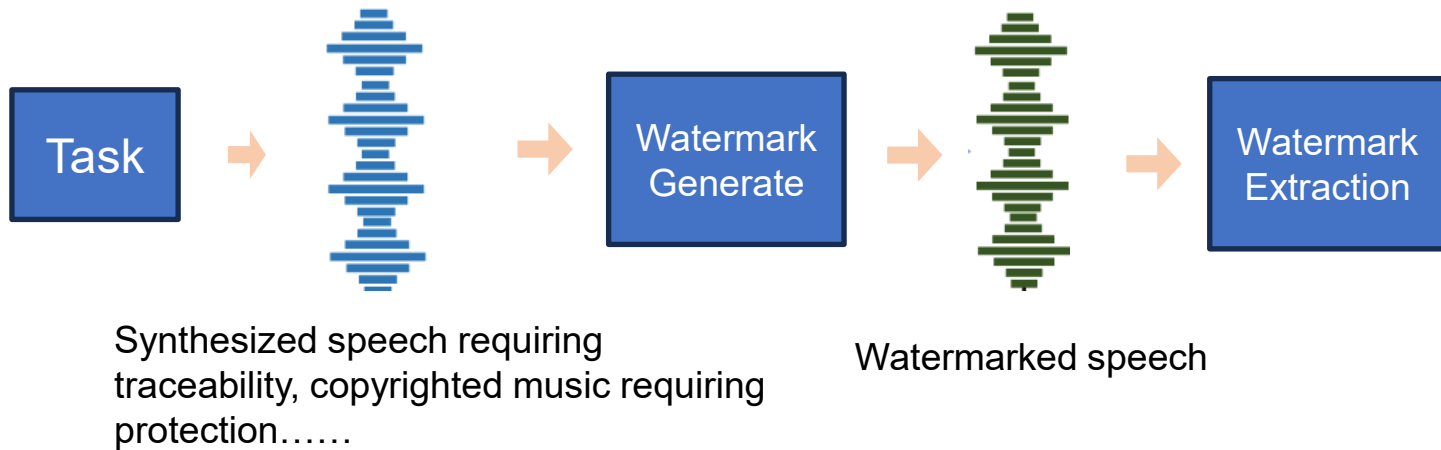
- 1: Watermark embedding without involving the spectrum
- 2: Frame-level localization of watermark segments
 - precision up to 1/16k second
- 3: Unified architecture for watermark detection and payload extraction
- 4: Maintains embedding capacity (16 bits / 1 s) and robustness



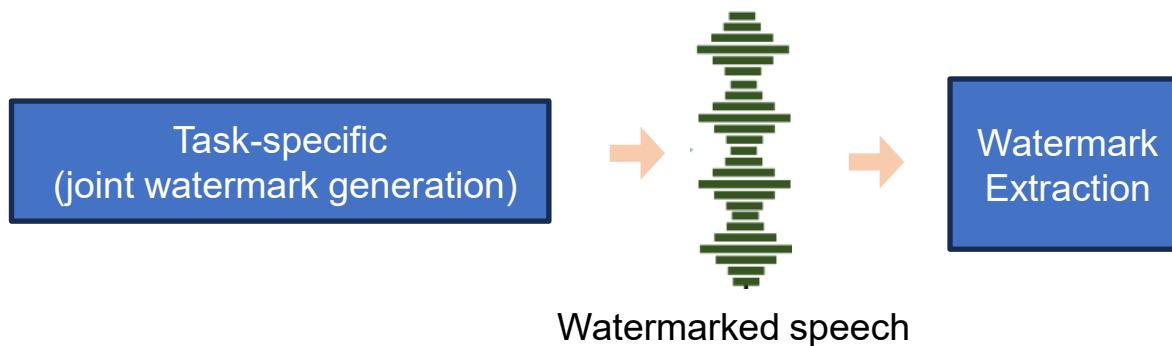
San Roman R, Fernandez P, Elshahar H, et al. Proactive Detection of Voice Cloning with Localized Watermarking[C]//ICML 2024-41st International Conference on Machine Learning. 2024, 235: 1-17.

Task-driven Watermarking

General watermarking methods are post-hoc, multi-stage, and cascaded, rather than end-to-end systems.



Why we need task-driven watermarking



Task-driven Watermarking



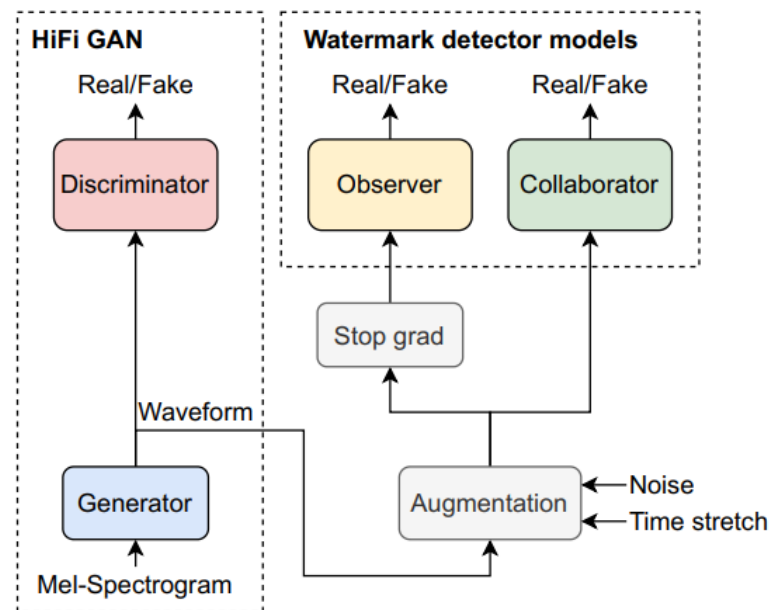
Collaborator Watermarking

1: Enhance detectability of real/fake labels during speech synthesis

integrate watermark indicators into vocoder training

2: Use a speech forgery detection model as the watermark detector

The indicator only reflects authenticity (real/fake) without recovering watermark content



Task-driven Watermarking

🧠 **Watermark embedded during generation**, improving **imperceptibility**

- Stage 1: Watermark embedding + codec integration
- Stage 2: VALLE in speech synthesis

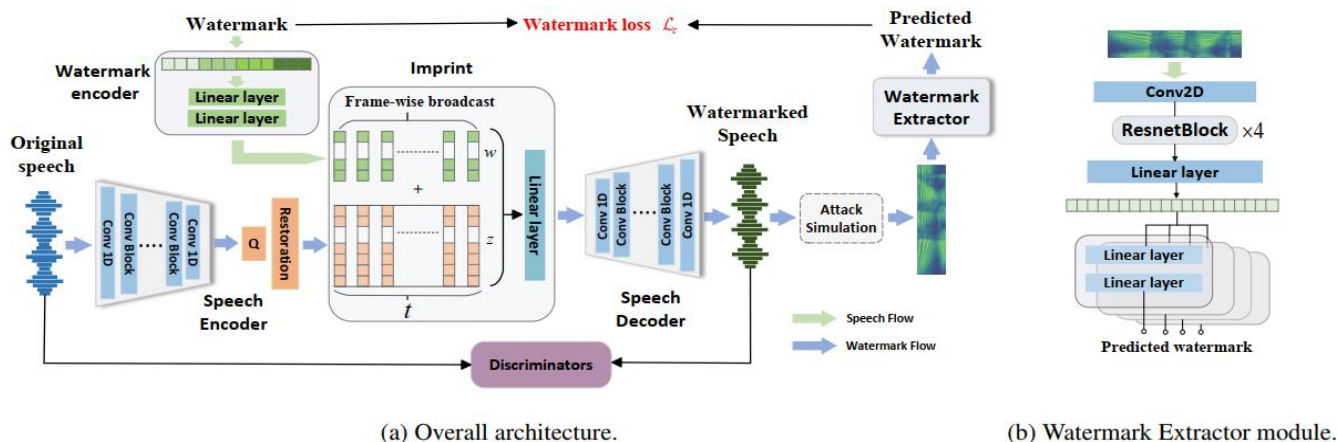


Figure 1: The first stage: Watermarking mechanism integrate into neural codec.

🧠 **Codec jointly training**

- Imprint watermarks into pre-decoder features

🧠 **TTS-specific Imprint Strategy**

- Frame-wise broadcast watermark features
- Full-span protection for temporal **robustness**
- Supports **flexibility** to variable-length inference

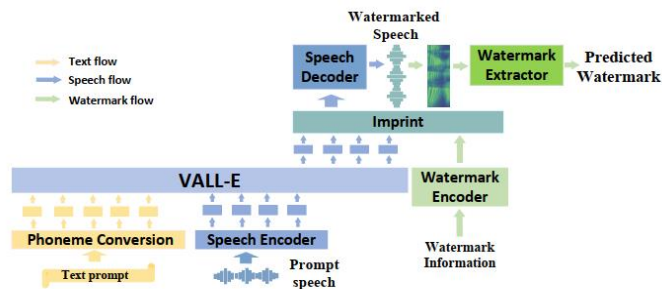


Figure 2: The second stage: Watermarking mechanism integrate into language model of VALL-E.

Task-driven Watermarking



Imperceptibility

Table 1: *Watermark Imperceptibility Metrics in Speech Reconstruction*

Model	PESQ \uparrow	STOI \uparrow	ViSQOL \uparrow
HiFicodec + WavMark(16bit)	3.197	0.947	3.880
TraceableSpeech(4@10)	3.641	0.950	4.060
TraceableSpeech(4@16)	3.569	0.948	3.985

¹ @ denotes the watermarking capacity. For example, 4@16 indicates 4-digit base-16, equivalent to the 16-bit capacity of WavMark used in the baseline. This annotation is applicable to other tables as well.

Table 2: *Speech Quality in Zero-Shot Speech Synthesis*

Model	WER(%) \downarrow	MOS \uparrow
VALL-E + WavMark(16bit)	10.80	3.554 \pm 0.19
TraceableSpeech(4@10)	9.61	3.959 \pm 0.18
TraceableSpeech(4@16)	10.47	3.905 \pm 0.17



• In both experiments, imperceptibility improved.



Robustness

- Resist temporal-editing attacks (resplicing).
- Robust even when 2/3 of the speech is randomly removed.

Table 3: *Watermark extraction accuracy (%) under various attacks*

Attack \ Model	Resplicing	Normal	RSP-90	Noise-W35	SD-01	AR-90	EA-0315	LP5000
VALL-E + WavMark(16bit)	No	100.00	99.76	91.41	100.00	100.00	94.53	100.00
TraceableSpeech(4@10)	No	100.00	100.00	100.00	100.00	100.00	100.00	100.00
TraceableSpeech(4@16)	No	98.97	98.82	98.95	99.12	99.46	97.71	98.84
VALL-E + WavMark(16bit)	Once	91.10	91.46	63.53	95.95	93.61	88.58	89.66
TraceableSpeech(4@10)	Once	100.00	100.00	100.00	99.90	100.00	100.00	100.00
TraceableSpeech(4@16)	Once	100.00	99.82	99.83	98.78	99.50	99.57	99.62
VALL-E + WavMark(16bit)	Twice	76.65	77.74	49.14	79.47	85.46	68.19	75.32
TraceableSpeech(4@10)	Twice	100.00	100.00	100.00	100.00	100.00	100.00	100.00
TraceableSpeech(4@16)	Twice	99.58	99.20	99.58	99.56	99.00	99.65	98.83

¹ The resplicing column mean the times of resplicing attack


Flexibility

- With a 4-bit 64-base watermark in 0.3s speech, TraceableSpeech extracts with 95%+ accuracy.

Table 4: *Watermark extraction accuracy (%) of larger capacity models under various speech durations (s)*

Model \ Duration	1.0	0.8	0.5	0.3	0.2	0.175	0.15	0.125	0.1
TraceableSpeech(4@32)	100.00	100.00	99.74	99.23	94.13	86.22	77.29	57.14	50.51
TraceableSpeech(4@64)	100.00	100.00	99.86	95.57	80.59	66.79	53.90	27.47	17.01

Task-driven Watermarking

 **Goal Scenario:** Authenticity verification in speech codec transmission (Watermark is embedded before compression (encoder side) and still successfully extracted after decoding)

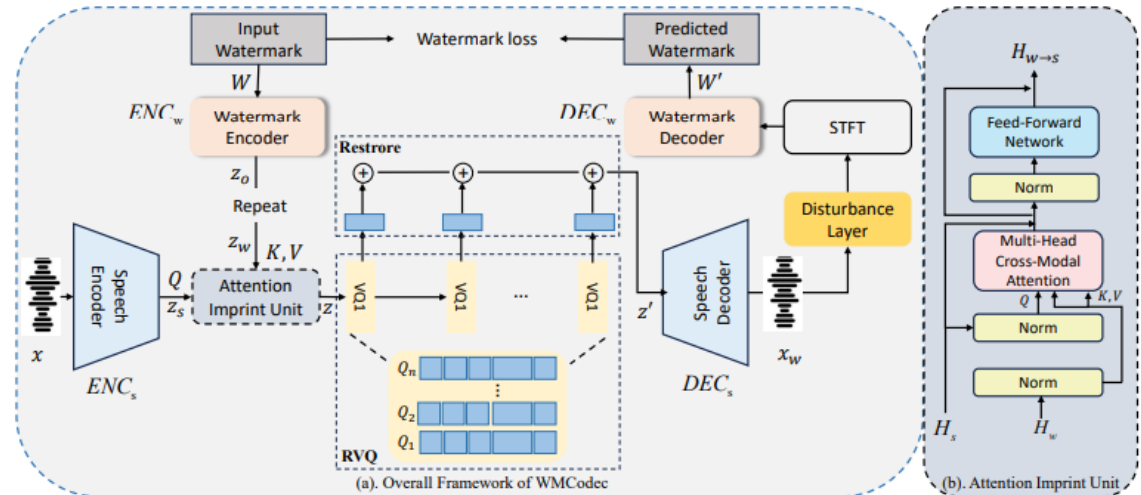
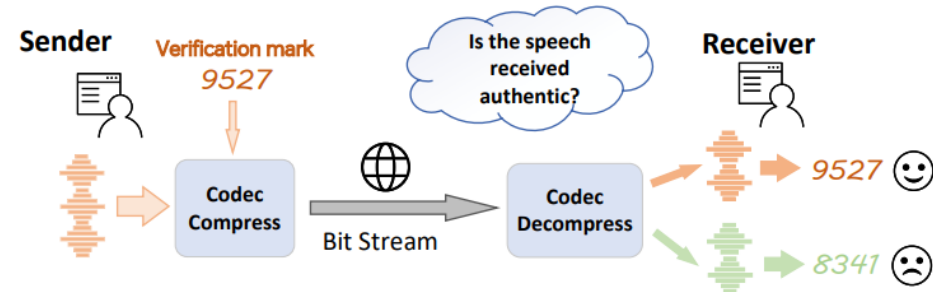
 **Challenges of prior SOTA :**

- Quantization distortion in codecs makes watermark extraction **significantly harder**
- **Simple fusion** between watermark and speech features limit accuracy

 **Core Design:**

- End-to-end training with pre-compression embedding + post-decoding extraction
- Iteratively fuses watermark info with speech through cross-modal attention (Attention Imprint Unit)

Fig. 1. Example of Watermark as Verification Marking for Codec Protection

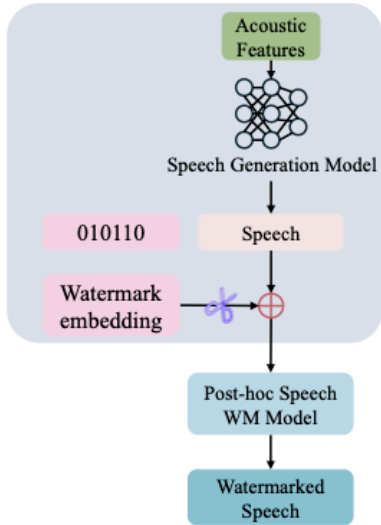


Zhou J, Yi J, Ren Y, et al. WMCodec: End-to-End Neural Speech Codec with Deep Watermarking for Authenticity Verification[C]//ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE

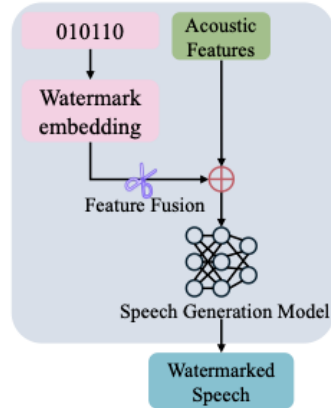
- Imperceptibility: Basic Demand.
- Capacity: Further Demand.
- Robustness: Generalization and Practicality out of the Dataset

Audio Watermarking Classification: A Different Perspective

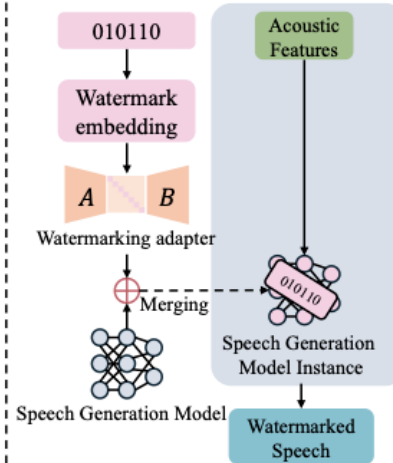
 Can be removed or manipulated  Open Source Released



(a) Audio-level Watermarking



(b) Feature-level Watermarking



(c) Parameter-level Watermarking

Applied as a Plugin in Speech
Synthesis Models **Parameter**

- Vocoder
- Codec Decoder

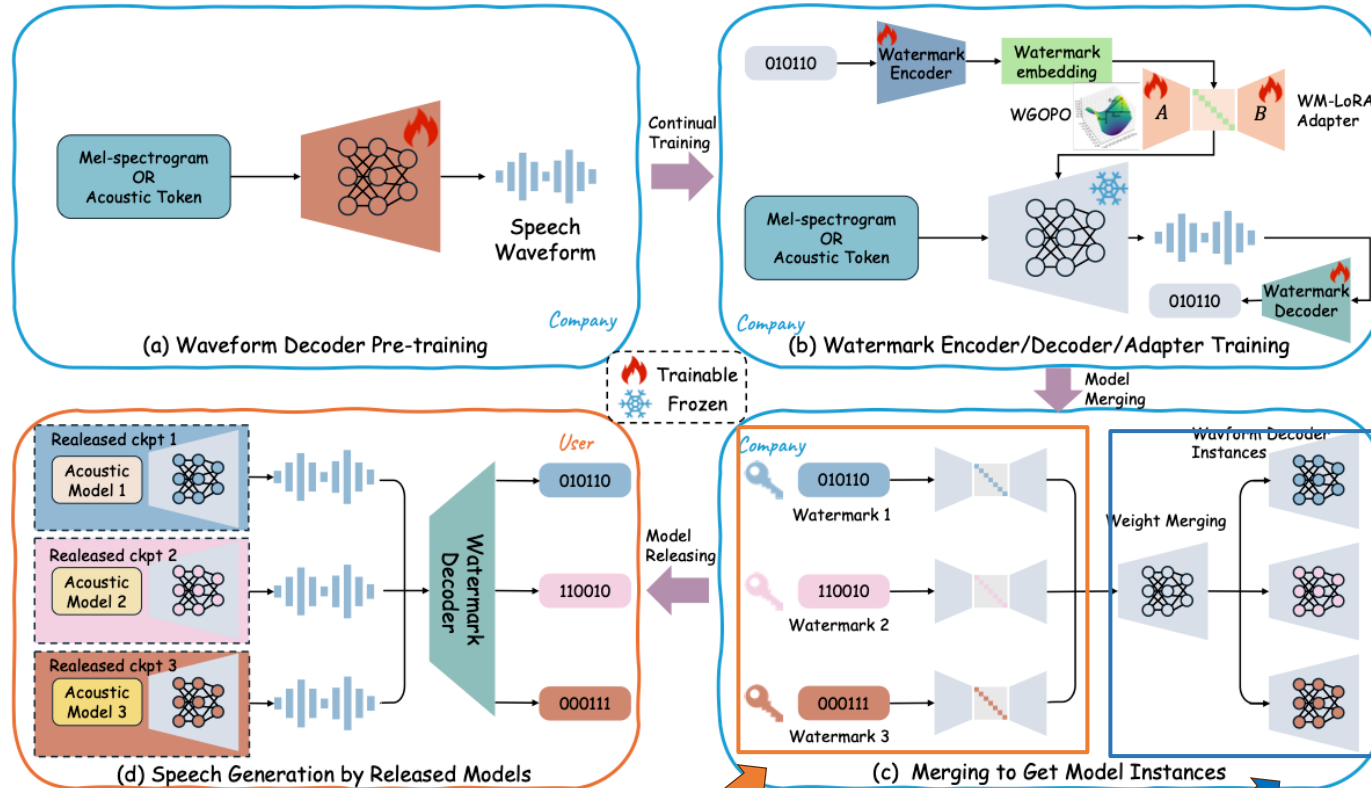
- Audio-level watermarking
- Feature-level Watermarking
- **Parameter-level Watermarking**

Open-Source White-Box Protection

Flexibility (modifiable before release)

Security (difficult to remove with after release)

Parameter-level Watermarking



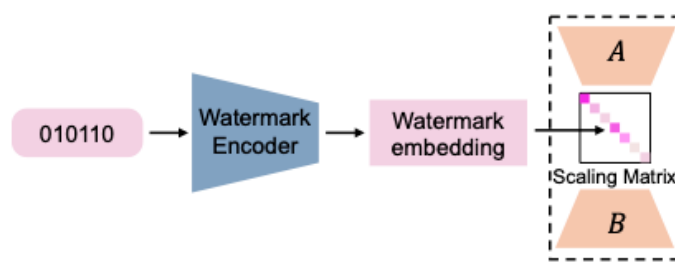
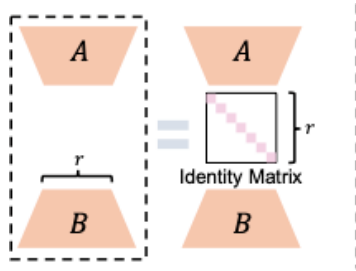
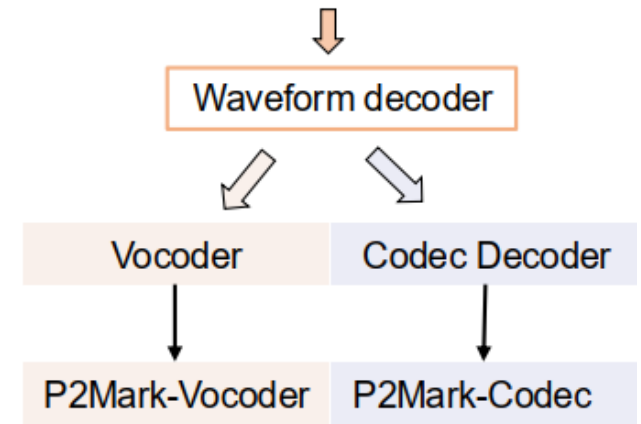
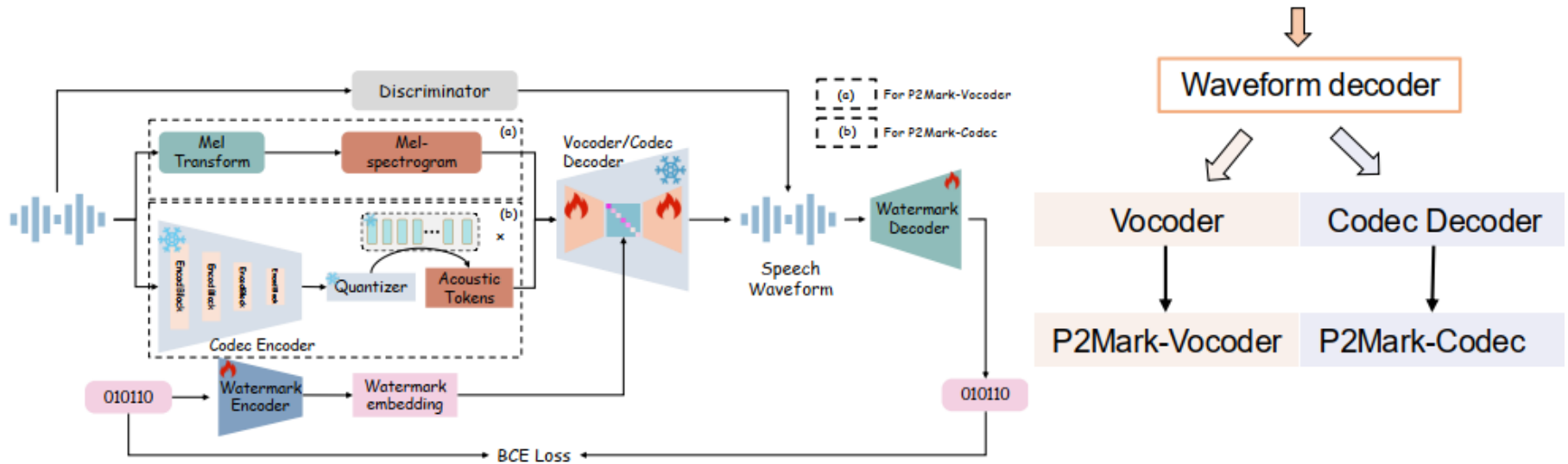
Flexibility

Safety

P2Mark: Plug-and-play Parameter-level
Watermarking for Neural Speech Generation

Parameter-level Watermarking

P2Mark: Plug-and-play Parameter-intrinsic Watermarking for Neural Speech Generation



- Add watermark embeddings as a diagonal matrix to LoRA;
- Replace decoder 1-D convolutions with LoRA ones;
- Jointly train with watermark encoder/decoder;

Thank you !