Detecting Voice Cloning Attacks via Timbre Watermarking

Chang Liu[1] Jie Zhang[2] Tianwei Zhang[2] Xi Yang[1] Weiming Zhang[1] Nenghai Yu[1]

[1]University of Science and Technology of China
[2]Nanyang Technological University

hichangliu@mail.ustc.edu.cn jie_zhang@ntu.edu.sg zhangwm@ustc.edu.cn

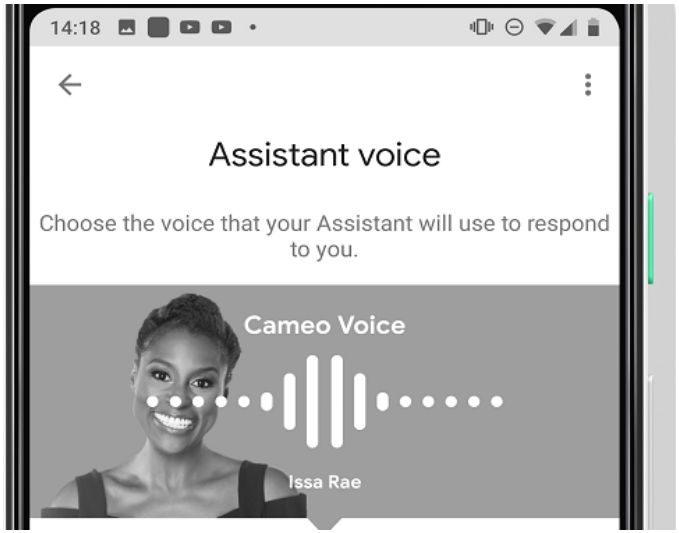
中国科学技术大学
University of Science and Technology of China

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

NDSS SYMPOSIUM 2024

Voice Cloning

Interview with Ai Musk and AI Jobs on YouTube: Is Ai a threat? Ai musk talks to AI Jobs, debating AI's threat to humanity.



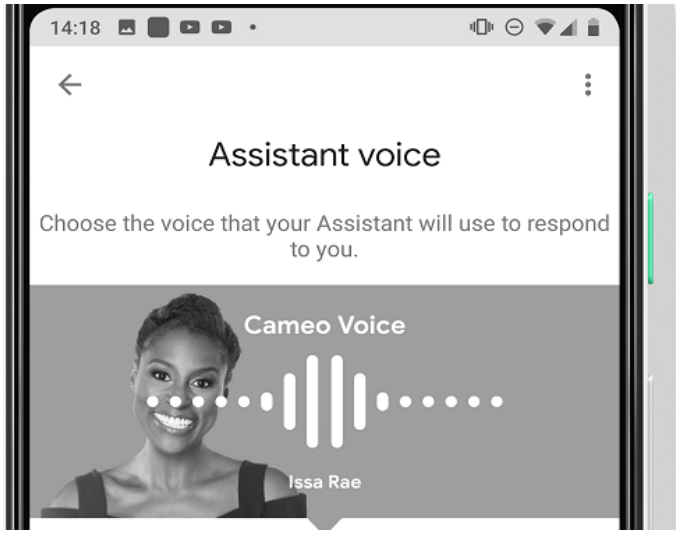
Google and other companies utilize high-quality, customized voice synthesis technology to offer voice assistant services.

From April 2023, a trend swept through major video platforms with once-dormant music icons making a collective comeback, releasing new songs at an astonishing pace—achieving in one month what previously took years.

High-quality customized voice cloning technology has been widely used in entertainment, commercial

Voice Cloning

Interview with Ai Musk and AI Jobs on YouTube: Is Ai a threat? Ai musk talks to AI Jobs, debating AI's threat to humanity.

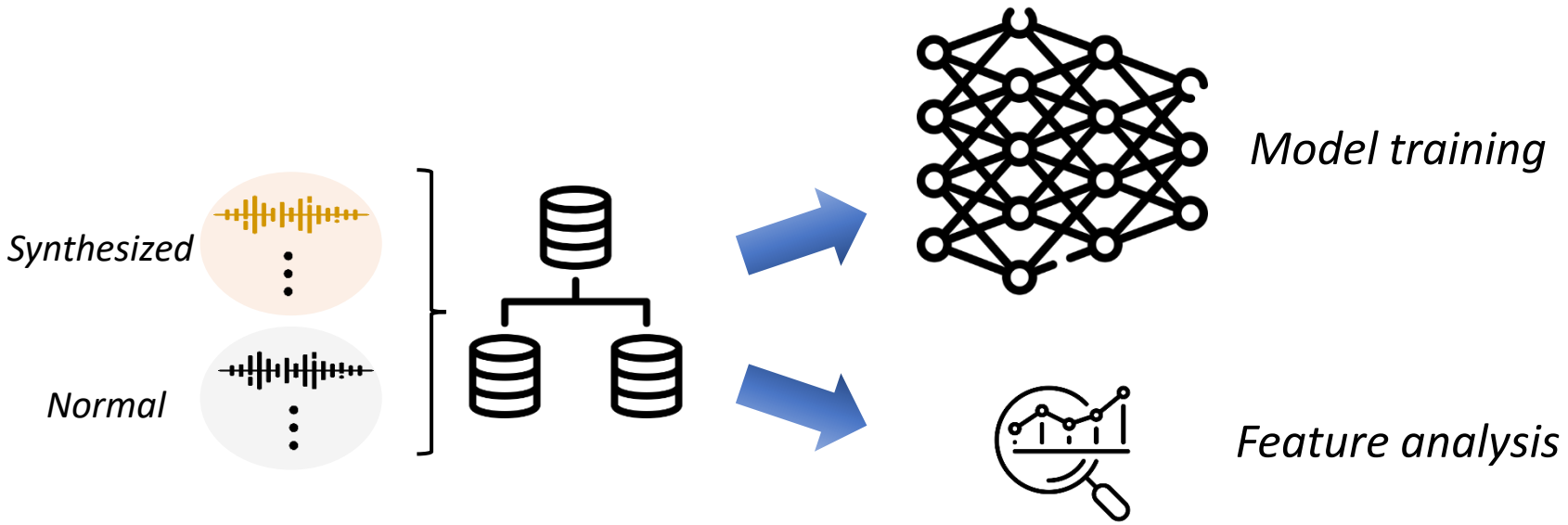


Google and other companies utilize high-quality, customized voice synthesis technology to offer voice assistant services.

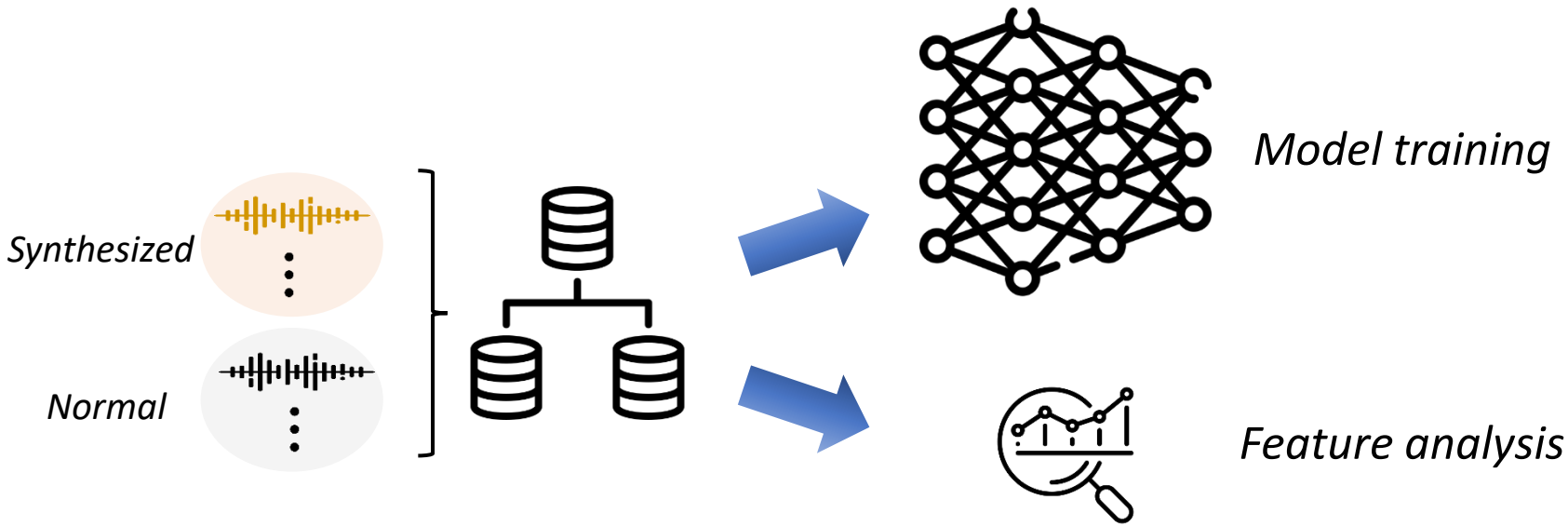
From April 2023, a trend swept through major video platforms with once-dormant music icons making a collective comeback, releasing new songs at an astonishing pace—achieving in one month what previously took years.

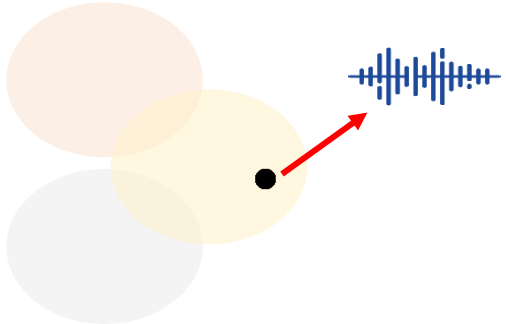
How to protect Timbre Rights?

Passive Detection-based Strategy



Passive Detection-based Strategy

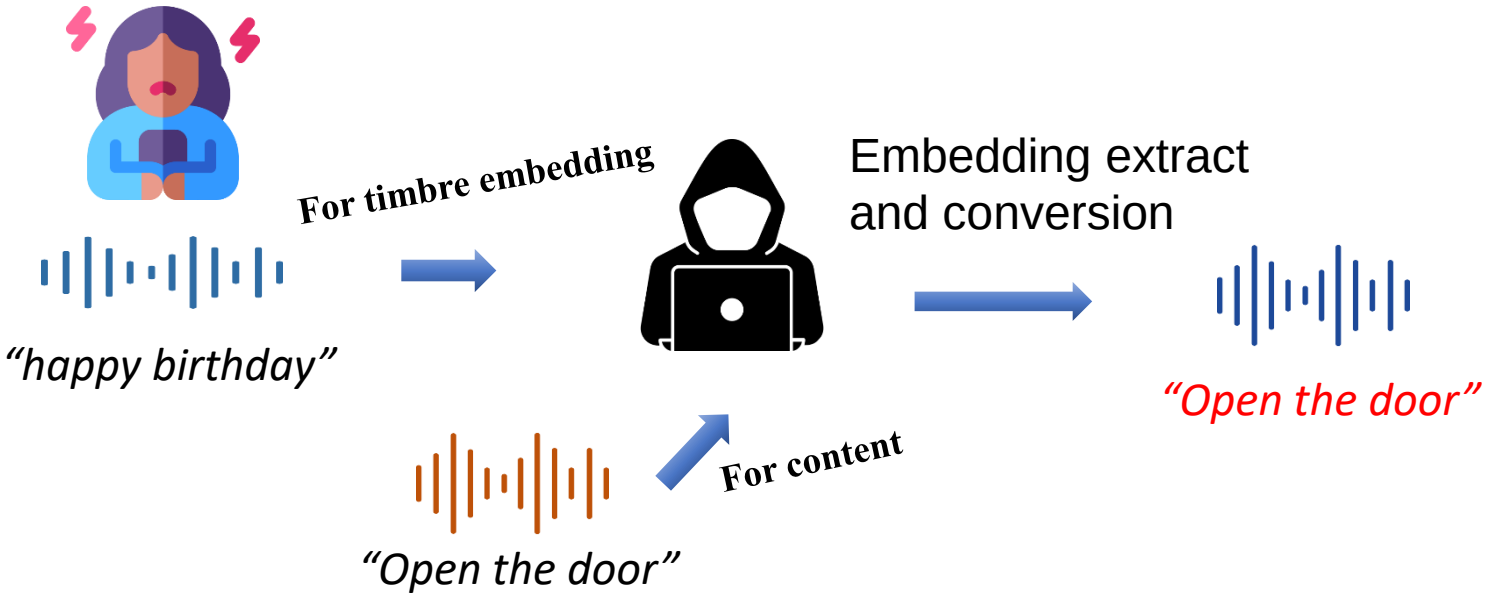




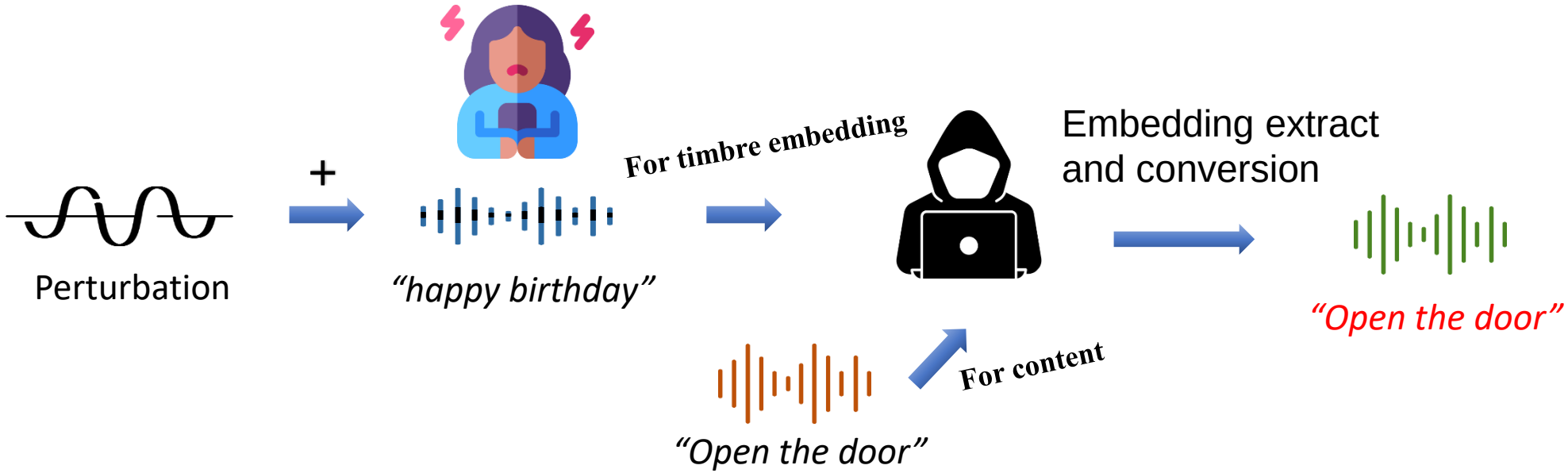
Synthesized?
Whose timbre*?*

- 😕 Generalizability and credibility is limited
- 😕 Can’t trace the original timbre

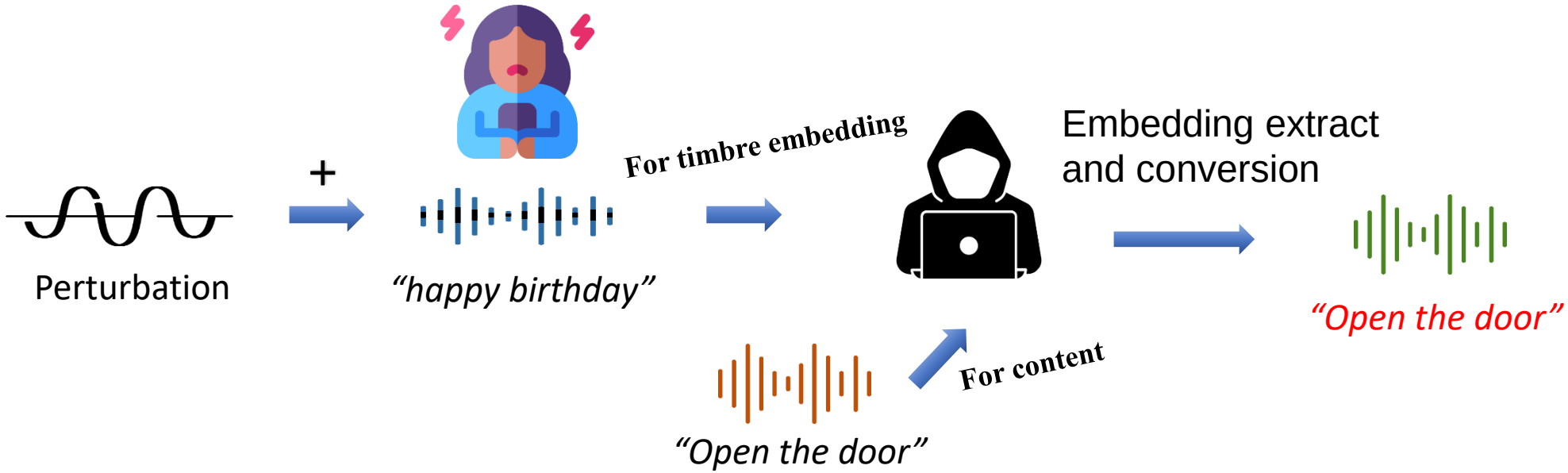
Proactive Prevention-based Strategy



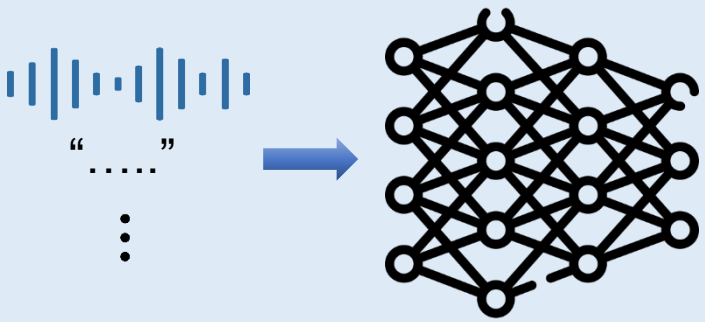
Proactive Prevention-based Strategy



Proactive Prevention-based Strategy

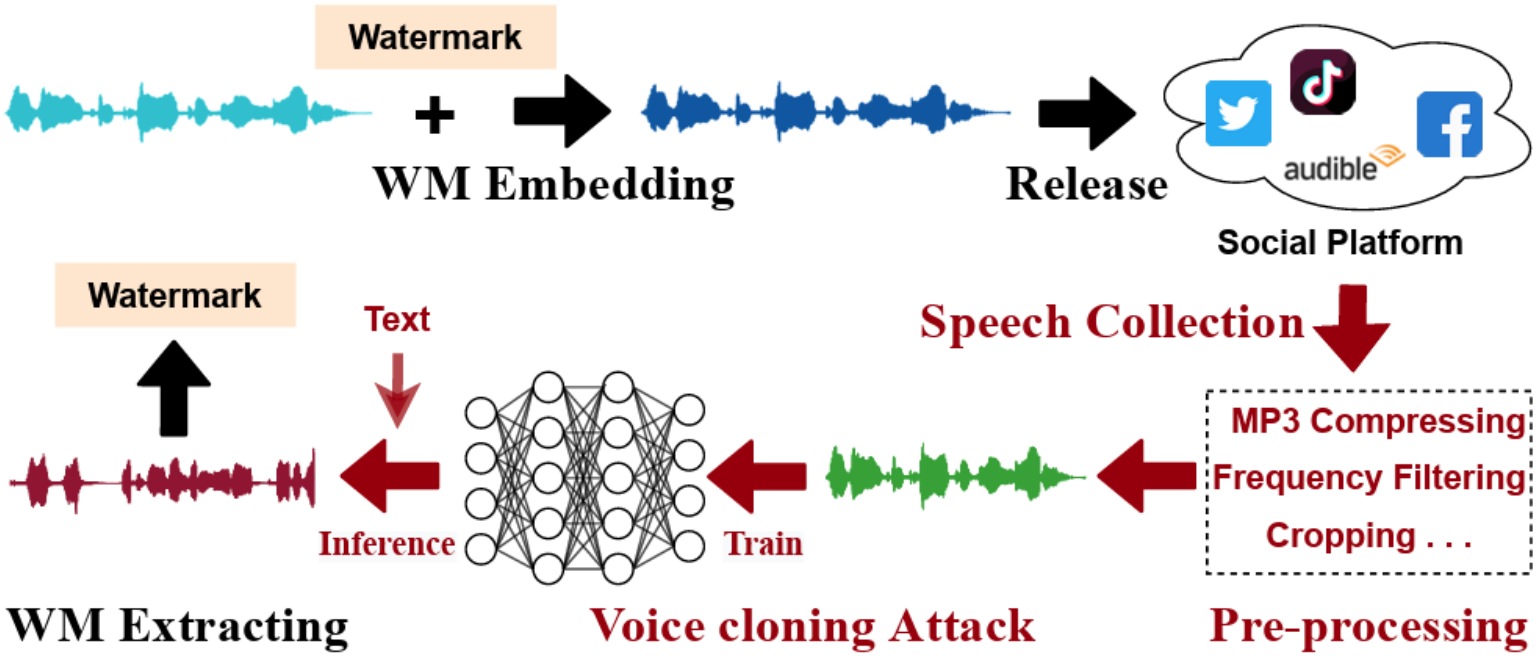


High quality TTS

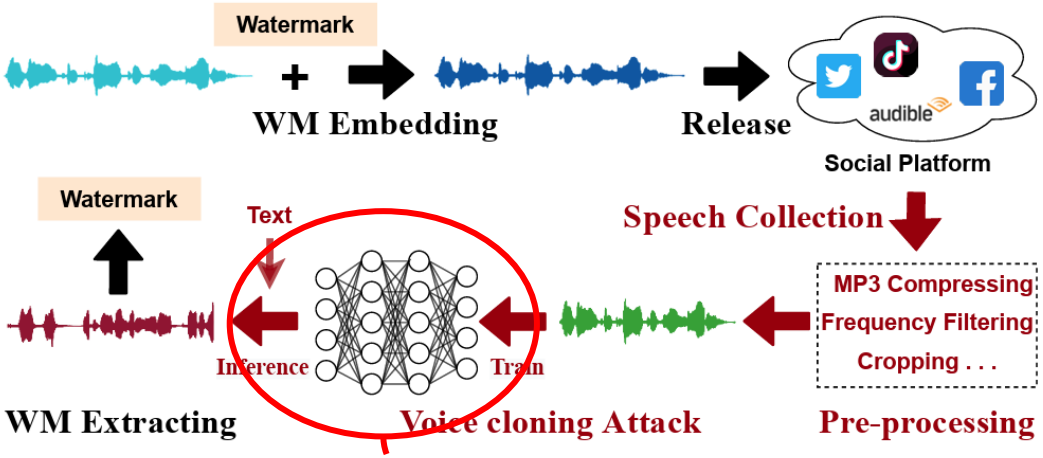


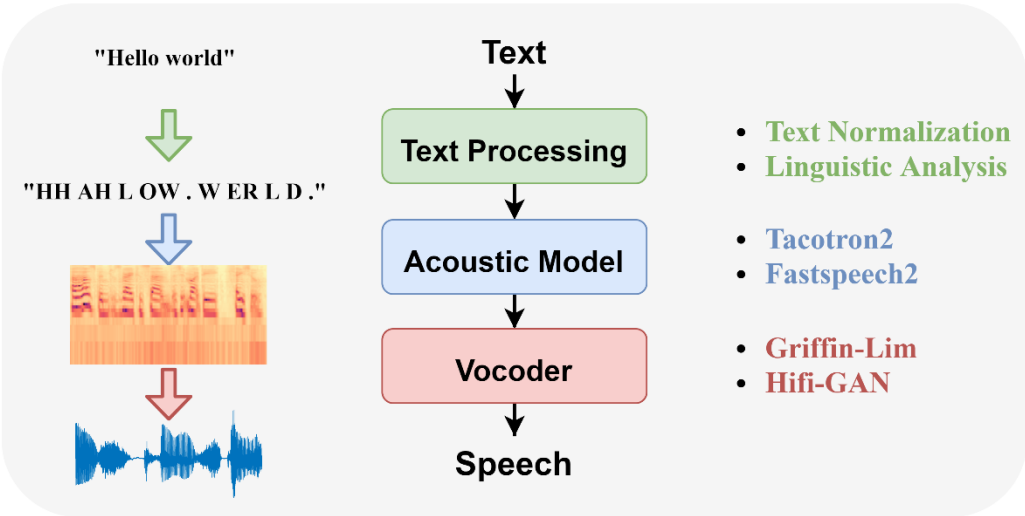
- A significant level of perturbation is required to achieve a effective defense
- It can only defend against clone models based on timbre decoupling and lacks defensive capability in high-quality TTS scenarios
- Can't trace the original timbre

Proposed Idea



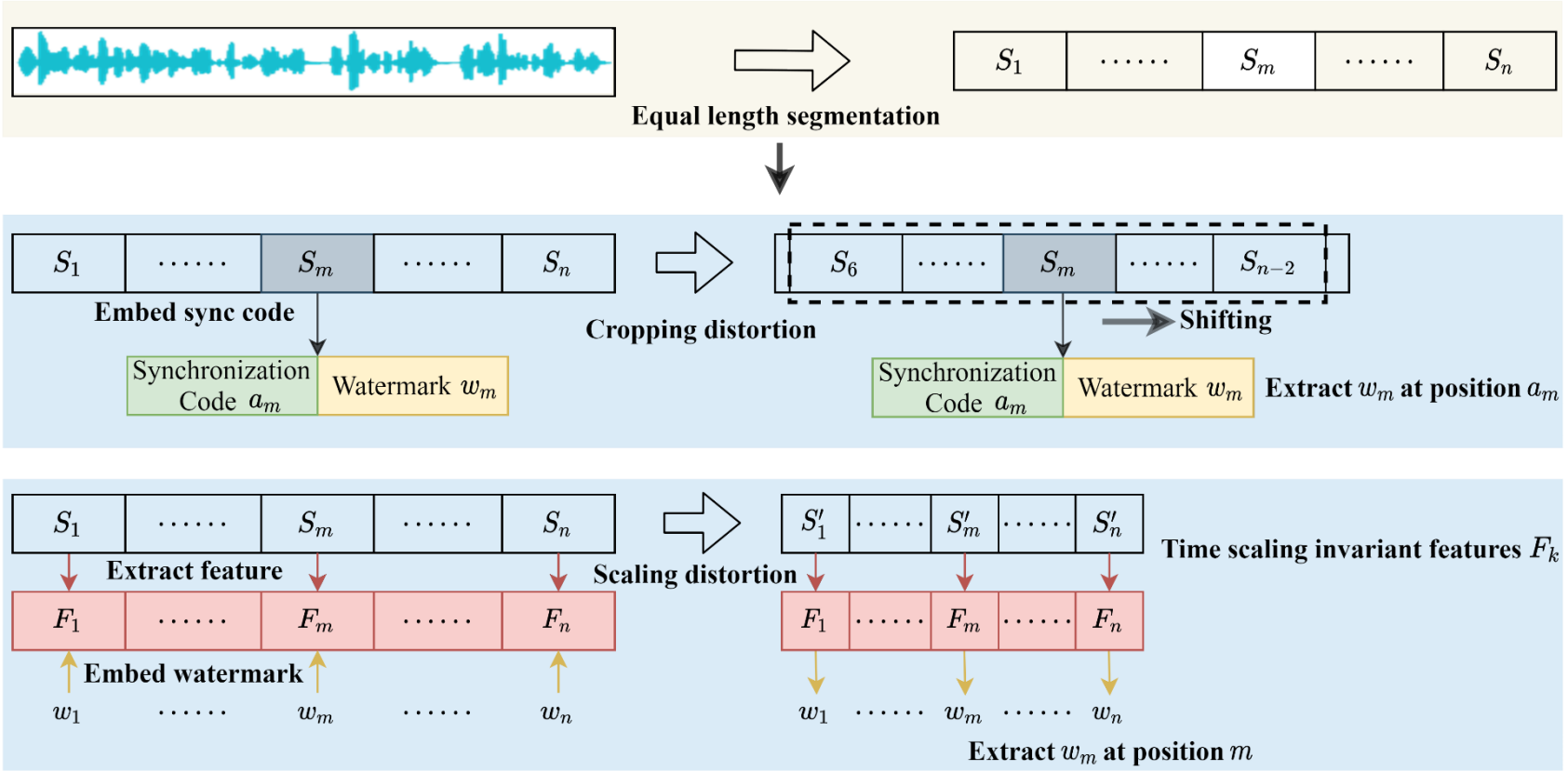
Voice Cloning





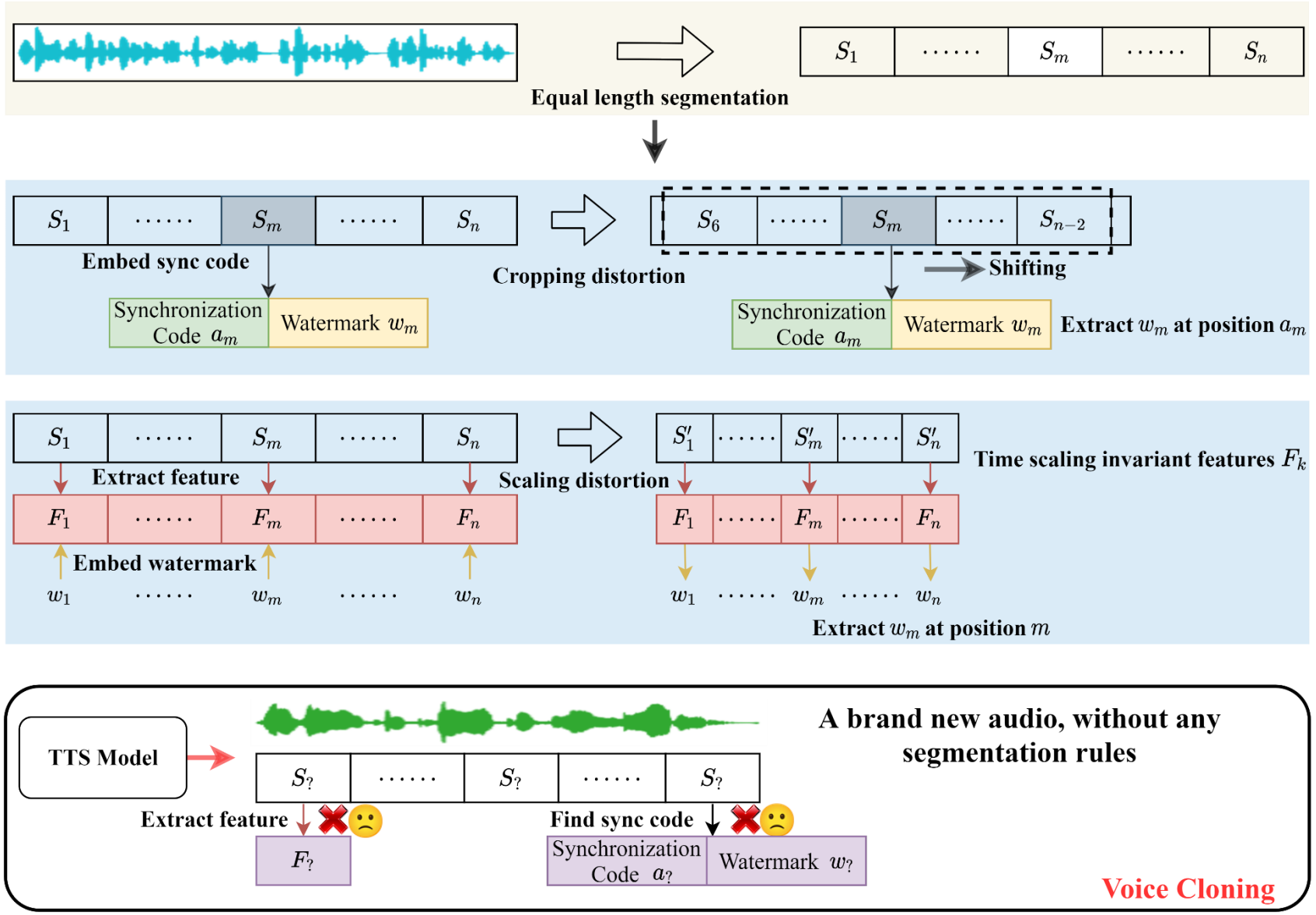
Traditional Audio Watermarking

Two types of solutions

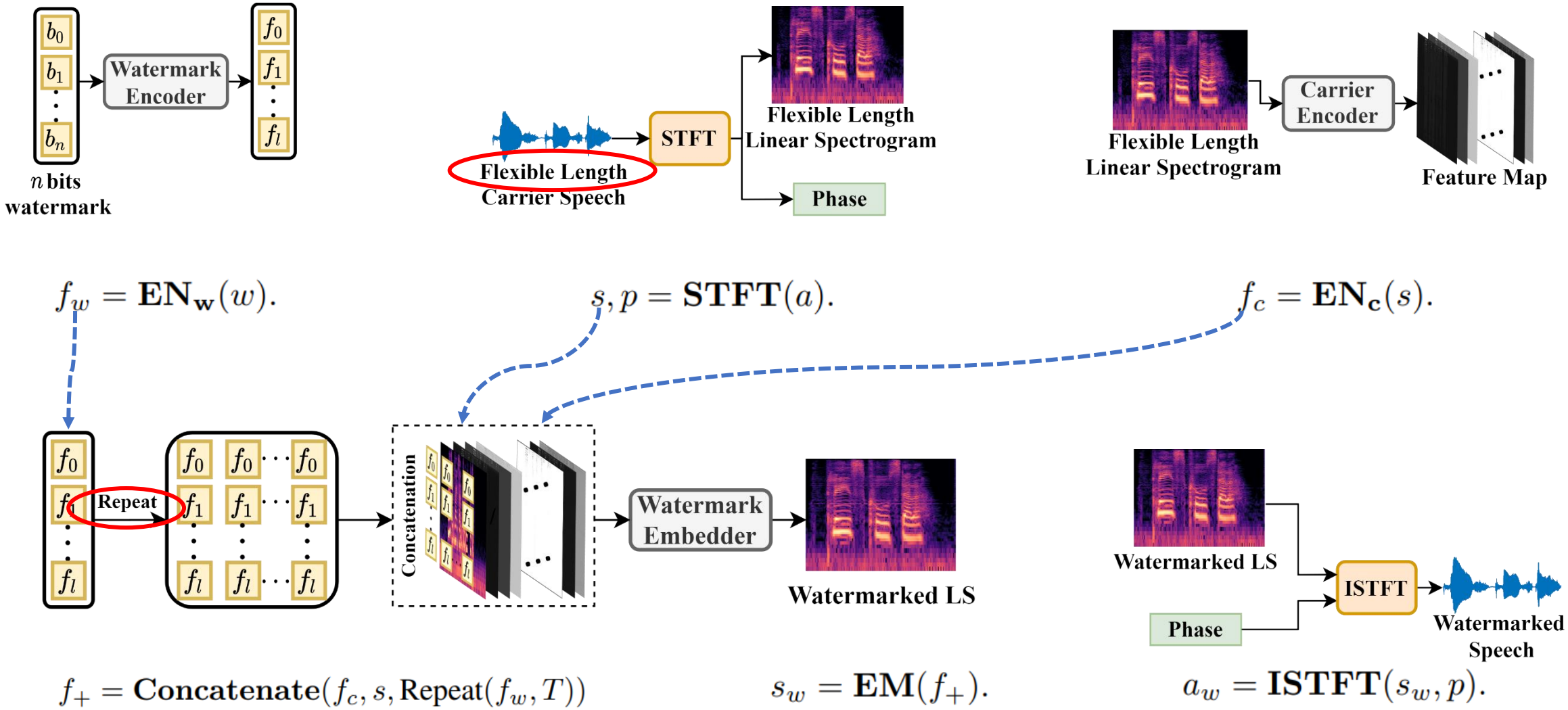


Traditional Audio Watermarking

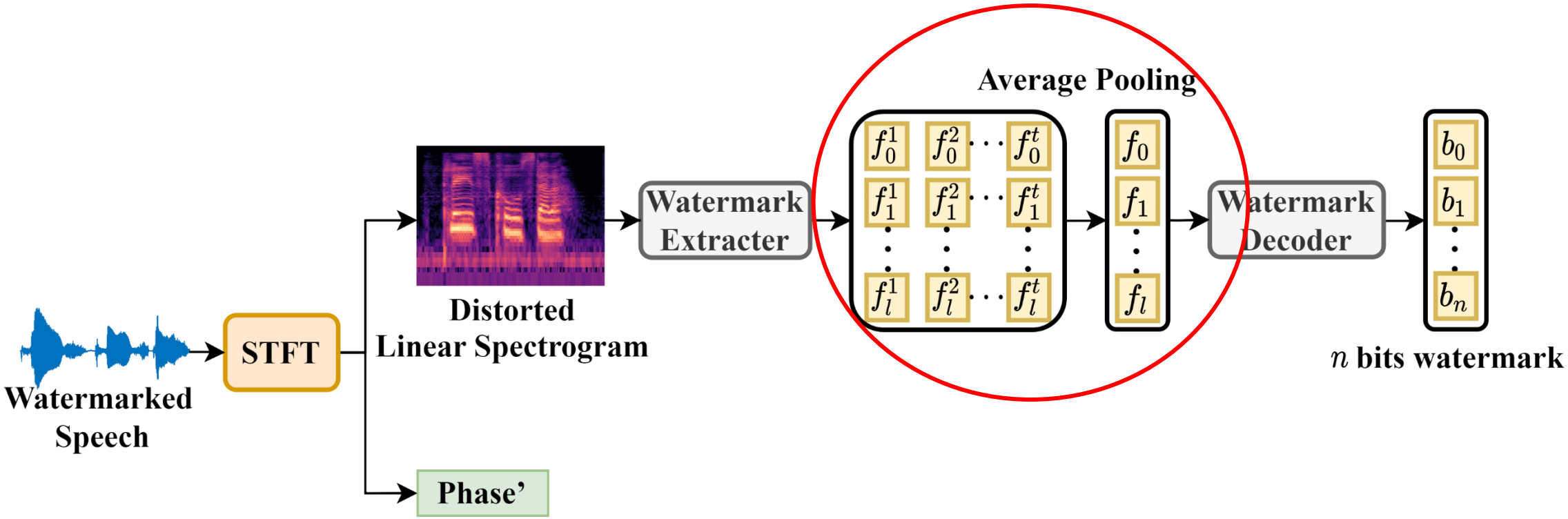
Two types of solutions



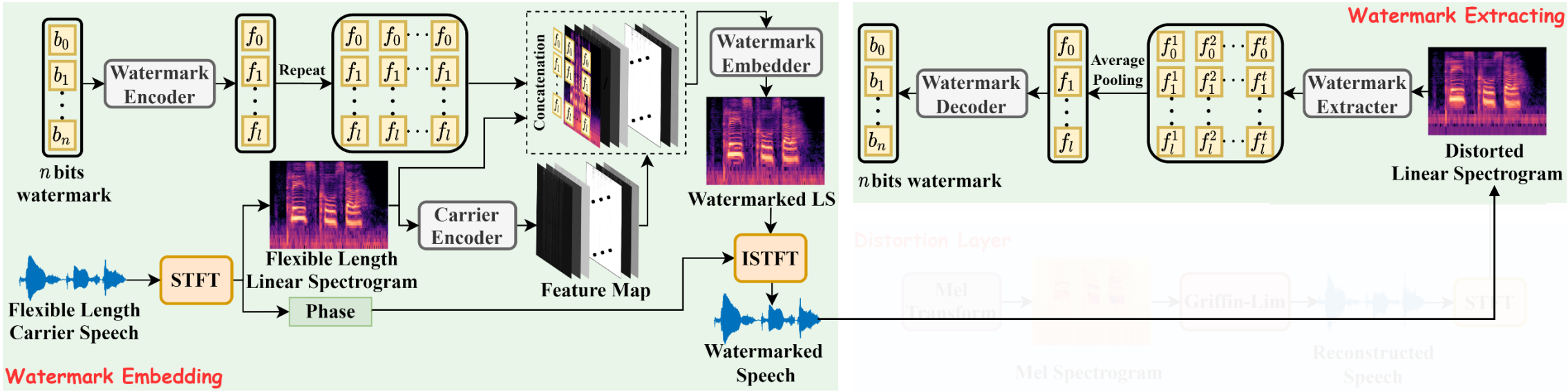
Goal 1: Time-dimension Independent



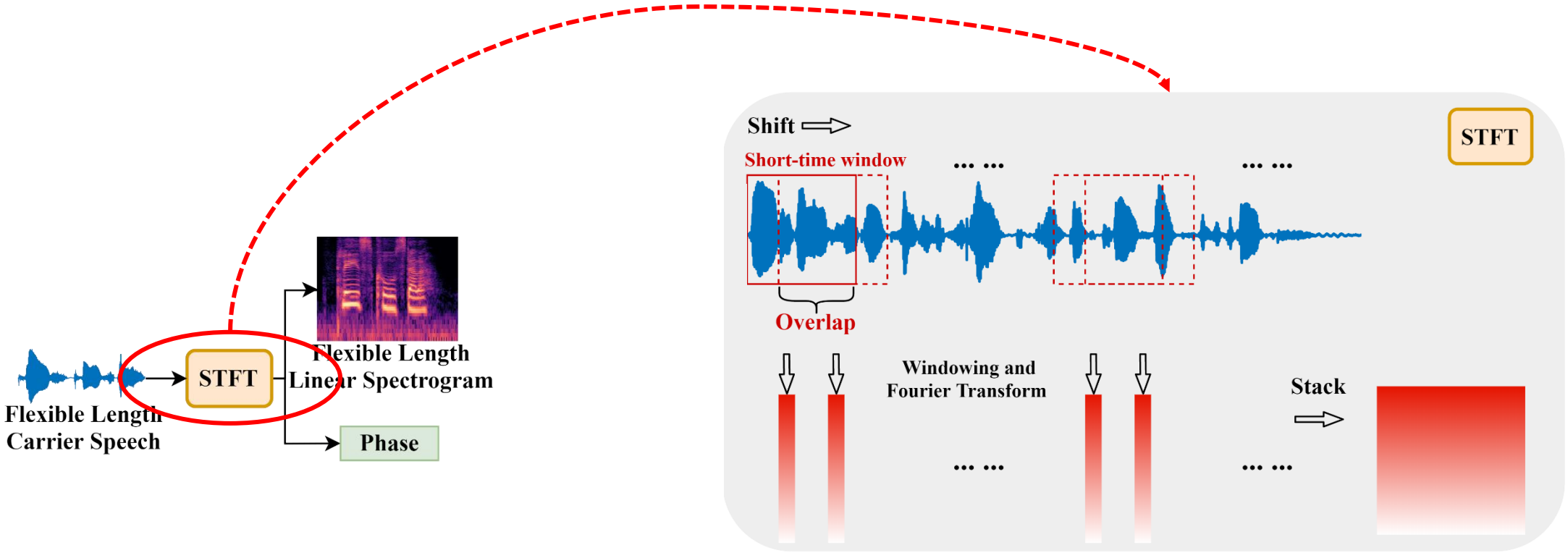
Goal 1: Time-dimension Independent



Goal 1: Time-dimension Independent



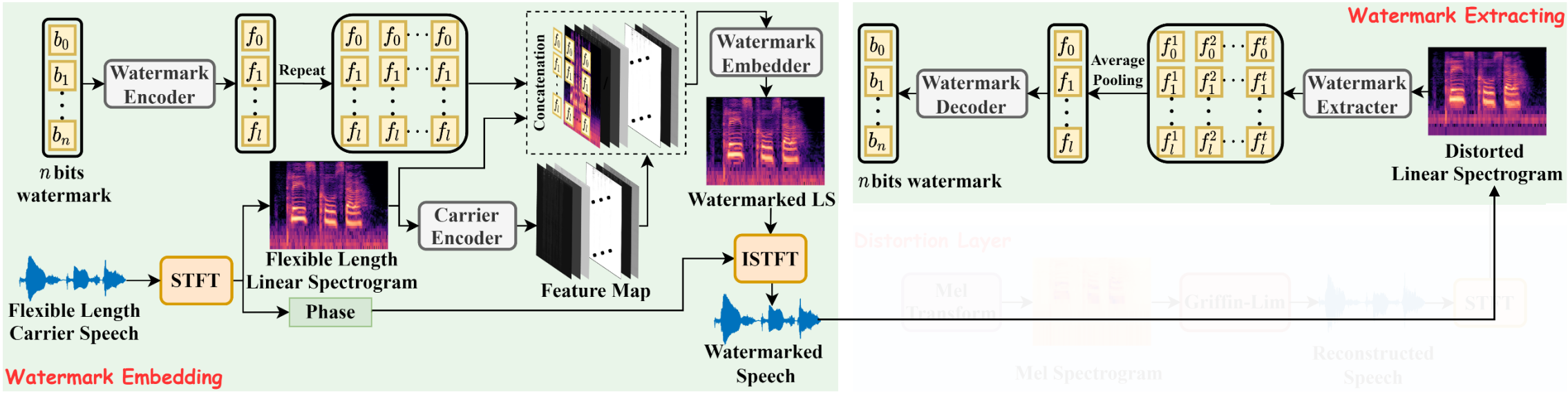
Goal 1: Time-dimension Independent



short-time effect

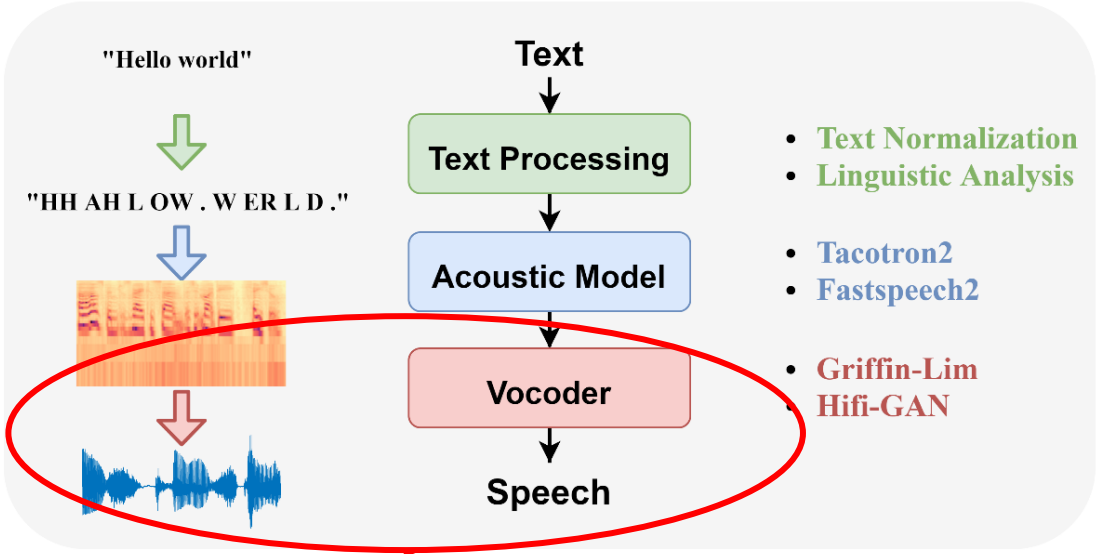
window overlapping effect

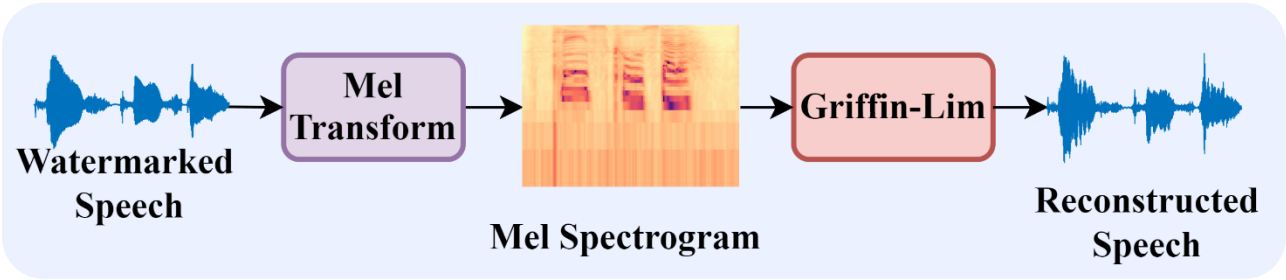
Goal 1: Time-dimension Independent



- Leveraging the short-time effect and the window overlapping effect of STFT
  - Embedding: Overlay the same watermark signal on the FFT coefficients at different moments in time.
  - Extraction: Take the average along the time axis, corresponding to the embedding strategy, to achieve time-axis-independent watermark embedding and extraction.

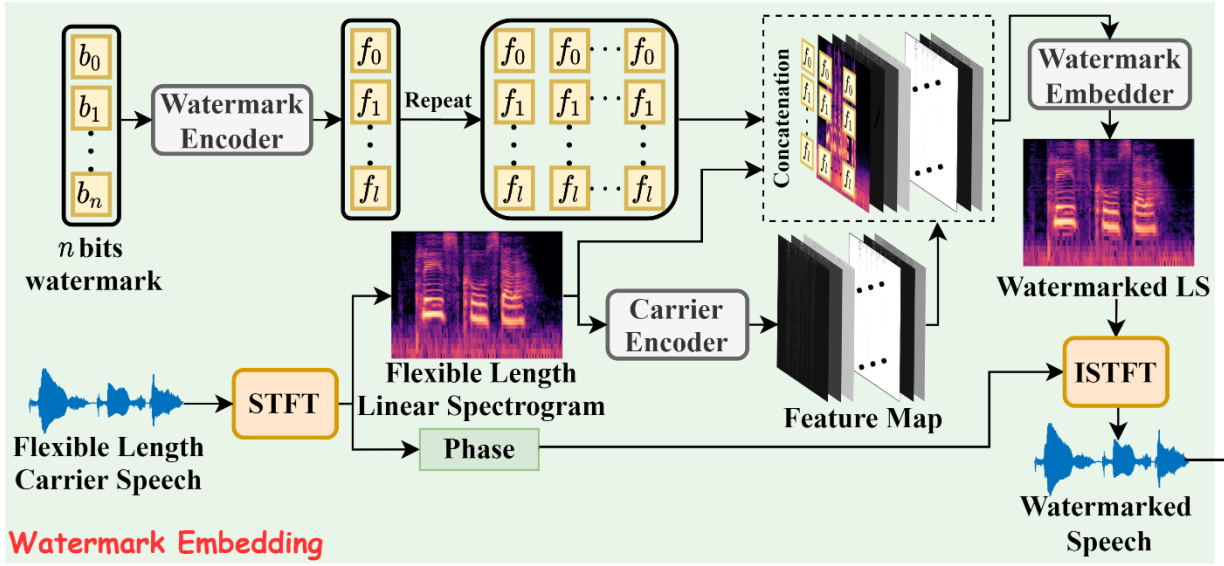
Goal 2: Voice Cloning Robustness

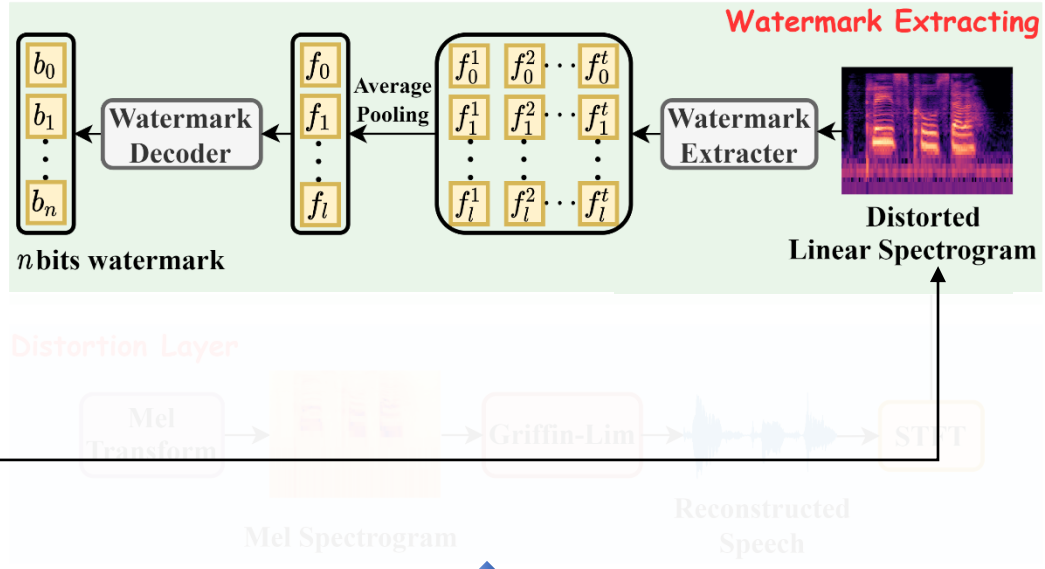




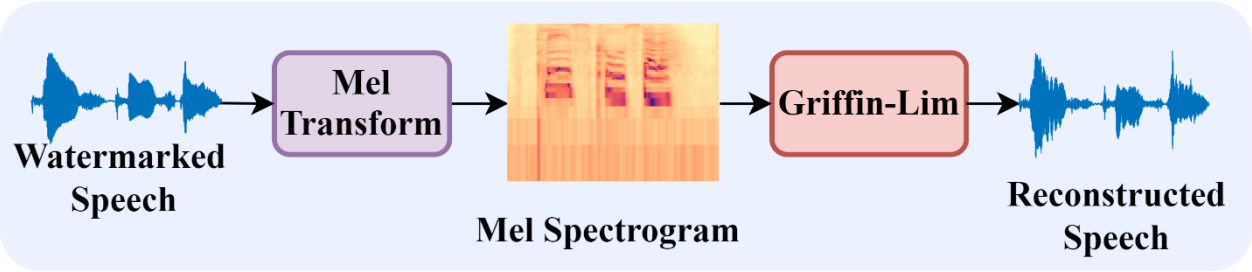
Waveform reconstruction distortion

Goal 2: Voice Cloning Robustness

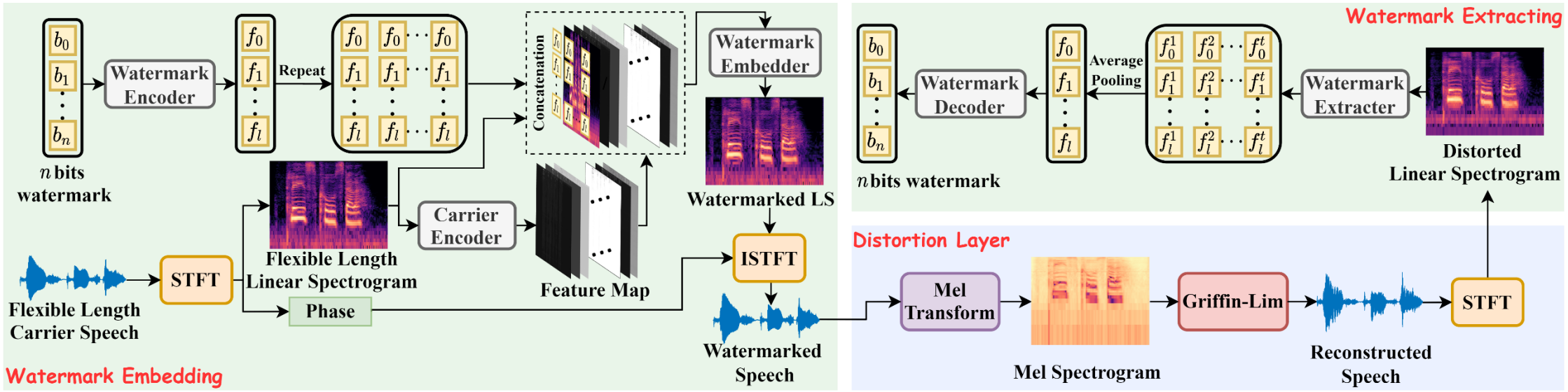




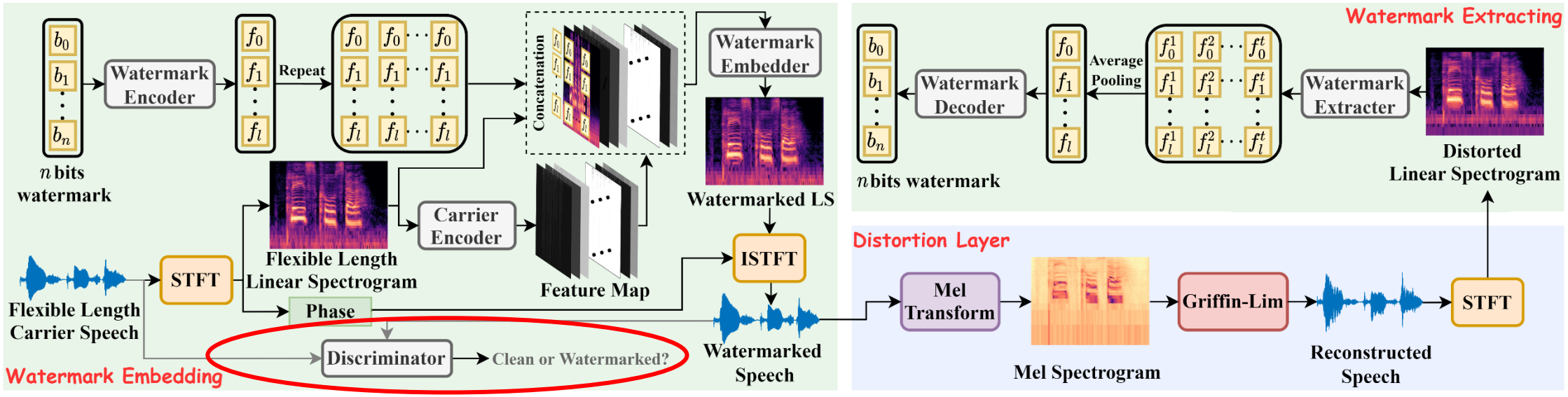
Insert



Goal 2: Voice Cloning Robustness



Goal 3: Fidelity



Experimental Setting

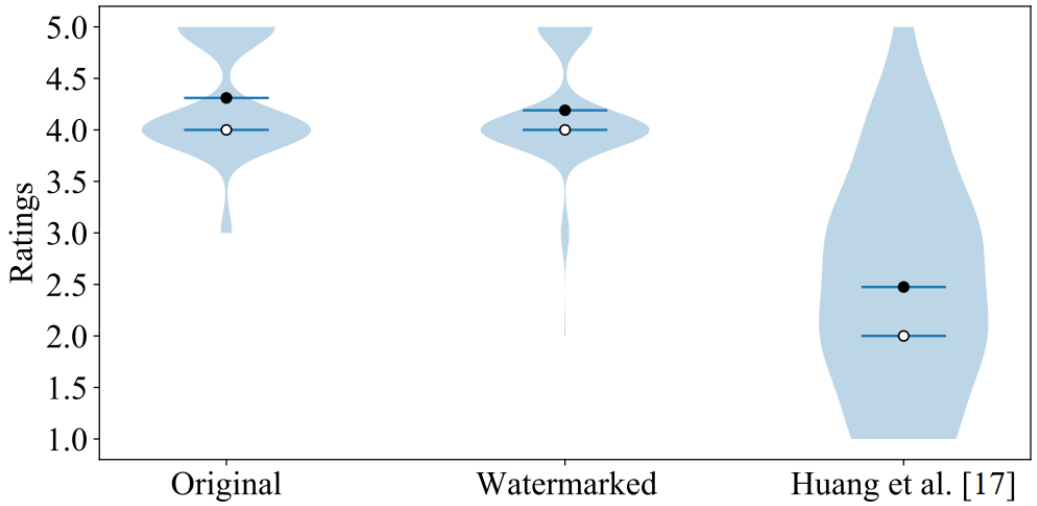
Setup

- **Voice cloning models:** FastSpeech2, Tacotron2 and VITS with LJSpeech as training set
- **Voice cloning API:**
  - PaddleSpeech, Voice-Clone-App with 10 segments as training set
  - so-vits-svc with 1 singing song as training set
- **Watermarking model training set:** LibriSpeech training set
- **Processing distortion testing set:** LibriSpeech test set
- **Voice cloning test:** 500 text segments from the LJSpeech test set

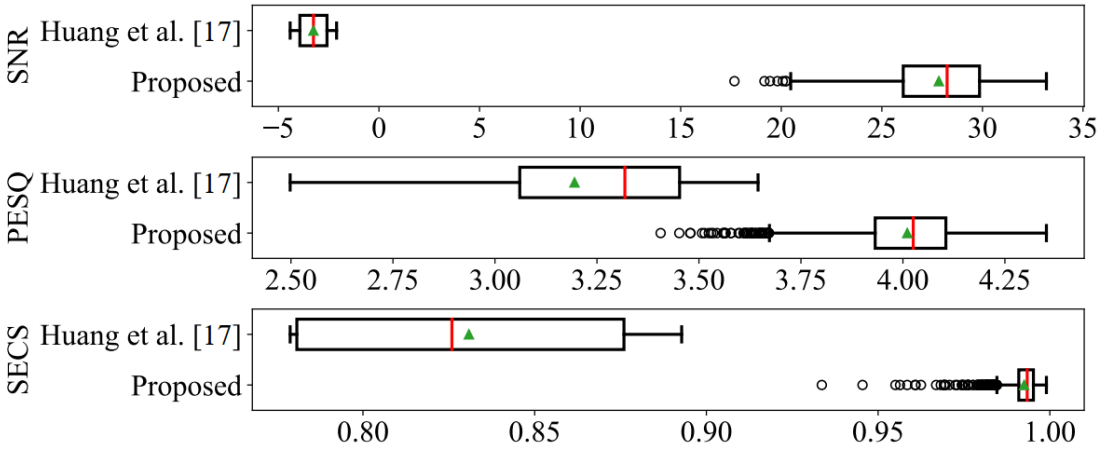
Evaluation metrics

- **Fidelity:** Signal-to-Noise Ratio (SNR), Perceptual Evaluation of Speech Quality(PESQ), Speaker Encoder Cosine Similarity (SECS), Mean Opinion Score(MOS) with five ratings
- **Robustness:** Bit recovery accuracy (ACC)

Fidelity Testing



Black dots are means and white dots are medians

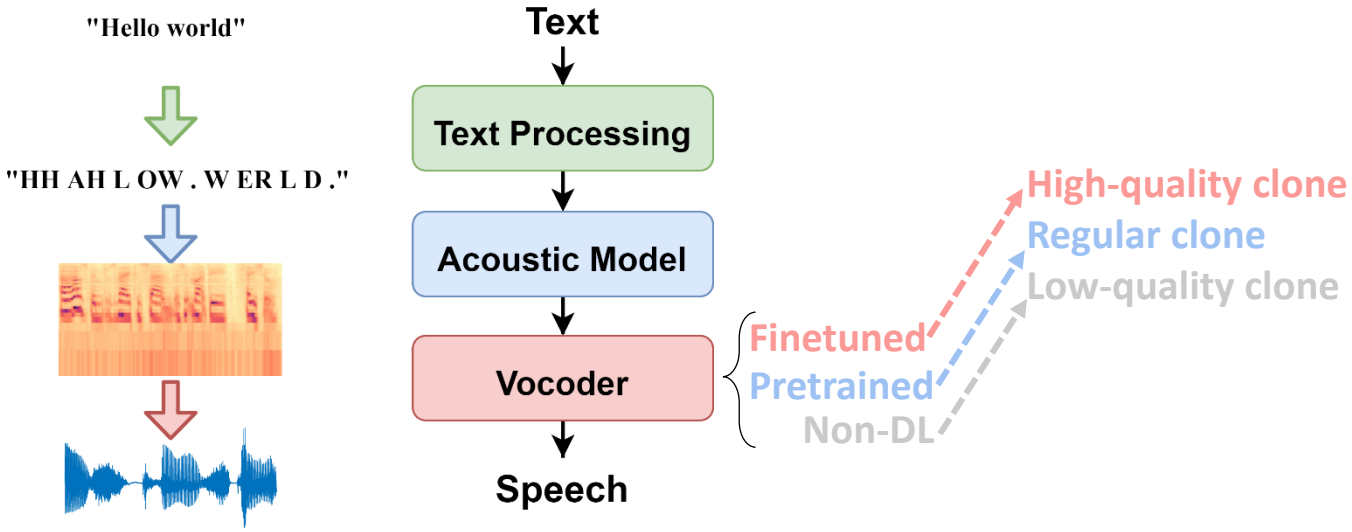


Green triangles represent the mean values and **red lines** indicate the median values.

Subjective experiments show that the watermarked audio is almost indistinguishable from the original audio.

Objective experiments indicate that the imperceptibility metrics of the watermarking scheme significantly surpass the baseline.

Robustness to Voice Cloning



Model		Quality		ACC↑
Acoustic Model	Vocoder	PESQ↑	SECS↑	
Fastspeech2* [8]	Hifi-GAN* [40]	1.0578	0.8957	1.0000
	Hifi-GAN [40]	1.0712	0.8965	0.9933
	Griffin-Lim [38]	1.1129	0.7034	1.0000
Tacotron2* [36]	Hifi-GAN* [40]	1.1143	0.8598	1.0000
	Hifi-GAN [40]	1.1136	0.8626	0.9988
	Griffin-Lim [38]	1.1971	0.7125	1.0000
VITS* [30] **(All in one)**		1.0342	0.9085	1.0000

* denotes using a watermarked dataset to train the acoustic model or fine-tune the vocoder, otherwise using a watermark-free dataset

The proposed watermarking method demonstrates strong robustness across various scenarios involving different acoustic models and vocoder combinations.

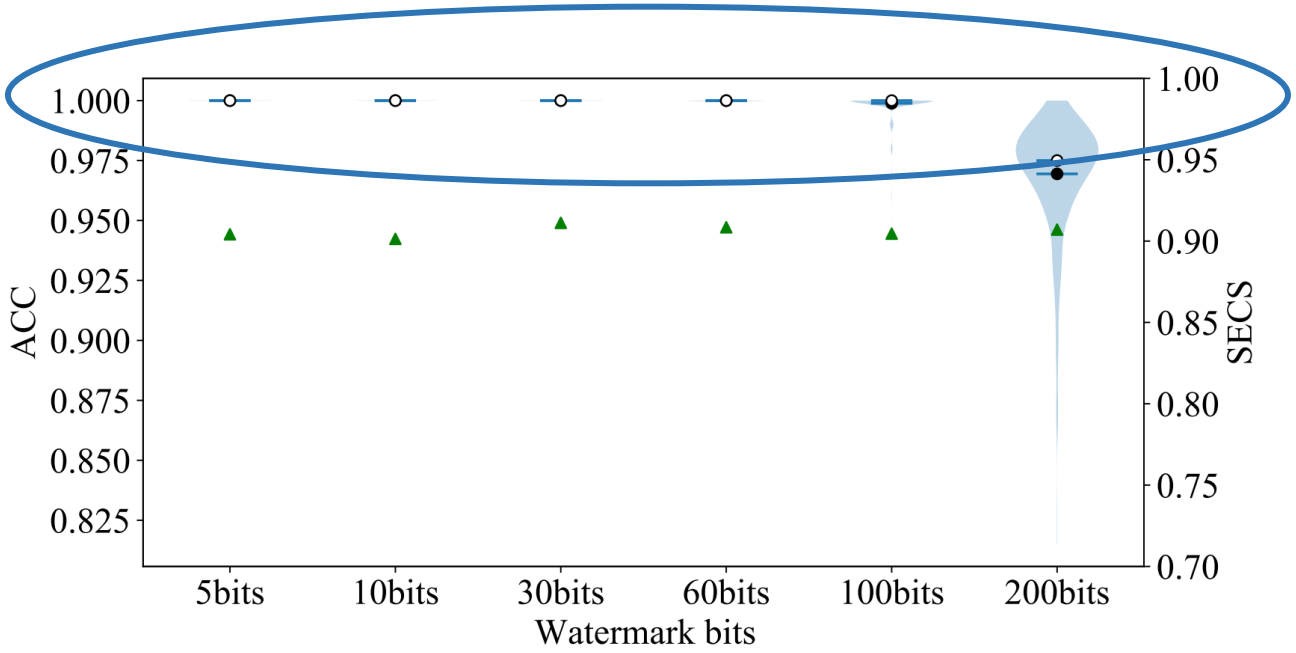
Robustness to Voice Cloning

Method	syn PESQ↑	syn SECS↑	wm SNR↑	wm ACC↑
FSVC [23]	0.9949	0.9139	21.1282	0.5554 (×)
RFDLM [21]	1.0303	0.9179	19.4668	0.5096 (×)
The Proposed	1.0342	0.9085	28.1650	1.0000 (✓)

Totally fail

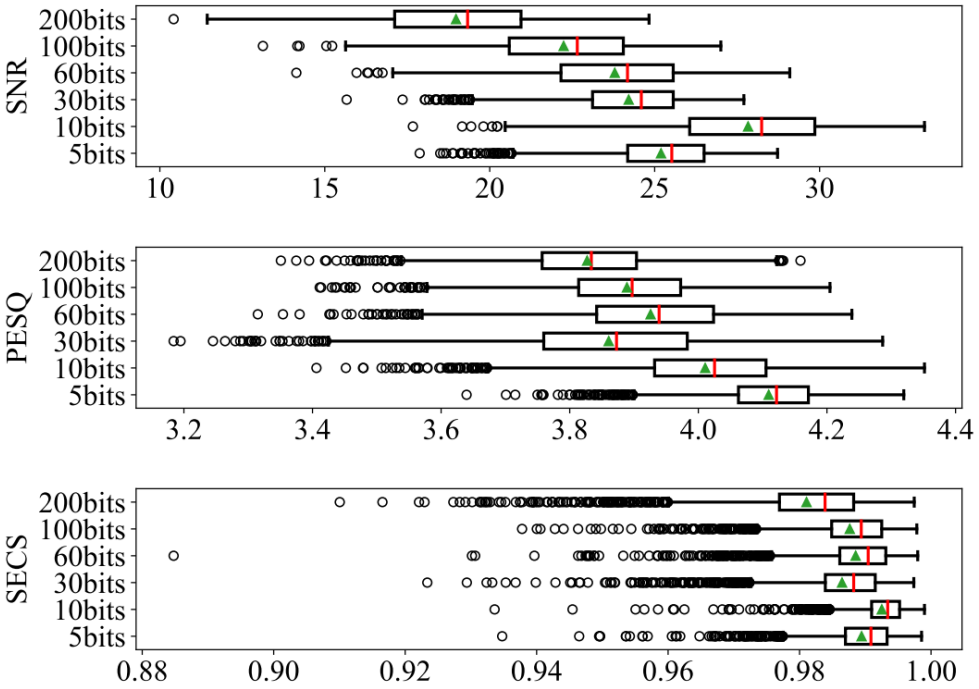
In the case of similar fidelity, the existing watermarking schemes can not resist speech cloning

Robustness to Voice Cloning



Black dots are mean accuracy and white dots are median accuracy
Green triangles represent the average SECS values of synthesized speech

It is possible to embed longer sequences of bits to address a wider range of scenarios



Green triangles represent the mean values and **red lines** indicate the median values.

Increasing the length of the embedded bits does not result in a noticeable degradation of audio quality.

Robustness to Real-world Black-box API

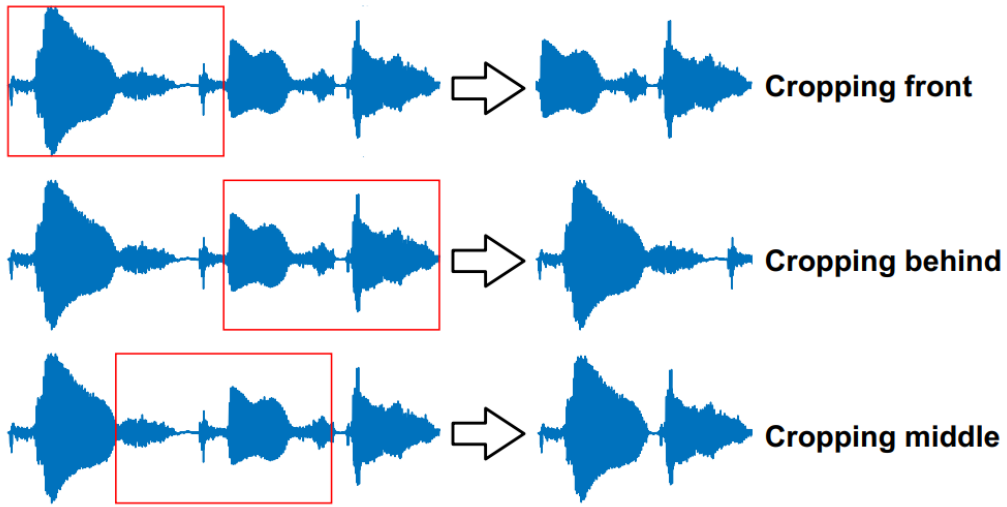
- PaddleSpeech (baidu aistudio)
- Voice-clone-App

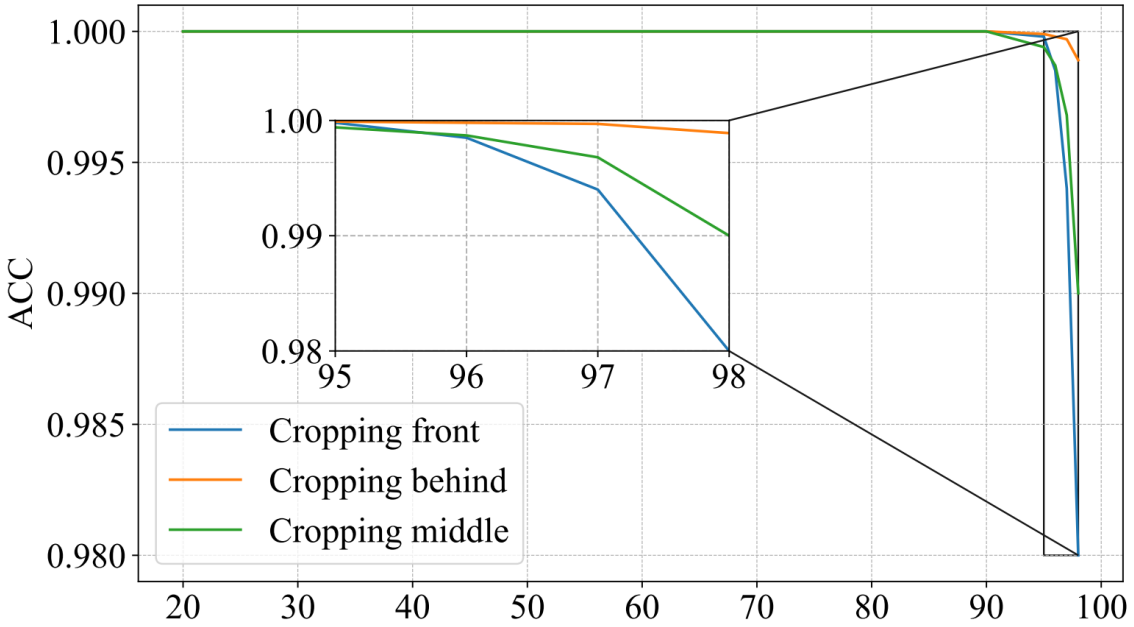
Service	Language	Metric	Speaker					
			P225	P226	P227	P228	P229	P230
PaddleSpeech [71]	English	PESQ↑	2.5958	2.7235	2.3573	2.3235	2.7419	1.7095
		SECS↑	0.8611	0.8701	0.8552	0.8537	0.8592	0.8519
		ACC↑	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
			D4	D6	D7	D8	D11	D12
	Chinese	PESQ↑	1.7642	1.9851	2.6490	2.0223	2.3808	1.2313
		SECS↑	0.7836	0.8034	0.7622	0.8219	0.7304	0.7103
		ACC↑	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
			P225	P226	P227	P228	P229	P230
Voice-Cloning-App [27]	English	PESQ↑	0.7809	1.5610	1.1913	1.1684	1.2601	1.2694
		SECS↑	0.7576	0.8564	0.7324	0.8781	0.8495	0.8799
		ACC↑	0.9000	0.9100	0.9000	0.9000	0.9500	0.9200

1. It performs well across scenarios involving different languages in the real world.

2. Low-quality synthetic speech does not significantly impact watermark extraction, still maintaining an extraction accuracy of over 90%.

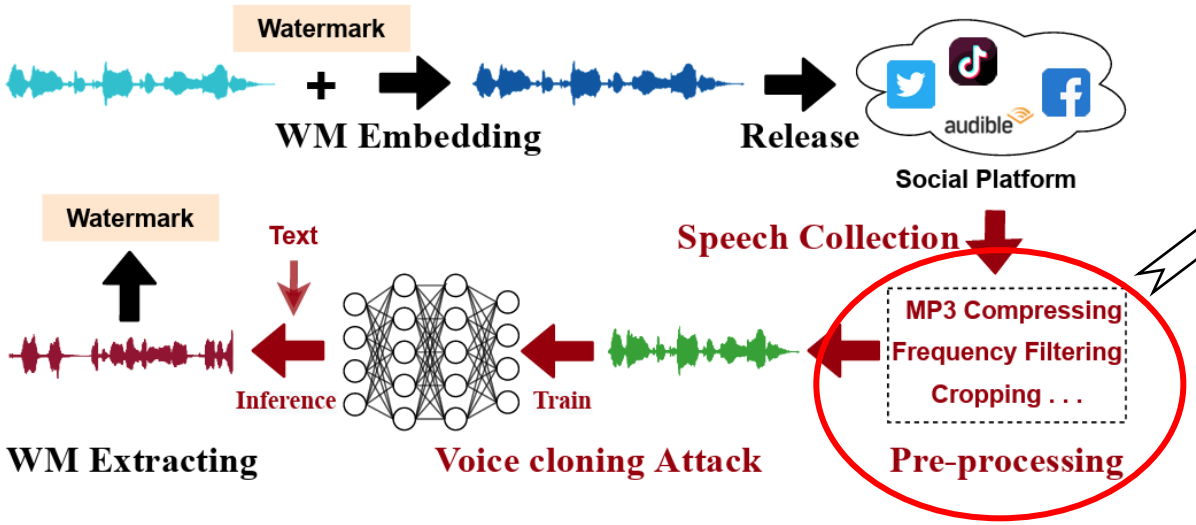
Robustness to Processing Distortions





The watermark can achieve 100% extraction even when 90% of audio is cropped.

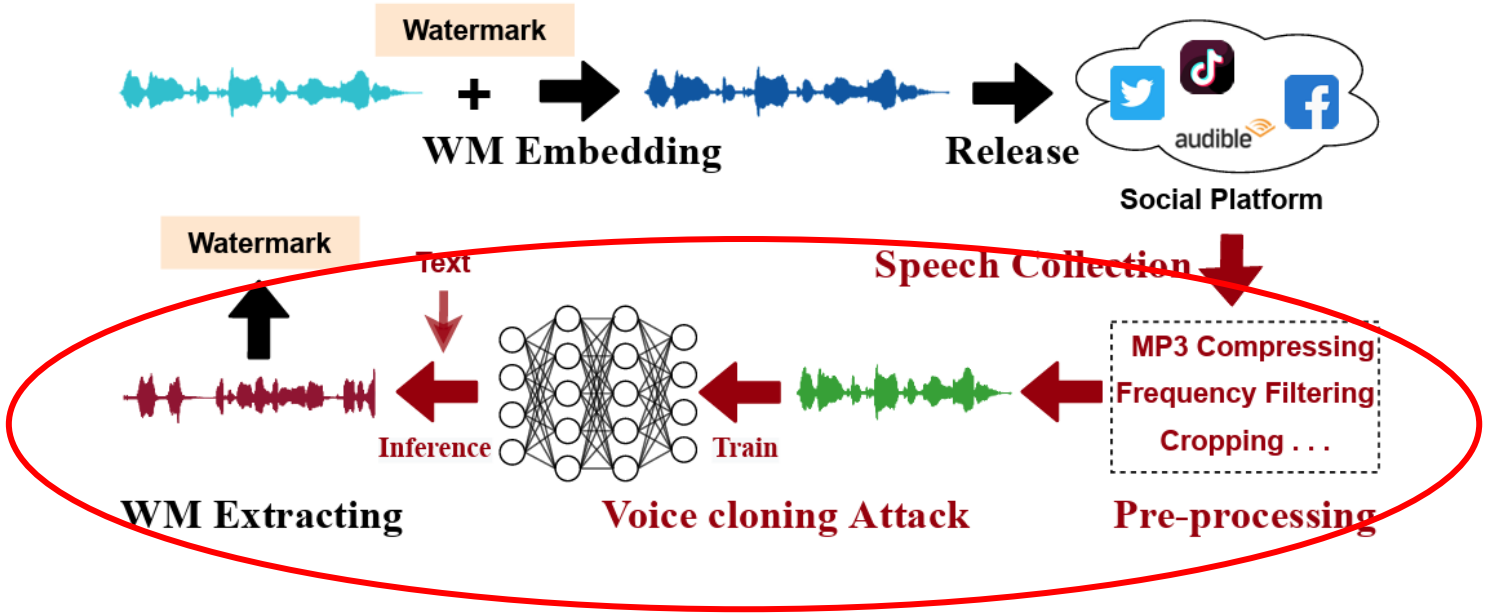
Robustness to Processing Distortions



Preprocessing	Parameter	Quality			ACC↑
		SNR↑	PESQ↑	SECS↑	
Resampling	16 kHz	34.8115	4.4967	1.0000	1.0000
	8 kHz	17.1642	4.4961	0.9025	0.9940
Amplitude Scaling	20%	1.9382	4.4918	0.9575	1.0000
	40%	4.4368	4.4973	0.9596	1.0000
	60%	7.9589	4.4986	0.9772	1.0000
	80%	13.9790	4.4991	0.9942	1.0000
MP3 Compression	8 kbps	9.0414	2.2115	0.7565	0.9186
	16 kbps	13.1554	3.3484	0.9552	0.9992
	24 kbps	15.2631	3.9259	0.9888	0.9999
	32 kbps	17.2272	4.0695	0.9962	1.0000
	40 kbps	18.7795	4.1902	0.9975	1.0000
	48 kbps	20.8746	4.3122	0.9986	1.0000
	56 kbps	22.8885	4.3813	0.9991	1.0000
	64 kbps	23.9958	4.4136	0.9992	1.0000
Recount	8 bps	22.9103	3.1708	0.9757	0.9995
Median Filtering	5 Samples	14.8666	3.6664	0.9459	1.0000
	15 Samples	8.9079	2.5726	0.7875	0.9933
	25 Samples	5.3999	2.1427	0.7338	0.9806
	35 Samples	3.2550	1.8721	0.6861	0.9402
Low Pass Filtering	2000 Hz	12.8558	3.8824	0.7280	0.9030
High Pass Filtering	500 Hz	3.7635	3.7919	0.6551	1.0000
Gaussian Noise	20 dB	20.0002	3.1287	0.9104	0.9962
	25 dB	24.9989	3.5182	0.9670	0.9995
	30 dB	29.9981	3.8662	0.9919	1.0000
	35 dB	34.9941	4.1277	0.9981	1.0000
	40 dB	39.9888	4.3038	0.9994	1.0000

The watermarking scheme can resist various processing operations. Considering the worst-case, it can achieve an extraction accuracy of over 90%.

Robustness to Processing + Voice Cloning



Robustness to Processing + Voice Cloning

Preprocessing	Parameter	Quality			ACC↑
		SNR↑	PESQ↑	SECS↑	
Resampling	16 kHz	34.8115	4.4967	1.0000	1.0000
	8 kHz	17.1642	4.4961	0.9025	0.9940
Amplitude Scaling	20%	1.9382	4.4918	0.9575	1.0000
	40%	4.4368	4.4973	0.9596	1.0000
	60%	7.9589	4.4986	0.9772	1.0000
	80%	13.9790	4.4991	0.9942	1.0000
MP3 Compression	8 kbps	9.0414	2.2115	0.7565	0.9186
	16 kbps	13.1554	3.3484	0.9552	0.9992
	24 kbps	15.2631	3.9259	0.9888	0.9999
	32 kbps	17.2272	4.0695	0.9962	1.0000
	40 kbps	18.7795	4.1902	0.9975	1.0000
	48 kbps	20.8746	4.3122	0.9986	1.0000
	56 kbps	22.8885	4.3813	0.9991	1.0000
	64 kbps	23.9958	4.4136	0.9992	1.0000
Recount	8 bps	22.9103	3.1708	0.9757	0.9995
Median Filtering	5 Samples	14.8666	3.6664	0.9459	1.0000
	15 Samples	8.9079	2.5726	0.7875	0.9933
	25 Samples	5.3999	2.1427	0.7338	0.9806
	35 Samples	3.2550	1.8721	0.6861	0.9402
Low Pass Filtering	2000 Hz	12.8558	3.8824	0.7280	0.9030
High Pass Filtering	500 Hz	3.7635	3.7919	0.6551	1.0000
Gaussian Noise	20 dB	20.0002	3.1287	0.9104	0.9962
	25 dB	24.9989	3.5182	0.9670	0.9995
	30 dB	29.9981	3.8662	0.9919	1.0000
	35 dB	34.9941	4.1277	0.9981	1.0000
	40 dB	39.9888	4.3038	0.9994	1.0000

	Pre-processing	PESQ↑	SECS↑	ACC↑
Regular	Resampling 16K	1.0775	0.9122	1.0000
	Mp3 Compression 64kbps	1.0347	0.9077	1.0000
	Combined	1.0776	0.9064	1.0000
Harmful	Mp3 Compression 8kbps	0.8284	0.6675	0.8996
	Low Pass Filtering 2000 Hz	1.0836	0.6481	0.9482
	Combined	1.0324	0.6567	0.9144

When combining watermark-erasuring preprocessing with voice cloning, the watermark still maintains high robustness.

Conclusion

- For the first time, we introduce the concept of "Timbre Rights" and propose a "Timbre Watermarking" scheme as an effective means of protection.
- To achieve "Timbre Watermarking", we propose a novel end-to-end voice cloning-resistant audio watermarking framework.
- Extensive experiments demonstrate that the proposed method can achieve robustness against traditional distortions and voice cloning distortion while guaranteeing the requirement of fidelity.

中国科学技术大学
University of Science and Technology of China

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

NDSS SYMPOSIUM/2024

THANK YOU!

Demo and code website: https://timbrewatermarking.github.io

Contact with any questions: hichangliu@mail.ustc.edu.cn