Proactive Detection of Speaker Identity Manipulation With Neural Watermarking

Wanying Ge Xin Wang Junichi Yamagishi

National Institute of Informatics, Tokyo, Japan

TL;DR

We embed speaker identity into speech as a watermark that can be verified upon reception.

While the proposed approach is robust against many attacks, neural codecs severely degrade detection performance.

Motivation

Deepfake speech and voice conversion technologies can seamlessly manipulate speaker identity, raising serious security and privacy concerns.

Rather than injecting random bits, we employ watermarking to embed traceable speaker information directly into the speech waveform. The **imperceptible yet detectable** speaker embedding can ensure security and traceability under various **transmission channels** and **identity manipulation attacks**.

Transmissions and attacks

- Transmission:
  - Echo
  - Gaussian noise
  - Low-pass filtering
  - Waveform quantization
  - Neural codec: DAC, EnCodec
- Identity manipulation attacks:
  - DSP-based attacks:
    - Clipping
    - Resampling
    - Pitch Shift
  - NN-based Attack:
    - k-Nearest Neighbors Voice Conversion (kNN-VC).

Findings

- The approach is effective at detecting manipulation attacks if important speaker information remains intact during transmission.
- However, if the speaker information is altered, the system can easily identify the manipulation.
- While neural codec transmissions do not distort speaker information, they interfere with watermark detection, leading to nearly random embedding reconstruction.

References

[1] B. Desplanques, J. Thienpondt, and K. Demuynck. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. pages 3830–3834, 2020.

[2] C. Liu, J. Zhang, T. Zhang, X. Yang, W. Zhang, and N. Yu. Detecting voice cloning attacks via timbre watermarking. In *Network and Distributed System Security Symposium*, 2024.

[3] S. Wang, P. Zhu, and H. Li. M-Vec: Matryoshka speaker embeddings with flexible dimensions. *arXiv preprint arXiv:2409.15782*, 2024.

Proposed framework

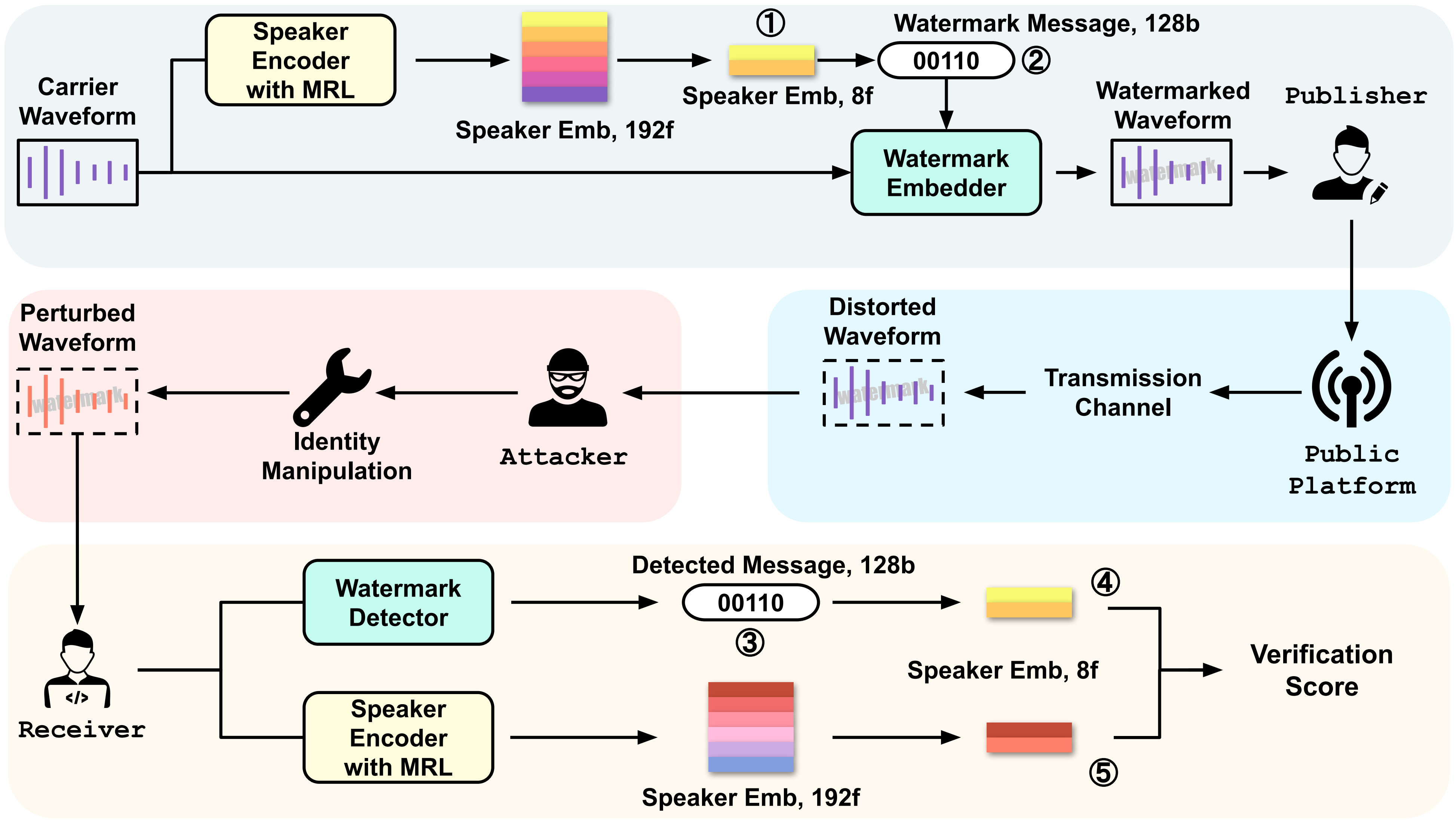

Figure 1. Proposed framework. Speaker embedding is first extracted from the carrier waveform and partially binarized (①) to form a watermark message (②). This message is then embedded into the waveform before transmission. After potential identity manipulation attacks, the receiver extracts the watermark (③) and reconstructs speaker embedding (④). Simultaneously, the speaker embedding is directly extracted from the distorted waveform (⑤). The similarity between these embeddings determines if the speaker's identity has been compromised. MRL: Matryoshka Representation Learning [3].

EER results of identity manipulation detection

	Column ID		1	2	3	4	5	6	7	8	9
			Mild Identity Manipulation				→		Aggressive Identity Manipulation		
	Transmission / Attack		Clipping	kNN	Pitch Shift (by semitones)				Resampling		
		Params	70%	-	2	4	6	8	4 kHz	8 kHz	22.05 kHz
	None	-	1.11	0.02	0.02	0.06	0.06	0.04	0.00	0.04	0.04
DSP	Echo	0.7	2.59	0.18	0.10	0.31	0.29	0.31	0.04	0.14	0.27
		0.5	2.48	0.10	0.06	0.18	0.23	0.14	0.00	0.12	0.14
		0.3	2.65	0.21	0.16	0.27	0.27	0.25	0.18	0.18	0.23
		0.1	3.14	0.27	0.23	0.37	0.29	0.37	0.00	0.14	0.29
	Gaussian	40 dB	1.27	0.06	0.06	0.12	0.08	0.14	0.00	0.00	0.14
		20 dB	7.06	3.96	3.75	4.27	4.45	4.01	3.96	4.02	4.04
		10 dB	22.26	14.61	15.57	16.84	16.78	16.00	16.41	16.66	15.43
		5 dB	35.51	22.79	27.37	27.60	28.95	27.33	28.48	28.66	26.12
	Lowpass	6.4 kHz	0.82	0.14	0.08	0.27	0.16	0.14	0.02	0.10	0.12
		4.8 kHz	8.41	5.38	4.94	6.22	5.97	5.79	5.17	5.46	5.58
		3.2 kHz	35.49	25.97	24.83	27.51	30.20	32.38	28.15	27.86	31.53
		1.6 kHz	45.92	34.51	41.01	40.13	42.53	33.63	41.36	42.06	43.76
	Quantization	64-bit	4.04	2.61	2.24	2.48	2.67	2.34	2.07	2.48	2.28
		32-bit	8.27	5.79	5.58	6.52	6.36	5.79	5.68	6.01	6.05
		16-bit	20.48	15.51	15.55	16.25	16.02	15.98	16.04	16.70	16.04
		8-bit	34.41	26.53	29.71	30.76	31.37	32.58	31.08	31.58	30.14
NN	DAC	-	44.09	38.22	41.01	38.20	41.71	39.39	41.55	39.84	41.51
	EnCodec	24 kbps	48.95	51.78	48.81	51.87	53.75	55.81	53.16	54.16	53.61
		12 kbps	49.96	53.34	49.55	53.10	54.92	58.25	54.29	55.21	54.66
		6 kbps	50.90	53.73	50.02	52.28	54.55	59.27	55.70	57.00	50.10
		3 kbps	54.37	57.35	51.66	55.17	58.64	65.31	59.77	64.28	49.20

Speaker encoder and Watermark model

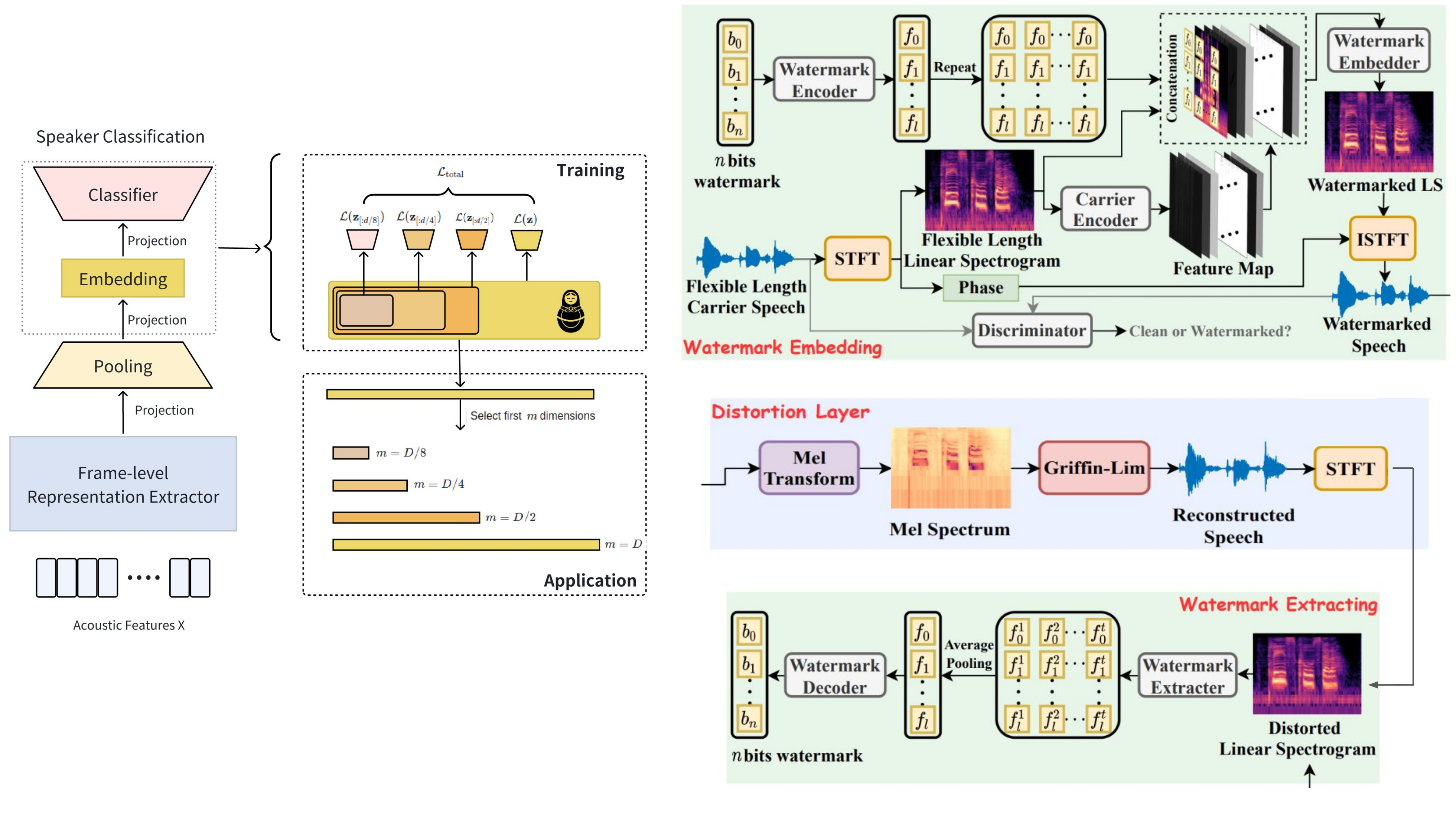

Figure 2. ECAPA-TDNN speaker encoder (left) and Timber watermarking model (right.). Figure reproduced from [1] and [3].