Combining Self-Embedding Audio Watermarking with Ultra-Low-Bitrate Neural Codecs

Yigitcan Özer, Xin Wang, Zhe Zhang, Junichi Yamagishi
National Institute of Informatics, Tokyo, Japan
Email: {yiitozer, wangxin, zhe, jyamagis}@nii.ac.jp

***Abstract*—Partial manipulation of speech recordings, where only localized segments of an utterance are altered, poses a significant challenge for content integrity verification, as reliable detection and localization of such edits becomes harder as the manipulated proportion decreases. Watermarking offers a proactive defense alternative by embedding auxiliary information prior to distribution; classical hash-based schemes achieve near-perfect detection and localization under ideal conditions, but the original content cannot be recovered once a segment is manipulated. Building on a prior self-embedding audio steganography framework, this work presents an initial exploration of proactive defense performance under ideal conditions, extending the investigation along three axes: frame-level localization, multi-bit least significant bit variants, and evaluation across multiple ultra-low-bitrate neural codec representations. By embedding a compact neural codec representation rather than a cryptographic hash, the framework additionally enables recovery of the manipulated regions, while supporting training-free detection and localization without spoofed examples. Experiments across four controlled manipulation types under ideal channel conditions show that the embedded payload, and hence an approximate reconstruction of the authentic content, is always fully recovered without bit errors. The results also indicate that the choice of neural codec is the dominant factor for detection and localization performance.**

***Index Terms*—audio watermarking, partial deepfake detection, self-embedding watermarking, self-recovery watermarking**

I. Introduction

Partial manipulation of speech recordings poses a fundamental challenge for audio integrity [1]. Unlike fully synthesized speech, such manipulations preserve most of the original recording while altering only a localized segment [2]. Proactive defense strategies, e.g., watermarking, address this by embedding auxiliary information into the signal prior to distribution [3], enabling subsequent integrity verification.

In classical watermarking, fragile, semi-fragile, and robust watermarks serve different security goals [4]. Robust watermarks are engineered to survive substantial signal processing and are appropriate for ownership tracing or provenance verification [5], [6]. Conversely, precise localization of altered segments rather requires fragile or semi-fragile paradigms. Since fragile watermarks are designed to break under the slightest waveform alterations, any local tampering selectively corrupts the embedded payload only within the modified interval. This sharp contrast between intact and broken payload segments provides a precise spatial indicator of the tampered boundaries under ideal channel conditions. Traditionally, these paradigms embed a cryptographic hash of each audio segment, which localizes tampering exactly but cannot recover the original spoken content once a segment has been overwritten [7], [8]. Recently, a neural semi-fragile watermarking approach has been proposed for proactive deepfake speech detection [9]. However, such learned schemes require training on spoofed examples and cannot recover the original content, which would require embedding the audio itself as the payload.

Self-embedding watermarking, also referred to as self-recovery or self-healing watermarking, embeds an authentication and recovery description derived from the host signal into the carrier signal itself. This idea has been extensively studied for image and video authentication and recovery [10]–[13]. In the audio domain, early speech self-embedding schemes modeled the problem as source–channel coding: a compressed representation of the speech signal, protected by channel coding and accompanied by frame-level hash information, is embedded into the signal, allowing tampered frames to be detected and subsequently reconstructed [14]. Subsequent work improved embedding transparency via auditory masking [15] and a further study addressed content replacement with duration changes [16]. These studies establish that audio self-embedding is possible, but they rely on digital signal processing (DSP)-based codecs to generate the recovery payload, e.g., G.723.1 at $6.4\,\mathrm{kbps}$ and MELP at $2.4\,\mathrm{kbps}$ in [14], and OPUS at $64\,\mathrm{kbps}$ in [16]. Since the payload must fit within the strictly limited embedding capacity, the codec bitrate directly governs the achievable redundancy, and hence the extent of manipulation from which the content can be recovered. Even the lowest-rate standardized DSP-based coders, such as MPEG-4 HVXC, operate at $2.0\,\mathrm{kbps}$ [17], whereas recent neural audio codecs have pushed speech coding below $1\,\mathrm{kbps}$ while preserving intelligible, speaker-consistent reconstruction [18], reducing the recovery payload to a fraction of the bitrates used in prior audio self-embedding schemes.

In this work, we revisit audio self-embedding in light of recent ultra-low-bitrate neural audio codecs [19]–[21]. We embed a compact neural codec representation of the carrier signal into itself; at verification time, the extracted payload is decoded to obtain a self-reconstruction of the authentic speech, and the mismatch between this reconstruction and the received signal is used for detection and localization (see Fig. 1). This design preserves the central advantage of

This work was supported by JSPS MEXT KAKENHI Grant 24H00732.

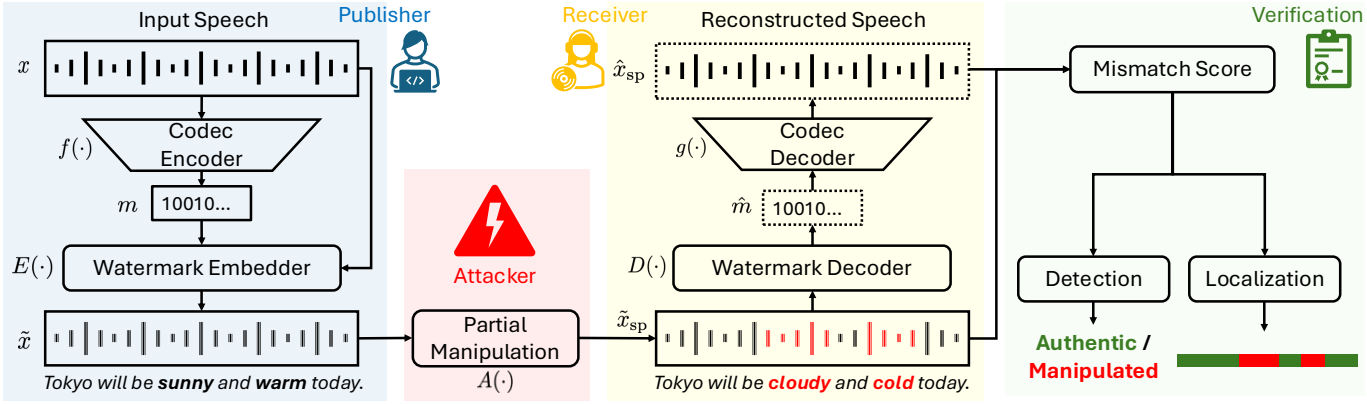

Fig. 1. Overview of the proposed self-embedding audio watermarking framework. At the publisher side, the carrier signal is encoded by a neural codec and embedded into itself via a watermark embedder. A partial manipulation is applied to the watermarked signal. At the receiver side, the embedded bitstream is recovered by a watermark decoder and passed to a codec decoder to reconstruct the authentic waveform. Finally, a mismatch score between the received signal and its reconstruction is used jointly for utterance-level detection and frame-level localization of manipulated segments.

self-recovery, which is the ability to reconstruct manipulated content, while allowing us to study whether recent sub-kbps neural codecs provide a practical payload for modern partial speech manipulation scenarios.

Our prior work introduced an initial version of this neural codec self-embedding framework for partial speech manipulation detection [22], using least significant bit (LSB) embedding [3] with SNAC as the compression backbone [19]. The present paper extends this framework in three directions: *(i)* we extend evaluation from utterance-level detection to frame-level localization; *(ii)* we generalize single-bit LSB embedding to multi-bit fragile variants ($k \in \{1, 2, 4\}$); *(iii)* we evaluate the framework across multiple ultra-low-bitrate neural codecs and manipulation types, including replacement, deletion, and insertion. Experiments show that detection and localization performance are governed primarily by neural codec reconstruction fidelity.

II. Self-Embedding Audio Watermarking using Neural Audio Codecs

The proposed framework embeds an ultra-low-bitrate neural codec representation of the carrier signal into itself, enabling a manipulated segment to be approximately reconstructed from the recovered payload as long as a sufficient portion of the watermarked signal remains intact. This section formalizes the embedding and verification pipeline, details the LSB embedding method, and analyzes the embedding capacity across codec configurations.

A. Problem Formulation

This formulation extends [22] to multiple bit depths and neural audio codecs. Let $\mathbf{x} \in \mathbb{R}^T$ denote a discrete-time speech signal of length T, sampled at f_s Hz. An embedder $E : \mathbb{R}^T \times \{0,1\}^M \to \mathbb{R}^T$ conceals a binary message $\mathbf{m} \in \{0,1\}^M$ within $\mathbf{x}$, producing a watermarked signal $\tilde{\mathbf{x}} = E(\mathbf{x}, \mathbf{m})$ that remains perceptually indistinguishable from the original. A decoder $D : \mathbb{R}^T \to \{0,1\}^M$ recovers the message as $\hat{\mathbf{m}} = D(\tilde{\mathbf{x}})$, where M denotes the total embedding budget (in bits) for a signal of length T.

The central constraint is that the embedding budget M is strictly limited by imperceptibility, satisfying $M \ll B \cdot T$, where B is the bit depth. As a result, embedding the waveform $\mathbf{x}$ itself, or even a lossless encoding of $\mathbf{x}$, is infeasible. More importantly, this constraint precludes redundancy through repeated embedding, which is required for robustness against localized manipulations.

We overcome this limitation by introducing a neural audio codec operating at bitrate ρ (bps) as a compression backbone. Let $f : \mathbb{R}^T \to \{0,1\}^N$ denote the codec encoder, where $N = \rho \cdot (T/f_s)$ is the number of bits in the compressed representation. For modern ultra-low-bitrate codecs ($\rho < 1.0\,\text{kbps}$), the following inequalities hold:

$$\rho \cdot \frac{T}{f_s} \ll M \ll B \cdot T, \tag{1}$$

where the left term is the codec representation size, the middle term is the embedding budget, and the right term is the raw waveform size in bits. Notably, the left inequality implies that the compressed representation $f(\mathbf{x})$ is sufficiently small to be embedded multiple times within the available embedding budget. This enables a self-embedding design with explicit redundancy. The message is defined as R consecutive repetitions of the codec representation:

$$\mathbf{m} = (m_1, m_2, \ldots, m_M) = \underbrace{f(\mathbf{x}) \parallel \cdots \parallel f(\mathbf{x})}_{R}, \tag{2}$$

where $\parallel$ denotes concatenation and $R = \lfloor M/N \rfloor$ is the maximum number of non-overlapping repetitions that fit within the embedding budget M. Equivalently, the i^{th} bit of the r^{th} repetition is $m_i^{(r)} = m_{(r-1)N+i}$ for $i = 1, \ldots, N$ and $r = 1, \ldots, R$, making the repetition structure of $\mathbf{m}$ explicit. This redundancy enables majority-vote decoding at extraction time, making the recovered codec representation robust to partial manipulations of the watermarked signal.

To model partial manipulations, we introduce a spoofing operator $A(\cdot)$ that modifies localized time intervals $\{\mathcal{I}_k\}_{k=1}^K$, with $\mathcal{I}_k = [t_{\text{start}}^{(k)}, t_{\text{end}}^{(k)}]$, by substitution, insertion, or deletion.

Since $\mathbf{m}$ is embedded within $\tilde{\mathbf{x}}$, the manipulation $A(\tilde{\mathbf{x}})$ simultaneously alters the signal content and corrupts the embedded payload within the affected intervals, yielding

$$\mathbf{m}' = f(\mathbf{x}) \parallel \cdots \parallel f'(\mathbf{x}) \parallel \cdots \parallel f(\mathbf{x}),$$

where $f'(\mathbf{x})$ denotes a repetition corrupted by the manipulation, and intact repetitions $f(\mathbf{x})$ remain unaffected. The decoder $D(\mathbf{y})$ recovers the N-bit codec representation $\hat{\mathbf{m}} \in \{0,1\}^N$ by majority voting across the R repetitions embedded in $\mathbf{m}'$:

$$\hat{m}_i = \mathbf{1}\left[\frac{1}{R}\sum_{r=1}^{R} m_i'^{(r)} \geq 0.5\right], \quad i = 1, \ldots, N, \tag{3}$$

where $m_i'^{(r)}$ denotes the i^{th} bit of the r^{th} repetition in $\mathbf{m}'$, yielding a reliable estimate of $f(\mathbf{x})$ from the intact copies. The framework makes no assumptions about the internal mechanism of $A(\cdot)$, and thus covers both TTS-based synthesis and waveform-level splicing operations.

At verification time, let $\mathbf{y}$ denote the received signal, where $\mathbf{y} = \tilde{\mathbf{x}}$ in the authentic case and $\mathbf{y} = \tilde{\mathbf{x}}_{\text{sp}} = A(\tilde{\mathbf{x}})$ denotes the spoofed signal under partial manipulations. A codec decoder $g : \{0,1\}^N \to \mathbb{R}^T$ reconstructs the authentic waveform from $\hat{\mathbf{m}}$, yielding the self-reconstruction $R(\mathbf{y}) = g(D(\mathbf{y}))$, and the mismatch score under a dissimilarity measure d is $s(\mathbf{y}) = d\big(R(\mathbf{y}), \mathbf{y}\big)$. Detection is formulated as the binary hypothesis test:

$$\delta(\mathbf{y}) = \begin{cases} 0, & s(\mathbf{y}) \leq \tau, \\ 1, & s(\mathbf{y}) > \tau, \end{cases} \tag{4}$$

where τ denotes a decision threshold, $\delta(\mathbf{y}) = 0$ indicates an *authentic* signal, and $\delta(\mathbf{y}) = 1$ indicates a *manipulated* signal.

Beyond utterance-level detection, the framework naturally extends to frame-level localization. By projecting local alignment costs (e.g., from DTW) onto the time axis of $\mathbf{y}$, manipulated regions can be identified directly from the resulting per-frame mismatch profile, as detailed in the next section.

B. Detection and Localization Score via Dynamic Time Warping (DTW)

For the proposed self-embedding watermarking approach, detection and localization scores are derived from the DTW alignment between the received signal $\mathbf{y}$ and its self-reconstruction $R(\mathbf{y}) = g(D(\mathbf{y}))$ (see § II-A). DTW provides a mechanism for aligning two time sequences under possible local temporal deviations [23]. It is particularly suitable in our setting, as partial speech manipulations introduce temporal inconsistencies between $\mathbf{y}$ and $R(\mathbf{y})$. By allowing non-linear temporal alignment, DTW separates structural mismatches from trivial timing offsets. We apply DTW to normalized log-mel spectrogram features, following [22].

Let π^* denote the optimal warping path obtained by $\text{DTW}(R(\mathbf{y}), \mathbf{y})$, and let c_t denote the local cosine distance between the two feature sequences at step t along π^*. The utterance-level detection score is defined as the mean of the smoothed per-step costs:

$$s_{\text{DTW}}(\mathbf{y}) = \frac{1}{|\pi^*|}\sum_{t \in \pi^*} c_t,$$

where higher values indicate a higher likelihood of manipulation.

For frame-level localization, the local cost c_t is projected back onto the time axis of the received signal $\mathbf{y}$ by aggregating costs per target frame along π^*. The resulting per-frame cost profile is smoothed and compared against binary ground-truth (GT) labels derived from word-level boundaries for frame-level localization evaluation.

C. Capacity and Repetition Analysis

A key design parameter of the proposed framework is the number of repetitions R of the codec representation that can be embedded within the available capacity, as this directly determines robustness to partial manipulation under majority-vote decoding. Let C denote the embedding capacity in bps of the chosen method, related to the total embedding budget M in § II-A by $M = C \cdot T/f_s$. Accordingly, we express all payload quantities in bps when analyzing repetition rates. The proposed framework is agnostic to this choice, and any embedding method satisfying the capacity constraint in (1) can serve as a drop-in replacement. The embedding method carries the same self-referential payload $\mathbf{m}$ defined in (2), where each repetition of the codec representation $f(\mathbf{x})$ is prepended with a 64-bit synchronization preamble. The number of repetitions per second is therefore $R = \lfloor C/(\rho + 64) \rfloor$, where ρ accounts for the codec representation size and the additional 64 bits for the synchronization preamble. The specific values of C for each embedding method are derived in § II-D, and the codec-specific values of ρ are detailed in § II-E.

D. Least Significant Bit (LSB) Embedding

In this work, we instantiate the framework with LSB encoding. Rather than applying LSB in its standard form, we adopt a temporally repetitive embedding strategy that enables robust payload recovery even when portions of the watermarked signal are manipulated, as detailed in the following.

LSB embedding encodes information by replacing the k least significant bits of each audio sample with payload bits, where $k \in \{1, 2, 4\}$ controls the capacity–distortion tradeoff. For a k-bit scheme, the k consecutive payload bits $(m_{ki}, \ldots, m_{ki+k-1}) \in \{0,1\}^k$ are packed into the k least significant bits of the audio sample x_i, introducing a maximum perturbation of $2^k - 1$ quantization steps per sample. The embedding capacity scales linearly with k as $C = k \cdot f_s$ bps; i.e., at $f_s = 16\,\text{kHz}$, the three variants yield capacities of 16,000, 32,000, and 64,000 bps for $k = 1, 2, 4$, respectively[1].

To account for partial manipulations to the watermarked signal, the codec representation $f(\mathbf{x})$ is embedded repeatedly across the full duration of the carrier, as explained in § II-A.

[1]The $k = 1$ case recovers the standard single-bit LSB scheme evaluated in our prior work [22].

TABLE I
NUMBER OF REPETITIONS (R) FOR EACH CODEC–BIT-DEPTH COMBINATION, COMPUTED FOR A 1 S CARRIER.

Codec	Bitrate (kbps)	Repetitions R (per second)		
		LSB-1	LSB-2	LSB-4
SNAC	0.98	15	30	61
SemantiCodec	0.65	22	44	89
TAAE	0.40	18	36	74

The r^{th} repetition begins at sample index $r\lceil N/k \rceil$, yielding R non-overlapping copies of the payload. At extraction time, each bit is recovered via majority voting across repetitions as defined in (3), yielding a reliable estimate of $f(\mathbf{x})$ from the intact copies even when some repetitions are corrupted by partial manipulations.

E. Ultra-Low-Bitrate Neural Audio Codecs

We evaluate three openly available ultra-low-bitrate neural codecs in this work. SNAC [19] extends the RVQ framework by introducing quantizers operating at multiple temporal resolutions, achieving $0.98\,\text{kbps}$ for speech at $24\,\text{kHz}$. SemantiCodec [20] adopts a fundamentally different architecture, decoupling semantic and acoustic information across two VQ layers and employing a latent diffusion model as the decoder [24], reaching bitrates as low as $0.31\,\text{kbps}$. The Transformer Audio AutoEncoder (TAAE) [21] departs from the convolutional paradigm by scaling a transformer-based encoder–decoder to approximately 950M parameters and replacing RVQ with Finite Scalar Quantization (FSQ) [25], achieving $0.4\,\text{kbps}$ and $0.7\,\text{kbps}$ for speech at $16\,\text{kHz}$. However, FSQ indices can exceed the `uint16` range. Consequently, each token requires `int32` storage (32 bits). The actual embedded payload sizes therefore differ substantially from the nominal bitrates, as reflected in Table I. Note that all codec–bit-depth pairs satisfy $R \geq 1$ with sufficient margin for majority-vote decoding, achieving exact payload recovery under all manipulation conditions considered in this work.

III. MANIPULATION CONDITIONS AND DATASET

To assess the performance limits of the proposed framework under ideal conditions for proactive defense, we evaluate across four controlled manipulation types that cover qualitatively distinct ways in which a watermarked signal may be altered. No channel degradation, compression, or additive noise is applied; the evaluation is designed to characterize what accuracy is achievable in the absence of such factors.

Direct replacement. One or more word-level segments of $\tilde{\mathbf{x}}$ are replaced with acoustically matched material drawn directly from authentic recordings of the same speaker, without any re-synthesis. This is a purely local manipulation detectable solely through payload mismatch.

TTS replacement. Synthesized segments produced by zero-shot voice cloning systems [26], [27] are substituted for the corresponding regions of $\tilde{\mathbf{x}}$, using GT word-level boundaries from AV-Deepfake1M annotations.

Deletion. One word segment is removed from $\tilde{\mathbf{x}}$ without replacement. The resulting temporal shift displaces all subsequent audio, disrupting the alignment of every payload repetition following the deletion point.

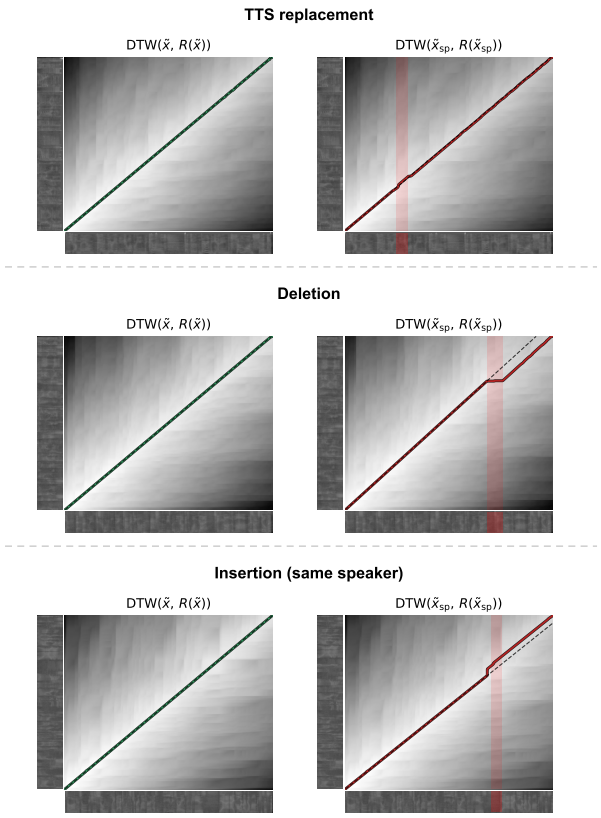

Fig. 2. DTW cost matrices and optimal warping paths for each manipulation type. **Left:** alignment between the intact watermarked signal $\tilde{\mathbf{x}}$ and its self-reconstruction $R(\tilde{\mathbf{x}})$, showing a near-diagonal path and low cost. **Right:** alignment between the manipulated watermarked signal $\tilde{\mathbf{x}}_{\text{sp}}$ and its self-reconstruction $R(\tilde{\mathbf{x}}_{\text{sp}})$, with the manipulated region highlighted in red.

Insertion. A word segment from a donor recording, drawn from the same or a different speaker, is inserted at a gap between two consecutive words of $\tilde{\mathbf{x}}$. As for deletion, the induced temporal shift disrupts all subsequent payload repetitions.

All manipulations are applied at the waveform level via an overlap-add concatenation procedure ($6\,\text{ms}$ frame shift, $12\,\text{ms}$ frame length, two-frame smoothing buffer), following [22]. Active speech levels are normalized; no codec or vocoder processing is applied. As illustrated in Fig. 2, the four manipulation types produce qualitatively distinct DTW distortion patterns. In the absence of manipulation the warping path closely follows the diagonal, replacement manipulations produce sharp local deviations, while deletion and insertion induce a global temporal shift in the warping path following the manipulation point.

Dataset. Experiments are conducted on an evaluation set derived from the validation split of the AV-Deepfake1M benchmark [28], retaining the $1{,}480$ utterances for which authentic recordings and Whisper ASR [29] word-level transcriptions are available. The validation split is used since GT labels and word-level metadata are publicly available only for training and validation partitions. Donor recordings for insertion are drawn from the VoxCeleb2 development and test splits [30].

TABLE II

UTTERANCE-LEVEL DETECTION EER (%) ↓ AND FRAME-LEVEL LOCALIZATION AUC↑ FOR REPLACEMENT AND STRUCTURAL MANIPULATIONS. DEL., INS. SAME, AND INS. DIFF. DENOTE DELETION, SAME-SPEAKER INSERTION, AND DIFFERENT-SPEAKER INSERTION, RESPECTIVELY.

		Detection EER (%) ↓						Localization AUC ↑					
		Replacement		Structural				Replacement		Structural			
Codec	Method	Direct	TTS	Del.	Ins. same	Ins. diff.	Mean	Direct	TTS	Del.	Ins. same	Ins. diff.	Mean
SNAC	LSB-1	8.56	10.68	18.57	18.51	19.41	15.15	0.912	0.889	0.662	0.868	0.864	0.839
	LSB-2	8.80	10.61	18.17	18.39	19.05	15.00	0.912	0.889	0.662	0.868	0.864	0.839
	LSB-4	8.80	10.75	18.50	18.43	19.21	15.14	0.912	0.889	0.662	0.868	0.864	0.839
SemantiCodec	LSB-1	8.40	9.78	17.47	18.13	18.55	14.47	0.911	0.892	0.661	0.868	0.862	0.839
	LSB-2	8.80	9.96	17.91	18.04	19.20	14.78	0.912	0.891	0.665	0.869	0.863	0.840
	LSB-4	8.08	9.97	18.05	18.12	19.30	14.70	0.912	0.893	0.662	0.867	0.864	0.840
TAAE	LSB-1	28.43	29.11	36.20	33.50	33.32	32.11	0.871	0.841	0.614	0.802	0.799	0.785
	LSB-2	28.33	29.08	36.20	33.46	33.16	32.05	0.871	0.841	0.614	0.802	0.799	0.785
	LSB-4	28.36	29.04	36.16	33.46	33.20	32.05	0.871	0.841	0.614	0.802	0.799	0.785

IV. EVALUATION

A. Experimental Conditions

Neural codecs. SNAC, SemantiCodec, and TAAE are used at $0.98\,\mathrm{kbps}$, $0.65\,\mathrm{kbps}$, and $0.4\,\mathrm{kbps}$, respectively (see Table I). For SNAC, signals are resampled to $24\,\mathrm{kHz}$ for codec encoding and decoding only, while watermarking embedding and all subsequent evaluation are performed at $16\,\mathrm{kHz}$.

LSB. LSB embedding is evaluated for $k \in \{1, 2, 4\}$ bits per sample, with a 64-bit synchronization preamble prepended to each payload repetition.

DTW-based scoring. Detection and localization scores are derived from DTW alignment between the received signal and its self-reconstruction, computed on 40-band log-mel spectrograms with a $16\,\mathrm{ms}$ frame shift. Cosine distance is used as the local dissimilarity measure, and the per-frame cost profile is smoothed prior to localization evaluation.

B. Evaluation Methodology

We evaluate utterance-level detection using the equal error rate (EER). The detection score is computed as the mean of the top 1% of smoothed per-frame DTW cosine costs between the received signal and its self-reconstruction, computed as $s_{\mathrm{DTW}}(\tilde{\mathbf{x}}, R(\tilde{\mathbf{x}}))$ for authentic signals and $s_{\mathrm{DTW}}(\tilde{\mathbf{x}}_{\mathrm{sp}}, R(\tilde{\mathbf{x}}_{\mathrm{sp}}))$ for manipulated signals. The decision threshold τ in (4) corresponds to the operating point at which the false positive rate equals the false negative rate. Lower EER values indicate better discriminative capacity between authentic and manipulated signals.

For frame-level localization, performance is measured by the Area Under the Receiver Operating Characteristic (ROC) curve (AUC). The AUC is computed between the smoothed per-frame DTW cost profile and binary GT frame labels derived from word-level boundaries provided by the dataset annotations. A perfect localization performance yields an AUC of 1.0, while an AUC of 0.5 corresponds to random guessing. Higher AUC values indicate better localization of manipulated regions.

C. Results

Imperceptibility. The watermarked signal $\tilde{\mathbf{x}}$ attains a wideband PESQ [31] of 4.64, the maximum attainable score, for all $k \in \{1, 2, 4\}$, confirming that the embedding remains perceptually transparent even at the highest bit depth.

Hash-based upper bound. As a reference, we evaluate a baseline which embeds a cryptographic hash of each local audio segment using the same LSB carrier method. Under ideal conditions, this baseline achieves near-perfect performance (detection EER near 0%, localization AUC above 0.999 across all manipulation types). This upper-bound performance is achieved because any localized modification results in a completely different decoded hash with high probability, making the detection of altered frames mathematically trivial. Such schemes, however, authenticate content without describing it, and thus cannot recover the original speech.

Detection and Localization. Across every codec, bit depth, and manipulation type, the embedded payload is recovered without bit errors after majority voting. The reported results therefore reflect the discriminability of the mismatch score rather than payload recovery failures. Table II reports utterance-level EER and frame-level AUC for all codec–bit-depth combinations. As anticipated, perfect accuracy is not achieved by any configuration, confirming that the codec-based approach trades the exactness of hash-based verification for recovery capability. SemantiCodec achieves the lowest mean EER, ranging from 14.47% to 14.78%, closely followed by SNAC at 15.00% to 15.15%. TAAE yields a considerably higher mean EER at 32.05% to 32.11%. Across all codecs, direct replacement is the most detectable manipulation, while deletion is consistently the hardest as it removes content without introducing external material. Detection also improves with manipulation duration, as shorter manipulations produce mismatch scores overlapping with the authentic distribution.

The localization results follow the same trend: SNAC and SemantiCodec achieve closely comparable mean AUC of 0.839–0.840 across all bit depths, while TAAE yields the lowest score of 0.785. Deletion is again the hardest manipulation to localize as removing content compresses the DTW path rather than producing a high-cost region. Direct and TTS replacement yield the highest AUC, between 0.841 and 0.912.

Recovery quality. Unlike classical hash-based schemes, which leave tampered regions completely unrecoverable, the

proposed framework successfully reconstructs the authentic speech within manipulated regions at codec-level fidelity, with zero payload bit errors. The receiver reconstructs $R(\tilde{x}_{sp}) = g(D(\tilde{x}_{sp}))$, successfully restoring the original spoken content. The reconstructions attain a wideband PESQ of 1.75, 1.63, and 1.65 for SNAC, SemantiCodec, and TAAE, respectively, on a scale from 1.0 to 4.64 where higher values indicate better perceptual quality. While these moderate PESQ values reflect the sample-level waveform deviations inherent to ultra-low-bitrate neural compression, these values do not compromise the usability of the restored signal as the reconstructed speech remains fully intelligible, natural, and speaker-consistent. Importantly, our analysis shows that reconstruction quality and detection performance are decoupled: SemantiCodec yields the lowest reconstruction PESQ but the lowest detection EER. Conversely, TAAE's transformer decoder prioritizes generating highly natural-sounding acoustic variations rather than maintaining sample-faithful alignment with the original source. This generative behavior increases the baseline mismatch score on authentic signals, which ultimately degrades the statistical separation required for accurate detection and localization.

V. Conclusion

We presented training-free self-embedding audio watermarking framework with ultra-low-bitrate neural codecs, in which a compact codec representation of the carrier is embedded into the signal itself. Across all LSB bit depth–codec configurations, the embedded payload is recovered without bit errors, so that the authentic content of manipulated regions can always be reconstructed at codec fidelity, with intelligible and natural-sounding recovered speech, which cannot be offered by hash-based verification. Detection and localization, obtained from the mismatch between the received signal and its self-reconstruction, remain imperfect even under ideal conditions, with deletion posing the greatest challenge. Our results indicate that performance is driven primarily by the codec's reconstruction characteristics, with LSB bit depth playing only a limited role. Future work includes semi-fragile embedding for robustness under channel distortions and further payload compression to increase embedding redundancy.

References

[1] J. He, J. Yi, J. Tao, S. Zeng, and H. Gu, "Manipulated regions localization for partially deepfake audio: A survey," arXiv, 2025.

[2] L. Zhang, X. Wang, E. Cooper, N. W. D. Evans, and J. Yamagishi, "The PartialSpoof database and countermeasures for the detection of short fake speech segments embedded in an utterance," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, pp. 813–825, 2023.

[3] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," *IBM Syst. J.*, vol. 35, no. 3–4, pp. 313–336, 1996.

[4] M. Steinebach and J. Dittmann, "Watermarking-based digital audio data authentication," *EURASIP J. Appl. Signal Process.*, vol. 2003, no. 10, pp. 1001–1015, 2003.

[5] R. S. Roman, P. Fernandez, H. Elsahar, A. Défossez, T. Furon, and T. Tran, "Proactive detection of voice cloning with localized watermarking," in *Proc. ICML*, 2024, pp. 43 180–43 196.

[6] C. Liu, J. Zhang, T. Zhang, X. Yang, W. Zhang, and N. Yu, "Detecting voice cloning attacks via Timbre Watermarking," in *Proc. NDSS*, 2024.

[7] G. Hua, J. Huang, Y. Q. Shi, J. Goh, and V. L. L. Thing, "Twenty years of digital audio watermarking: A comprehensive review," *Signal Process.*, vol. 128, pp. 222–242, 2016.

[8] D. Renza, D. M. B. L., and C. Lemus, "Authenticity verification of audio signals based on fragile watermarking for audio forensics," *Expert Syst. Appl.*, vol. 91, pp. 211–222, 2018.

[9] D. Yoon and T. Toda, "Neural semi-fragile watermarking for proactive deepfake speech detection," in *Proc. APSIPA ASC*, 2025, pp. 2092–2097.

[10] J. Fridrich and M. Goljan, "Images with self-correcting capabilities," in *Proc. ICIP*, 1999, pp. 792–796.

[11] P. Yogarajah, J. V. Condell, K. Curran, and P. McKevitt, "Video authentication: A self-embedding steganography approach," in *Proc. IMVIP*, 2011, pp. 174–189.

[12] L. Rakhmawati, S. Suwadi, and W. Wirawan, "Blind robust and self-embedding fragile image watermarking for image authentication and copyright protection with recovery capability," *Int. J. Intell. Eng. Syst.*, vol. 13, no. 5, 2020.

[13] D. Singh, S. K. Singh, and S. S. Udmale, "An efficient self-embedding fragile watermarking scheme for image authentication with two chances for recovery capability," *Multimedia Tools Appl.*, vol. 82, no. 1, pp. 1045–1066, 2023.

[14] S. Sarreshtedari, M. A. Akhaee, and A. Abbasfar, "A watermarking method for digital speech self-recovery," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 11, pp. 1917–1925, 2015.

[15] A. Menendez-Ortiz, C. Feregrino-Uribe, and J. J. Garcia-Hernandez, "Self-recovery scheme for audio restoration using auditory masking," *PLOS ONE*, vol. 13, no. 9, pp. 1–23, 2018.

[16] J. J. Gomez-Ricardez and J. J. Garcia-Hernandez, "A low distortion audio self-recovery algorithm robust to discordant size content replacement attack," *Computers*, vol. 10, no. 7, p. 87, 2021.

[17] M. Nishiguchi, K. Iijima, A. Inoue, Y. Maeda, and J. Matsumoto, "Harmonic vector excitation coding of speech at 2.0–4.0 kbps," in *Proc. ICCE*, 1998, pp. 208–209.

[18] P. Mousavi, G. Maimon, A. Moumen, D. Petermann, J. Shi, H. Wu, H. Yang, A. Kuznetsova, A. Ploujnikov, R. Marxer, B. Ramabhadran, B. Elizalde, L. Lugosch, J. Li, C. Subakan, P. Woodland, M. Kim, H. yi Lee, S. Watanabe, Y. Adi, and M. Ravanelli, "Discrete audio tokens: More than a survey!" *Trans. Mach. Learn. Res.*, pp. 1–54, 2025.

[19] H. Siuzdak, F. Grötschla, and L. A. Lanzendörfer, "SNAC: Multi-scale neural audio codec," in *Proc. Audio Imagination Workshop at NeurIPS*, 2024.

[20] H. Liu, X. Xu, Y. Yuan, M. Wu, W. Wang, and M. D. Plumbley, "SemantiCodec: An ultra low bitrate semantic audio codec for general sound," *IEEE J. Sel. Topics Signal Process.*, vol. 18, no. 8, pp. 1448–1461, 2024.

[21] J. D. Parker, A. Smirnov, J. Pons, C. Carr, Z. Zukowski, Z. Evans, and X. Liu, "Scaling transformers for low-bitrate high-quality speech coding," in *Proc. ICLR*, 2025.

[22] Y. Özer, Z. Zhang, W. Ge, X. Wang, and J. Yamagishi, "A training-free proactive defense against partial speech manipulation via self-embedding steganography," 2026, accepted to Proc. Interspeech.

[23] M. Müller, "Dynamic time warping," in *Information Retrieval for Music and Motion*. Springer, 2007, pp. 69–84.

[24] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," in *Proc. ICML*, 2023, pp. 21 450–21 474.

[25] F. Mentzer, D. C. Minnen, E. Agustsson, and M. Tschannen, "Finite scalar quantization: VQ-VAE made simple," in *Proc. ICLR*, 2024.

[26] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. ICML*, 2021, pp. 5530–5540.

[27] E. Casanova, J. Weber, C. D. Shulby, A. C. Júnior, E. Gölge, and M. A. Ponti, "YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," in *Proc. ICML*, 2022, pp. 2709–2720.

[28] Z. Cai, S. Ghosh, A. P. Adatia, M. Hayat, A. Dhall, T. Gedeon, and K. Stefanov, "AV-Deepfake1M: A large-scale LLM-driven audio-visual deepfake dataset," in *Proc. ACMMM*, 2024, pp. 7414–7423.

[29] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023, pp. 28 492–28 518.

[30] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[31] ITU-T, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," International Telecommunication Union, Tech. Rep. P.862, 2001.