

# SemBind: Inspectable Self-Supervised and Spectral Placement for Reference-Based Speech Watermarking

Sanjeev Kumar, Aditya Kumar Sahu, Banashree Mandal, Biswajit Maity

**Abstract**—Speech watermarking schemes decide where to embed from signal energy or leave the decision implicit inside a trained network, so the effect of placement on speech content is rarely measured. This letter is a controlled study of that effect, built on SemBind, a reference-based STFT watermark whose placement is explicit: a frozen HuBERT model gives a closed-form leave-one-out contribution map from one forward pass, which is fused with spectral suitability to place a 256-bit, three-replica, SHA-256-bound payload. Holding the embedding rule fixed, we compare placement rules across a strength sweep on LibriSpeech. On the full 2,620-utterance LibriSpeech test-clean corpus, a six-level sweep from 87 to 20 dB evaluates five placement rules, including random placement. At target 40 dB, joint and spectral-only placement change 8.85% and 8.28% of Whisper tiny.en transcripts, compared with 18.36% for high-contribution and 16.98% for random placement. At 30 dB, the corresponding rates are 21.18%, 19.12%, 30.46%, and 29.69%. Low-contribution placement produces fewer changes than its high-contribution control from 61 to 20 dB, showing that the HuBERT map is directionally informative; however, the current joint fusion does not consistently exceed spectral-only placement.

**Index Terms**—Speech watermarking, embedding-location selection, self-supervised learning, HuBERT, content preservation, provenance.

## I. INTRODUCTION

**W**ATERMARKING supports ownership identification and provenance of speech recordings [1]. Besides the payload and the extractor, a scheme must decide *where* to embed. Classical spread-spectrum, echo, and patchwork schemes choose locations from energy, transform statistics, masking, or fixed bands [2]–[5]. Neural schemes [6]–[11] learn the embedder and detector jointly, and recent letters continue with adversarial, generative, and reinforcement-learned designs [12]–[14]. In all of these, placement cannot be varied independently of encoding, so its effect on content is rarely isolated.

Self-supervised models such as HuBERT [15] and wav2vec 2.0 [16] encode phonetic and speaker content per frame, which suggests a content-aware alternative to energy-based placement: embed where a frame contributes least to the utterance representation. This letter tests that hypothesis directly; to our knowledge, it is the first study of speech watermarking that varies placement alone while holding the payload, embedding rule, and extractor fixed.

The authors are with the Department of Computer Science and Engineering, SRM University-AP, Amaravati, India (e-mail: sanjeevkumar@srmmap.edu.in; biswajit.maity@srmmap.edu.in; aditya.sahu@srmmap.edu.in).

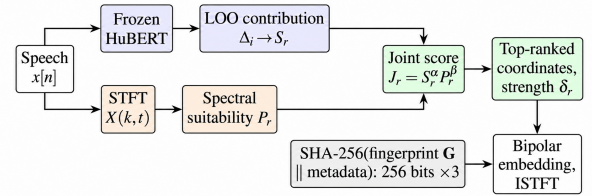


Fig. 1. SemBind embedding. HuBERT only produces the contribution map and the fingerprint; embedding is in the STFT domain.

The contributions are: (i) SemBind, a reference-based watermark with an explicit placement stage, built on a closed-form leave-one-out contribution map from one HuBERT pass; (ii) a controlled comparison in which only placement varies, from 87 to 20 dB SNR, with classical and neural baselines; and (iii) the finding, explained by a local-SNR analysis, that energy-aware placement, not self-supervised contribution, governs transcript damage, i.e., that representational redundancy alone does not identify safe locations. Blind extraction and robustness are out of scope.

## II. PROPOSED METHOD

### A. Representations

Let  $X(k, t)$  be the STFT of the 16-kHz input  $x[n] \in [-1, 1]$  (periodic Hann window,  $N = 1024$ , hop  $H = 320$ , centered reflective padding, no normalization, 513 one-sided bins). The frozen HuBERT-base model [15] (hubert-base-ls960, last layer) gives frame representations  $\mathbf{E}_1, \dots, \mathbf{E}_T \in \mathbb{R}^{768}$  with mean  $\mathbf{G}$ . Removing frame  $i$  gives, in closed form, the contribution  $\Delta_i$  and representation suitability  $S_i$ :

$$\Delta_i = 1 - \cos\left(\mathbf{G}, \frac{T\mathbf{G} - \mathbf{E}_i}{T - 1}\right), \quad S_i = 1 - \min_{j \neq i} \cos(\Delta_j). \quad (1)$$

This needs one forward pass and  $O(Td)$  operations. A frame with small  $\Delta_i$  (e.g., a pause) is redundant for the utterance summary and is expected to be safer to perturb; high-contribution placement is the negative control.

### B. Spectral Suitability and Joint Placement

HuBERT frame  $r$  (receptive field centered at sample  $rH + 200$ ) maps to the single centered STFT frame  $t = r + 1$ ,

**Algorithm 1** SemBind embedding

---

**Require:** clean speech  $x$ ; owner  $o$ , timestamp  $\tau$ , key  $\kappa$   
**Ensure:** watermarked speech  $x_w$

- 1:  $X \leftarrow \text{STFT}(x)$ ;  $X' \leftarrow X$ ;  $\mathbf{E}_{1:T} \leftarrow \text{HuBERT}(x)$
- 2:  $S_r, P_r, J_r, \delta_r \leftarrow (1)-(3)$
- 3:  $\mathbf{b} \leftarrow \text{SHA256}(\mathbf{G} \parallel o \parallel \tau)$
- 4:  $\mathcal{C} \leftarrow \text{ALLOCATE}(J, \kappa)$   $\triangleright m$  bins per top region
- 5: **for**  $i = 0, \dots, 255$  **and**  $q = 0, \dots, R-1$  **do**
- 6:     update  $X'(c_{i,q})$ ,  $c_{i,q} = \mathcal{C}_{i+256q}$ , by (4)
- 7: **end for**
- 8: **return**  $x_w \leftarrow \text{ISTFT}(X')$

---

since both use a 20-ms stride. With admissible bins  $\mathcal{F} = \{1, \dots, 511\}$  (DC and Nyquist excluded),

$$P_r = \min_{\mathcal{F}} \left( \frac{1}{|\mathcal{F}|} \sum_{k \in \mathcal{F}} \log(1 + |X(k, r+1)|) \right) \quad (2)$$

measures local signal headroom (not a psychoacoustic model). Placement and strength follow from the joint score

$$J_r = S_r^\alpha P_r^\beta, \quad \delta_r = \delta_{\min} + (\delta_{\max} - \delta_{\min})J_r, \quad (3)$$

where  $\alpha = 0.7$  and  $\beta = 0.3$  weight the two criteria and  $\delta_{\min} = 0.002$ ,  $\delta_{\max} = 0.008$  (unnormalized STFT units). Spectral-only, representation-only, and high-contribution placement replace  $J_r$  by  $P_r$ ,  $S_r$ , and  $1 - S_r$ .

*Allocation with replica separation.* The top  $\lceil \rho T \rceil$  regions by  $J_r$  ( $\rho = 0.5$  is the selected fraction) each contribute  $m = \lceil 256R / \lceil \rho T \rceil \rceil$  distinct bins drawn from  $\mathcal{F}$  by a keyed pseudo-random generator, appended in ranking order to a list  $\mathcal{C}$  of  $256R$  coordinates ( $R = 3$ ). Replica  $q$  of bit  $i$  takes entry  $i + 256q$ . Because each region occupies  $m$  consecutive entries and replicas are 256 entries apart, the three replicas of a bit lie in distinct regions whenever  $m \leq 256$ , i.e.,  $\lceil \rho T \rceil \geq 3$ , which holds for any utterance longer than about 0.12 s.

### C. Payload, Embedding, and Verification

The payload  $\mathbf{b} \in \{0, 1\}^{256}$  is the SHA-256 digest [17] of a canonical JSON object holding  $\mathbf{G}$  (four decimals), an owner identifier  $o$ , and a UTC timestamp  $\tau$ . Each coordinate  $c \in \mathcal{C}$  carrying bit  $b_i$  in region  $r$  is modified, and each replica is read back against the reference, as

$$\text{Re } X'(c) = \text{Re } X(c) + (2b_i - 1)\delta_r, \quad (4)$$

$$\hat{b}_{i,q} = \mathbf{1}[\text{Re } \tilde{X}(c_{i,q}) > \text{Re } X(c_{i,q})], \quad (5)$$

so bit 1 raises and bit 0 lowers the coefficient. Verification (Algorithm 2) needs the clean reference  $x$  and the trusted metadata  $(o, \tau, \kappa)$  from the owner's provenance record: it recomputes  $\mathcal{C}$  and the expected  $\mathbf{b}$  from  $x$ , so no coordinates are stored. SHA-256 binds content and metadata for integrity; it does not by itself prove who embedded the mark, which would require a keyed MAC or a signature. The default configuration uses the adaptive  $\delta_r$  of (3); the controlled experiments fix  $\delta_r \equiv \delta$  to isolate placement. Selection and embedding take 52 ms per utterance on one CPU core beyond one HuBERT pass (real-time factor 0.007).

**Algorithm 2** SemBind reference-based verification

---

**Require:** observed  $\tilde{x}$ ; clean reference  $x$ ; trusted  $(o, \tau, \kappa)$   
**Ensure:** ACCEPT or REJECT

- 1: recompute  $\mathbf{G}$ ,  $J_r$ , and  $\mathcal{C} \leftarrow \text{ALLOCATE}(J, \kappa)$  from  $x$
- 2:  $\mathbf{b} \leftarrow \text{SHA256}(\mathbf{G} \parallel o \parallel \tau)$   $\triangleright$  expected payload
- 3:  $\tilde{X} \leftarrow \text{STFT}(\tilde{x})$ ;  $X \leftarrow \text{STFT}(x)$
- 4: **for**  $i = 0, \dots, 255$  **do**
- 5:      $\hat{b}_i \leftarrow \text{majority}_q \hat{b}_{i,q}$  by (5)
- 6: **end for**
- 7: **return** ACCEPT if  $\hat{\mathbf{b}} = \mathbf{b}$ , else REJECT

---

TABLE I  
RECOVERY AND FIDELITY ON 200 UTTERANCES. STFT ROWS USE IDENTICAL FIXED STRENGTH ( $\delta = 0.005$ ).

| Method                      | Recovery | SNR (dB) $\uparrow$ | PESQ $\uparrow$ | STOI $\uparrow$ |
|-----------------------------|----------|---------------------|-----------------|-----------------|
| SemBind (joint)             | 100%     | 86.99               | 4.64            | 1.0000          |
| Spectral-only               | 100%     | 87.00               | 4.64            | 1.0000          |
| Low contribution            | 100%     | 87.14               | 4.64            | 1.0000          |
| High contribution           | 100%     | 87.05               | 4.64            | 1.0000          |
| Random                      | 100%     | 87.07               | 4.64            | 1.0000          |
| DWT-SS [2]                  | 93%      | 86.99               | 4.64            | 0.9999          |
| AudioSeal <sup>a</sup> [10] | 100%     | 27.13               | 4.50            | 0.9980          |

<sup>a</sup>Blind detector, 16-bit message (accuracy 100%), native strength.

## III. EXPERIMENTAL SETUP

LibriSpeech test-clean [18] contains 2,620 utterances from 40 speakers. The primary placement sweep evaluates all 2,620 utterances under five deterministic rules: joint ( $J_r$ ), spectral-only ( $P_r$ ), low contribution ( $S_r$ ), high contribution ( $1 - S_r$ , a negative control), and random placement. For each rule, the embedding strength is fixed within a condition and is swept over target SNRs of 87, 61, 40, 30, 27, and 20 dB. Content damage is measured by the number and percentage of utterances whose Whisper tiny.en [19] transcript differs from the clean reference, together with mean  $\Delta\text{WER}$ . Decoding is greedy, with no temperature fallback and no previous-text conditioning. The full-corpus strength sweep measures content damage and waveform SNR; it does not re-evaluate payload recovery at every strength. The fixed-strength recovery and baseline comparisons use the deterministic 200-utterance subset. Baselines are a reference-based DWT spread-spectrum scheme after Cox et al. [2] with the same payload, three replicas, and SNR, and AudioSeal [10] at its native operating point.

## IV. RESULTS

### A. Recovery and Fidelity

At 87 dB every STFT rule recovers the full payload in every utterance, whereas DWT-SS with the same payload, replicas, and SNR fails on 14 of 200, so in this matched setting coefficient-level bipolar embedding is the more reliable carrier. PESQ and STOI are at their ceilings, and none of the four swept rules changes a single transcript (Fig. ??), so at transparent strength placement has no measurable content cost.

Table II positions SemBind among the baselines rather than ranking them. AudioSeal provides blind, processing-robust detection of a 16-bit message at 27 dB; SemBind verifies a 256-bit payload at 87 dB only when the reference is available. The

TABLE II  
OPERATING POINTS OF THE COMPARED SCHEMES (200 UTTERANCES).  
PAYLOAD, EXTRACTION MODE, AND ROBUSTNESS DIFFER, SO THE ROWS  
ARE NOT A RANKING.

| Scheme         | Placement  | Extraction | Payload | SNR (dB) | Recovery |
|----------------|------------|------------|---------|----------|----------|
| DWT-SS [2]     | fixed band | reference  | 256 b   | 86.99    | 93%      |
| AudioSeal [10] | learned    | blind      | 16 b    | 27.13    | 100%     |
| SemBind        | explicit   | reference  | 256 b   | 86.99    | 100%     |

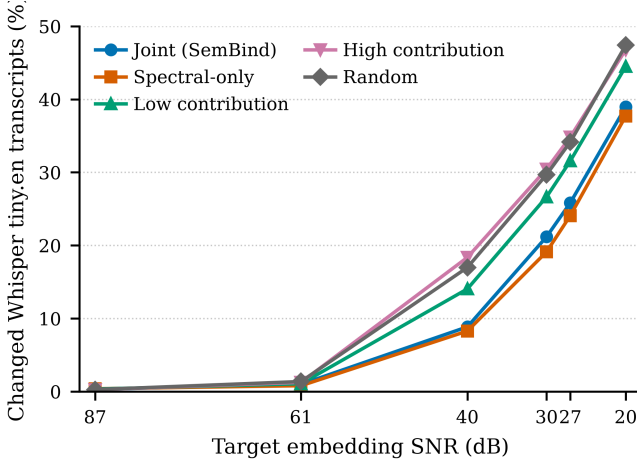


Fig. 2. Full-corpus strength sweep on 2,620 utterances per rule. Transcript-change rate under Whisper tiny.en is shown for all five placement rules. Lower target SNR corresponds to stronger embedding.

SNR gap reflects these different goals, not overall superiority; SemBind’s distinguishing property is that its placement can be read and varied directly.

### B. When Does Placement Matter?

Across the full corpus, transcript changes are low but nonzero at the 87 dB target (7–10 of 2,620 utterances, depending on placement) and remain below 1.4% at 61 dB. Placement effects become substantial from 40 dB downward. At 40 dB, joint and spectral-only placement change 8.85% and 8.28% of transcripts, respectively, compared with 18.36% for high-contribution and 16.98% for random placement. At 30 dB, the corresponding rates are 21.18%, 19.12%, 30.46%, and 29.69%.

Spectral-only placement has fewer changed transcripts than joint placement at five of the six tested strengths. Thus, the present multiplicative fusion should not be described as consistently improving on spectral-only placement. Its clearest benefit is relative to unguided random placement at 61–20 dB, where joint placement produces substantially fewer transcript changes (Fig. 2).

HuBERT contribution remains informative as a directional diagnostic: low-contribution placement yields fewer changed transcripts than high-contribution placement at each target from 61 to 20 dB (28 vs. 33, 368 vs. 481, 698 vs. 798, 827 vs. 913, and 1166 vs. 1226). This result shows that the representation-derived contribution map separates relatively safer from riskier frames under stronger embedding. However,

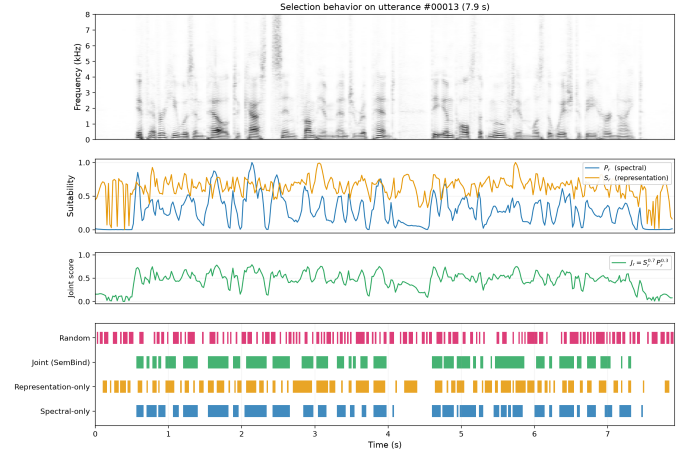


Fig. 3. Placement on one utterance (7.9 s): spectrogram;  $P_r$ ,  $S_r$ , and  $J_r$ ; and the frames used by each rule. Low-contribution (representation-only) selection fills the pauses; joint and spectral-only selection avoid them because  $P_r \approx 0$  there.

spectral suitability is the stronger operational criterion, and the current joint fusion does not consistently outperform spectral-only placement.

### C. Why Spectral Placement Wins

The explicit placement stage explains the sweep. Low-contribution placement fills the leading and trailing pauses (Fig. 3, 0–0.5 s and after 7.4 s), as the redundancy argument predicts: pauses contribute least to  $\mathbf{G}$ , so  $S_r$  peaks there. But for  $m_r$  coordinates of strength  $\delta$  in frame  $t = r+1$ , the local signal-to-watermark ratio is

$$\text{SNR}_r = 10 \log_{10} \frac{\sum_{k \in \mathcal{F}} |X(k, t)|^2}{m_r \delta^2}, \quad (6)$$

which falls decibel for decibel with frame energy at fixed  $\delta$ . This first-order expression assumes a fixed  $\delta$  and only the frame’s own energy over  $\mathcal{F}$ ; it ignores overlap-add spreading into adjacent frames and auditory masking, so it ranks frames rather than predicting ASR errors. A pause therefore receives the most exposed perturbation in the utterance, consistent with the higher counts of low-contribution placement. Because  $P_r$  in (2) grows with log-magnitude, ranking by  $P_r$  approximately ranks frames by  $\text{SNR}_r$ : spectral placement is, in effect, local-SNR maximization. It excludes pauses ( $P_r \approx 0$ ), and because  $J_r = S_r^\alpha P_r^\beta$  is multiplicative, joint placement excludes them too and, as Fig. 3 shows, covers nearly the same voiced frames as spectral-only placement. Frames redundant for the representation are thus not perceptually safe; local energy is what masks the watermark from the recognizer.

### D. Ablations and Operational Boundary

**Ablations** (87 dB): fusion weights from (0, 1) to (0.75, 0.25) and adaptive versus fixed strength change SNR by at most 0.95 dB (86.05–87.00 dB);  $R = 1$  and  $R = 2$  gain 4.76 and 1.75 dB but fail exact recovery on 1 and 4 of 200 utterances, so  $R = 3$  is kept.

TABLE III  
BIT ACCURACY (%) AFTER PROCESSING (200 UTTERANCES, 87 DB).

|          | AWGN (dB) |      |      |      | Scale | Resamp. | LPF  | MP3  | PCM   |
|----------|-----------|------|------|------|-------|---------|------|------|-------|
|          | 40        | 30   | 20   | 10   |       |         |      |      |       |
| Bit acc. | 57.6      | 52.4 | 50.9 | 50.4 | 65.5  | 81.5    | 77.1 | 61.9 | 100.0 |

The watermark survives storage as 16-bit PCM with 100% bit accuracy, so it persists in standard WAV files. Because each bit is the sign of a small displacement relative to the reference, additive noise, gain change, resampling, filtering, and compression lower bit accuracy to 50.4–81.5% (Table III), far from the exact match that verification requires; surviving such processing requires robust designs [20], [21]. SemBind therefore targets provenance verification of unmodified speech when the reference is available.

## V. DISCUSSION

*Implications.* The results suggest three guidelines for content-aware placement. First, any content score should be gated by a local-energy term, as (6) shows that a fixed perturbation is most exposed where content is sparse; the multiplicative fusion of SemBind does this automatically. Second, placement should be compared at strengths where the recognizer responds (here  $\leq 40$  dB), since at transparent strength every rule is indistinguishable. Third, the number of changed transcripts is a more sensitive and more robust content measure than mean  $\Delta$ WER, whose median was zero at every point of the sweep.

*Limitations.* The full-corpus placement sweep covers 2,620 utterances of read English speech from 40 speakers, but only one completed ASR model, Whisper tiny.en, is available. PESQ and STOI are ceiling-level in the clean-channel setting and therefore do not establish perceptual differences between placement rules. A completed second-ASR evaluation or a non-saturated perceptual measure is required before claiming model-independent content preservation. Extraction requires the reference, so SemBind addresses provenance verification of unmodified speech rather than detection after processing.

## VI. CONCLUSION

Making the placement stage of a reference-based speech watermark explicit allows its effect on content to be measured in isolation. On the full 2,620-utterance sweep, spectral-only and joint placement substantially reduce transcript changes relative to high-contribution and random placement at stronger embedding levels. The frozen HuBERT contribution map is directionally informative: low-contribution placement yields fewer changed transcripts than high-contribution placement from 61 to 20 dB. However, the current joint fusion does not consistently outperform spectral-only placement. Content-aware placement should therefore be energy-gated, while future work should test whether a revised semantic term can add value beyond local spectral suitability. Code and the selected coordinate lists will be released upon publication.

## REFERENCES

- [1] G. Hua, J. Huang, Y. Q. Shi, J. Goh, and V. L. L. Thing, "Twenty years of digital audio watermarking—a comprehensive review," *Signal Processing*, vol. 128, pp. 222–242, 2016.
- [2] I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamon, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. Image Process.*, vol. 6, no. 12, pp. 1673–1687, 1997.
- [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] B.-S. Ko, R. Nishimura, and Y. Suzuki, "Time-spread echo method for digital audio watermarking," *IEEE Trans. Multimedia*, vol. 7, no. 2, pp. 212–221, 2005.
- [5] I. Natgunanathan, Y. Xiang, Y. Rong, W. Zhou, and S. Guo, "Robust patchwork-based embedding and decoding scheme for digital audio watermarking," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 8, pp. 2230–2238, 2012.
- [6] K. Pavlović, S. Kovačević, I. Djurović, and A. Wojciechowski, "Robust speech watermarking by a jointly trained embedder and detector using a DNN," *Digit. Signal Process.*, vol. 122, p. 103381, 2022.
- [7] G. Chen, Y. Wu, S. Liu, T. Liu, X. Du, and F. Wei, "WavMark: Watermarking for audio generation," *arXiv:2308.12770*, 2023.
- [8] P. Li, X. Zhang, J. Xiao, and J. Wang, "IDEAW: Robust neural audio watermarking with invertible dual-embedding," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2024, pp. 4500–4511.
- [9] M. K. Singh, N. Takahashi, W.-H. Liao, and Y. Mitsufuji, "SilentCipher: Deep audio watermarking," in *Proc. Interspeech*, 2024, pp. 2235–2239.
- [10] R. San Roman, P. Fernandez, H. Elshahar, A. Défossez, T. Furon, and T. Tran, "Proactive detection of voice cloning with localized watermarking," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2024.
- [11] H. Li, Z. Wu, X. Xie, J. Xie, Y. Xu, and H. Peng, "VoiceMark: Zero-shot voice cloning-resistant watermarking approach leveraging speaker-specific latents," in *Proc. Interspeech*, 2025, pp. 5108–5112.
- [12] Y. He, H. Wang, Y. Qiu, and H. Cao, "ASSMark: Dual defense against speech synthesis attack via adversarial robust watermarking," *IEEE Signal Processing Letters*, vol. 32, pp. 1870–1874, 2025.
- [13] Y. Feng, X. Zhang, F. Feng, G. Zhang, and L. Xu, "Robust and imperceptible watermarking framework for generative audio models," *IEEE Signal Processing Letters*, vol. 32, pp. 3196–3200, 2025.
- [14] X. Fang, J. Li, J. Liu, and L. Zou, "Reinforced temporally sparse gating for high-fidelity speech watermarking," *IEEE Signal Processing Letters*, vol. 33, pp. 1936–1940, 2026.
- [15] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.
- [16] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 12 449–12 460.
- [17] National Institute of Standards and Technology, "Secure hash standard (SHS)," National Institute of Standards and Technology, Federal Information Processing Standards Publication FIPS PUB 180-4, 2015.
- [18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2015, pp. 5206–5210.
- [19] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLevey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023, pp. 28 492–28 518.
- [20] H. Liu, M. Guo, Z. Jiang, L. Wang, and N. Z. Gong, "AudioMarkBench: Benchmarking robustness of audio watermarking," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 37, 2024.
- [21] Y. Özer, W. Choi, J. Serrà, M. K. Singh, W.-H. Liao, and Y. Mitsufuji, "A comprehensive real-world assessment of audio watermarking algorithms: Will they survive neural codecs?" in *Proc. Interspeech*, 2025, pp. 5113–5117.