2026/09/29, 14:00–16:00
Spoofing and Deepfake Detection 2

Interspeech 2026

A Training-Free Proactive Defense Against Partial Speech Manipulation via Self-Embedding Steganography

Yigitcan Özer, Zhe Zhang, Wanying Ge, Xin Wang, Junichi Yamagishi
National Institute of Informatics, Tokyo
yiitozer@nii.ac.jp

Motivation & Key Idea

Partial speech manipulation
Only short segments of an authentic utterance are replaced (e.g., single words): the speaker identity is kept, the linguistic content changes.

Passive detectors struggle
They rely on synthesis artifacts that fade as the manipulated region shrinks, thus word-level manipulations are detected at near-random rates.

Our proposal: a proactive, training-free defense
Prior to distribution, the clean speech carries a compressed representation of *itself* (self-embedding steganography). The receiver reconstructs this copy and compares it with what was received: any discrepancy exposes the manipulation, and the low-bitrate neural codec restores the original content.

Self-Embedding Steganography

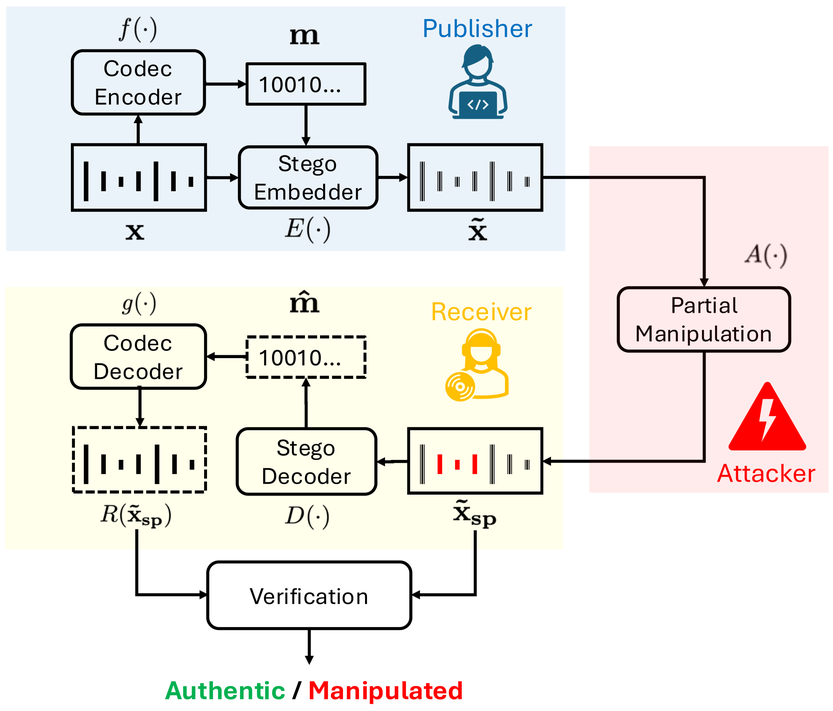


Fig. 1. Self-embedding framework: The SNAC code [4] of $\boldsymbol{x}$ is hidden in itself; the receiver decodes it into $R(\boldsymbol{y})$ and compares it with the received $\boldsymbol{y}$.

Temporally repetitive LSB embedding

- Bits $\boldsymbol{m} \in \{0,1\}^M$ are written into the least-significant bit (LSB) of the samples, repeating every $P \geq M$ samples; the decoder majority-votes over all R copies:

$$R = 1 + \left\lfloor \frac{T - M}{P} \right\rfloor$$

$$\hat{m}_i = 1 \text{ if } \frac{1}{R}\sum_{r=1}^{R} m_i^{(r)} \geq 0.5, \quad \text{else } 0$$



Capacity at 16 kHz: 1 s = 16,000 LSB positions = 15 frames of ■ 64-bit sync + ■ 980-bit SNAC payload [4] (0.98 kbps) ⇒ **100 % bit-exact recovery**, with and without attacks.

Detection Score via Dynamic Time Warping

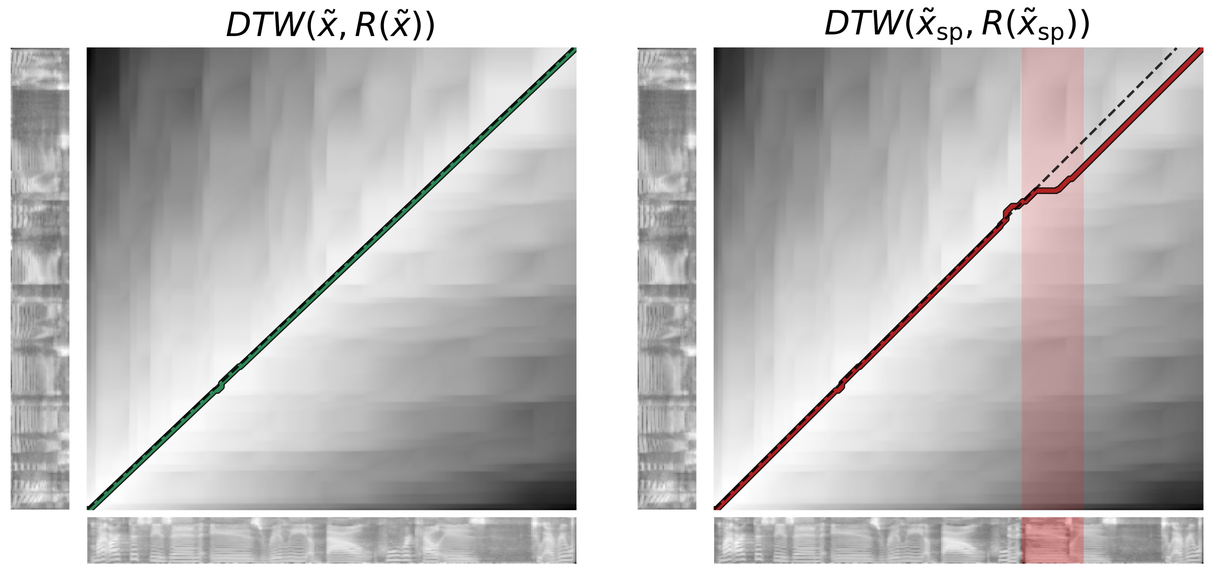


Fig. 2. Warping paths computed with DTW between the received signal and its self-reconstruction. **Left:** no manipulation. **Right:** after a single-word swap. The manipulated segment is highlighted in red.

- DTW [5] aligns the received signal $\boldsymbol{y}$ with its self-reconstruction $R(\boldsymbol{y})$ on normalized log-mel spectrograms using the cosine distance.
- **No manipulation:** the path stays on the diagonal and the accumulated cost stays low. **Manipulated:** the path deviates locally and the cost rises around the swapped region.
- Per-utterance score = mean local cost c_t along the optimal path π^*; higher s_{DTW} ⇒ more likely manipulated. The threshold τ is set at the EER operating point.

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

Experiments & Results

Setup

- **Data:** real recordings of the AV-Deepfake1M validation split [3].
- **Threat model:** one or two words are swapped for re-synthesized segments from other utterances of the same speaker. Re-synthesis methods: GriffinLim, HiFiGAN, HNSincNSF, HNSincNSFHiFi, WaveGlow.
- **Baselines** (pretrained, audio-only): LAV-DF [1], LAV-DF+ [2] and a ResNet word-level detector [6].
- **Metric:** EER (%), lower is better.

Table 1. EER (%) under the word-swapping attack, per re-synthesis model; baselines LAV-DF [1], LAV-DF+ [2], ResNet [6].

Re-synthesis	Ours	LAV-DF	LAV-DF+	ResNet
Single-word swapping				
GriffinLim	**9.95**	50.17	49.90	47.46
HiFiGAN	**9.38**	50.04	49.90	47.52
HNSincNSF	**9.52**	50.10	49.90	46.97
HNSincNSFHiFi	**9.52**	50.31	49.90	47.52
WaveGlow	**8.91**	50.37	50.03	43.98
Two-word swapping				
GriffinLim	**4.97**	50.07	48.34	45.89
HiFiGAN	**5.00**	50.03	48.56	46.44
HNSincNSF	**5.07**	49.96	48.34	45.67
HNSincNSFHiFi	**5.07**	50.03	48.34	45.45
WaveGlow	**4.44**	50.11	48.62	42.07

≈ 50 %
EER, passive detectors

4.4–10 %
EER, our proactive defense

Follow-up work

Audio samples

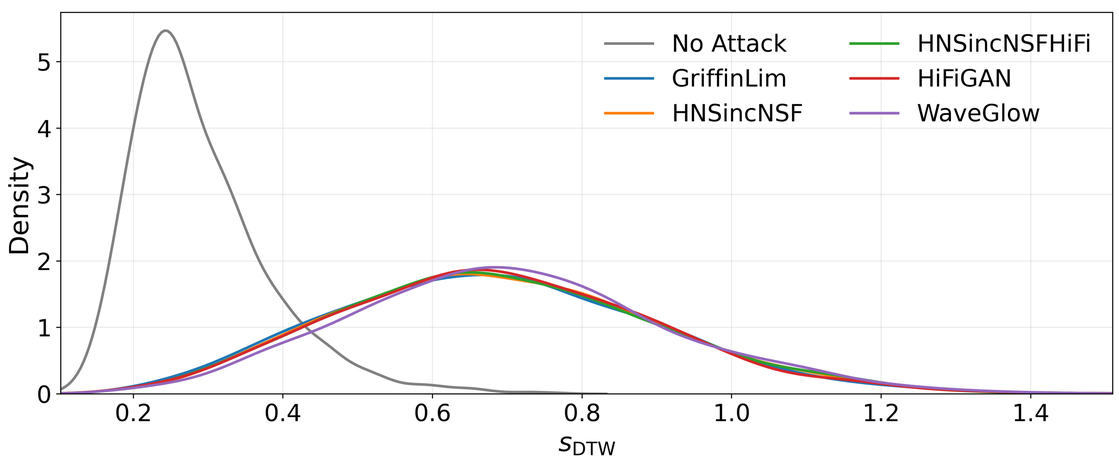


Fig. 3. Distributions of s_{DTW}: authentic (grey) vs. single-word swaps, per re-synthesis model.

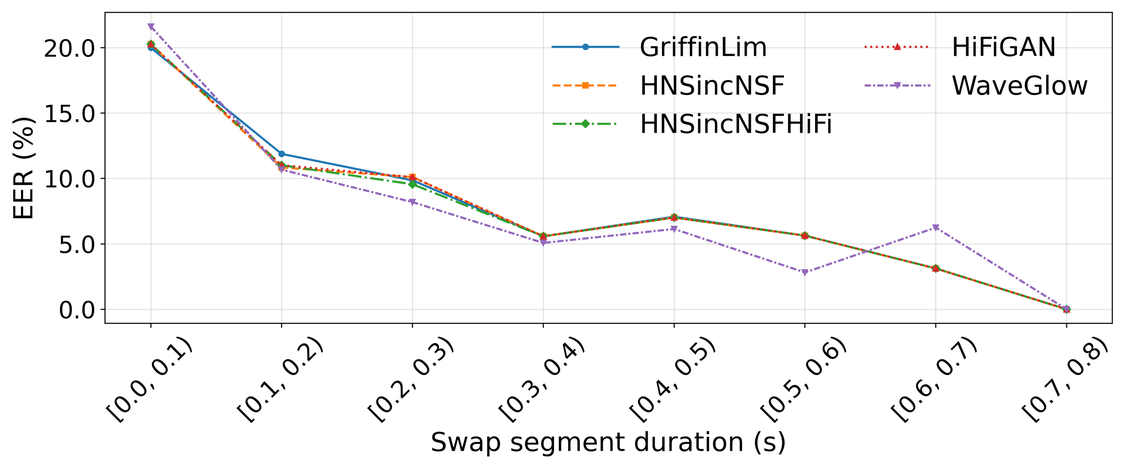


Fig. 4. EER as a function of the swapped-word duration (0.1 s bins), per re-synthesis model.

References [1] Z. Cai et al., "Do you really mean that? Content-driven audio-visual deepfake dataset and multimodal method for temporal forgery localization," *DICTA*, 2022. [2] Z. Cai et al., "Glitch in the matrix: A large-scale benchmark for content-driven audio-visual forgery detection and localization," *CVIU*, 2023. [3] Z. Cai et al., "AV-Deepfake1M: A large-scale LLM-driven audio-visual deepfake dataset," *ACM MM*, 2024. [4] H. Siuzdak et al., "SNAC: Multi-scale neural audio codec," *NeurIPS Workshop on AI-Driven Speech, Music, and Sound Generation*, 2024. [5] M. Müller, "Dynamic time warping," in *Information Retrieval for Music and Motion*, Springer, 2007. [6] H. M. Tran et al., "Deepfake word detection by next-token prediction using fine-tuned Whisper," *Interspeech*, 2026.

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