Feature Fusion for Indonesian Spoof Diarization Using Attractor Tokens and Cross Attention

Hibah P2MI ITB 2026

Dessi Puji Lestari[1]*, Naufal Adnan[1], Angelica Kierra Ninta Gurning[1]

[1] School of Electrical Engineering and Informatics (STEI), Institut Teknologi Bandung

* Corresponding author: dessipuji@staff.stei.itb.ac.id

Abstract

Speech forensics must not only decide whether an utterance is spoofed, but also locate manipulated segments and identify the spoofing method. This task, spoof diarization, has been studied mainly for English.

We introduce the first Indonesian spoof diarization dataset: 3,939 partially spoofed utterances built from Common Voice and Prosa.ai speech using seven TTS systems. Within an attractor-token and cross-attention framework, we combine self-supervised (SSL), LFCC, and pathological features through four fusion mechanisms.

Against the SSL baseline (JIbona 0.0214, JERspoof 0.2323), three-feature dual-encoder fusion gives the best JIbona (0.0205, 4.21% relative improvement) and three-feature gated fusion the best JERspoof (0.2100, 9.60%). Fusion effectiveness depends on both the feature combination and the mechanism.

Keywords: attractor tokens, cross attention, feature fusion, Indonesian speech, spoof diarization.

Introduction

Modern TTS and voice conversion produce speech that closely resembles human voices, which creates risks for speaker verification and forensics. A partially spoofed utterance may replace only a single word, yet still change its meaning, and the fake region can be masked by surrounding bona fide speech.

Spoof localization finds when spoofing occurs. Spoof diarization goes further and asks what spoofed when: it locates spoofed regions and groups them by spoofing method. The 3C model set the first baseline, and attractor tokens with cross-attention later improved it substantially on English data.

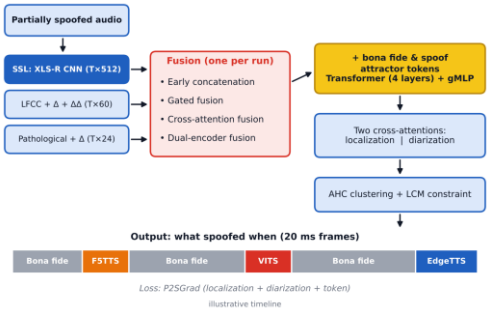
For Indonesian, prior work covered spoof detection (InaSAS) and localization with pathological features, but no diarization dataset or benchmark exists. The contribution of complementary features within the attractor-token framework has also not been studied systematically.

Contributions: (1) the first Indonesian spoof diarization dataset with 20 ms labels and seven TTS attacks; (2) integration of SSL, LFCC, and pathological features with four fusion mechanisms; (3) an analysis of how feature complementarity interacts with the fusion mechanism.

Research Method

Dataset. Bona fide speech from Common Voice and Prosa.ai; fully spoofed speech from Google TTS, Microsoft Edge TTS, F5TTS, Chatterbox, Facebook MMS, VITS, and Piper (plus FreeVC). Words are replaced using Whisper timestamps and fuzzy matching, with cross-fading; one utterance may mix several attacks. 3,939 partial spoofs, each paired with a bona fide utterance, labelled at 20 ms resolution.

Split. Stratified 8:1:1 by speech source and attack: 3,150 train, 394 dev, and 395 eval utterances.



Picture 1. Attractor-token framework with optional fusion of SSL, LFCC, and pathological features

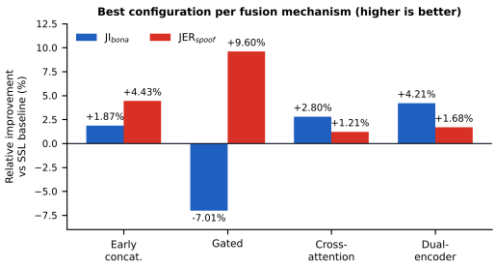
Backbone. Following Koo et al., bona fide and spoof attractor tokens are prepended to the frame sequence, processed by a Transformer and gMLP, then used in two cross-attentions for localization and diarization, followed by AHC clustering with an LCM constraint. Training uses P2SGrad loss.

Features. SSL from the Wav2Vec 2.0 XLS-R CNN encoder (T×512); LFCC with deltas (T×60); pathological jitter, shimmer, and GNE with deltas (T×24). SSL is always the primary stream.

Fusion. Early concatenation, gated fusion, cross-attention fusion, and dual-encoder fusion; three-feature runs use separate or concatenated auxiliary streams.

Metrics. JIbona for bona fide segmentation and JERspoof for spoof-method clustering (both lower is better), macro-averaged per audio, with separate early stopping for each.

Training. Latent size 256, AdamW (lr 10^{-4}) with reduce-on-plateau, batch 16, up to 512 frames (about 10 s), early stopping patience 15.



Picture 2. Relative improvement of the best configuration per fusion mechanism over the SSL baseline

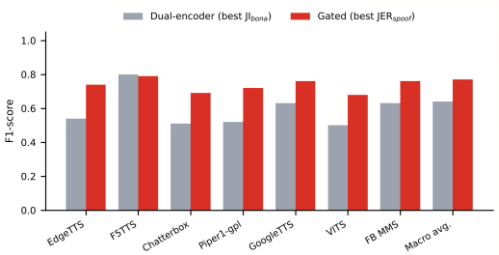
Discussion & Result

Single features. SSL is strongest (JIbona 0.0214, JERspoof 0.2323), ahead of LFCC (0.0460, 0.3363) and pathological features (0.1077, 0.3883).

Best fusion. Dual-encoder, three features: JIbona 0.0205 (−4.21%). Gated, three features with separate auxiliary: JERspoof 0.2100 (−9.60%). No mechanism wins on both metrics (Picture 2).

Errors. The best JIbona model keeps FRR at 0.48% and FAR at 4.56%. F5TTS, with zero-shot voice cloning, is the hardest to separate from bona fide speech (20.9% of its frames missed). Gated fusion raises macro F1 from 0.64 to 0.77; VITS-family systems remain mutually confused (Picture 3).

Trade-off. JERspoof-optimized models need more epochs than JIbona-optimized ones (85 vs 31 for the SSL baseline).



Picture 3. Per-class F1 of the best JIbona and best JERspoof models

Conclusion

This work presents the first Indonesian spoof diarization dataset and a systematic study of feature fusion in an attractor-token framework. Selected fusions improve different metrics over the SSL baseline: dual-encoder fusion for bona fide segmentation and gated fusion for spoof-method discrimination.

No single mechanism is universally best. Early concatenation suits similarly scaled features, gated fusion suits heterogeneous inputs, cross-attention suits sparse but informative features, and dual-encoder fusion preserves modality-specific information. Zero-shot voice cloning produces the most deceptive spoofs.

Limitations and future work. Read speech from two sources only; no speaker-disjoint split; all attacks seen in training; 10-second input limit; different checkpoints optimal for each metric. Next steps: dialects and noisy conditions, multi-objective early stopping, features targeting voice cloning, and real forensic audio.

Acknowledgement. Funded by Hibah P2MI ITB (STEI.PPMI-1-5-2026). Data use approved by the Prosa.ai Data Ethics Committee (GEN-C-2022-0007).

Key references. [1] Zhang et al., Spoof diarization, Interspeech 2024. [2] Koo et al., Token-based attractors in spoof diarization, arXiv 2025. [3] Ilman et al., Indonesian deepfake localization, O-COCOSDA 2025. [4] Mawalim et al., InaSAS, APSIPA TSIP 2025.