



FAR-CAP-DSF System for the VoicePrivacy 2026 Challenge

Farthest-Neighbor Aggregation and Reweighting with Content-Adaptive Prosody and Distance-Adaptive Stochastic Fusion

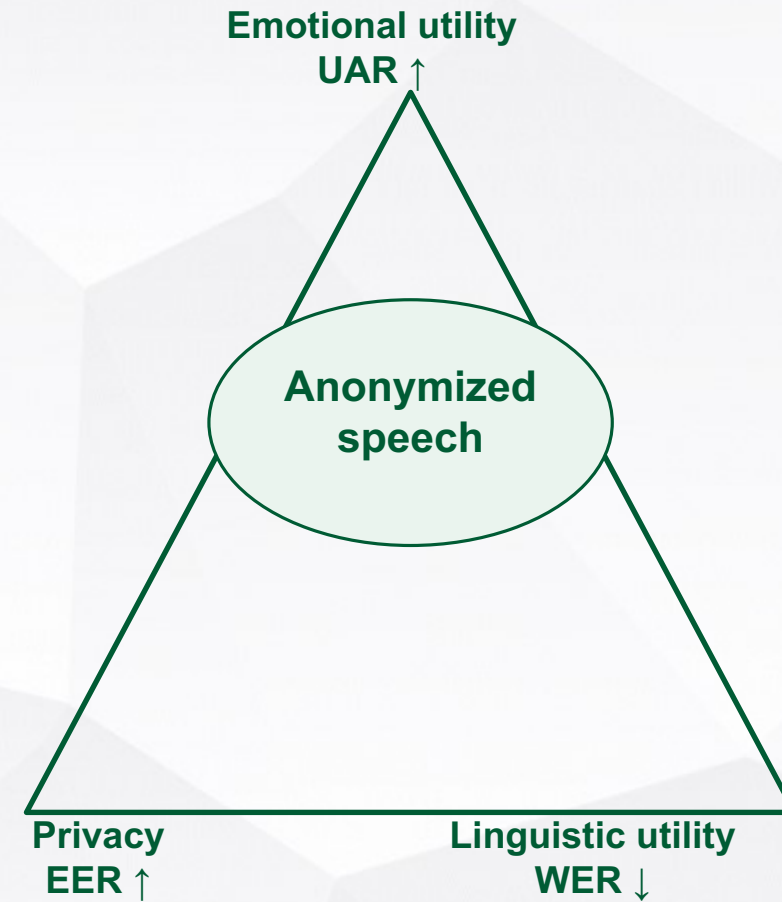
Hanbing Tian · Yi Guo · Fuqiang Hu
Xihua University, China

Remote oral presentation · SPSC 2026 / VoicePrivacy Challenge 2026 · 26 September 2026

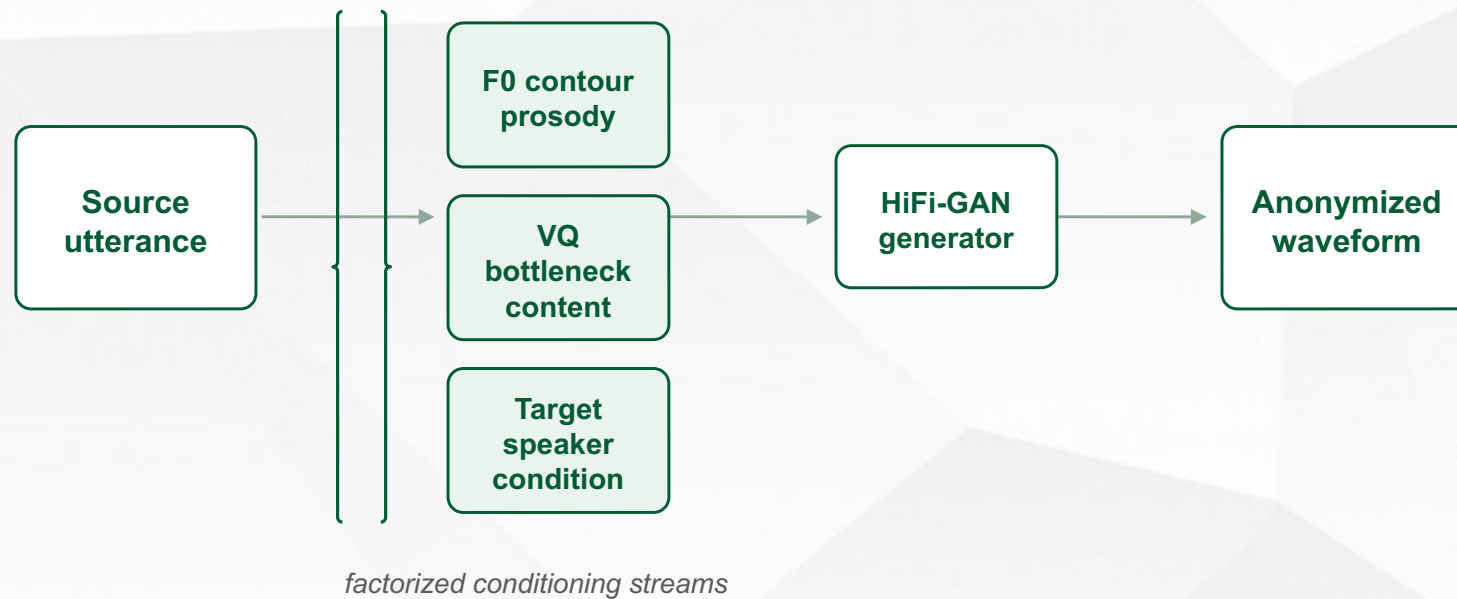
Problem

Speech carries more than words: it also contains biometric and paralinguistic cues.

- Speaker identity should be concealed.
- Linguistic content should remain intelligible.
- Emotional information should be preserved.
- Privacy gains should persist after attacker adaptation.



Key question: can an inference-time system improve adapted-attacker privacy without severe utility collapse?

B5 interface**Threat model****Lazy-informed attacker**

Fixed WavLM–ECAPA verifier; applied without system-specific retraining.

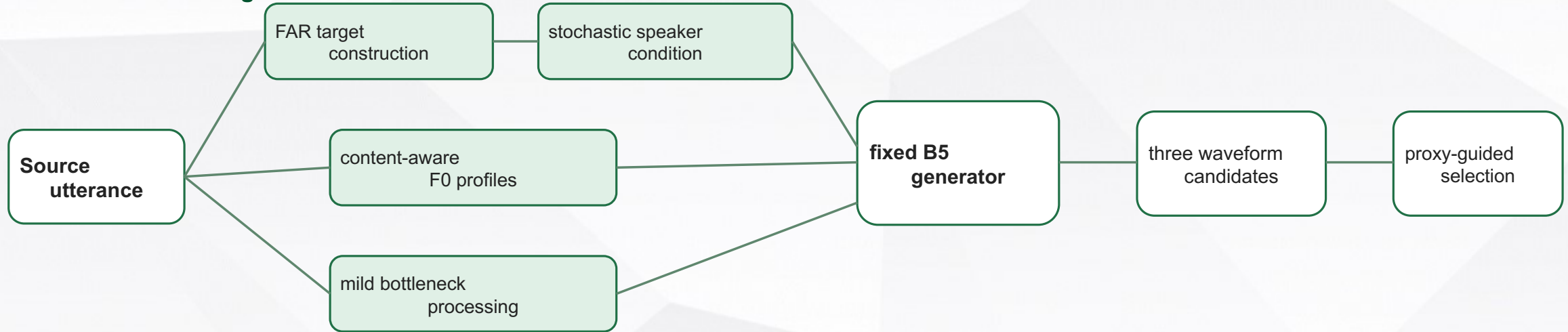
Semi-informed attacker

Same base verifier adapted to anonymized speech generated by the target system.

FAR-CAP-DSF keeps the pretrained B5 encoder and generator fixed; only inference-time conditioning is modified.

System overview

Parallel conditioning streams



No new training

The B5 encoder, speaker inventory, and HiFi-GAN generator are retained. The proposed system changes what is fed to the generator, not the generator itself.

Coordinated modifications

Speaker condition, prosody, bottleneck processing, and candidate selection are combined as a complete system configuration rather than isolated single-factor changes.

Design objective: stronger adapted-attacker privacy with controlled linguistic and emotional utility loss.

1. Farthest-neighbor target construction

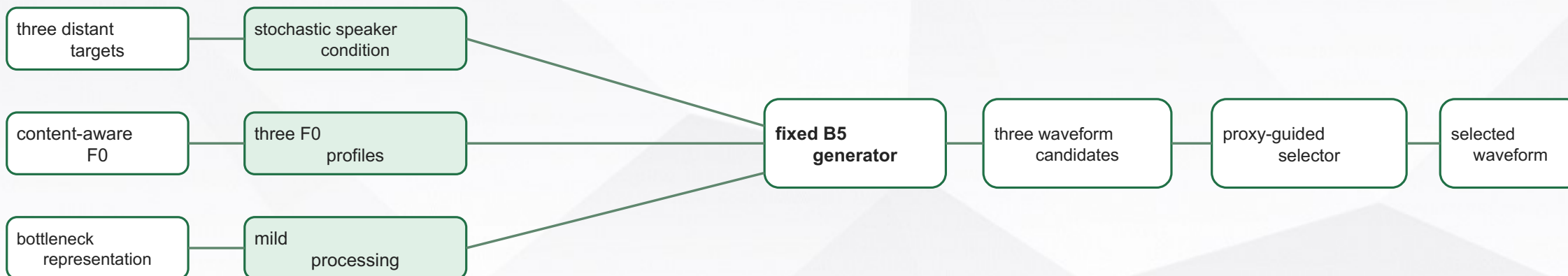
For each source utterance, the source identity is excluded. Up to 100 speaker candidates are sampled, and the three candidates farthest from the source embedding are retained.

2. Distance-adaptive stochastic fusion

The three one-hot targets are converted into a soft speaker condition. The condition is reweighted by source distance, perturbed by Gaussian noise, clipped, and renormalized.

3. Content-aware prosody + selection

Voiced F0 is moved toward a local mean with a strength guided by a vowel-likeness score. Three F0 profiles are generated and candidates are selected using privacy, content, emotion, and signal-quality proxies.



These components are evaluated as a complete configuration, not as independent causal ablations.

The selector is a routing heuristic with privacy, content, emotion, and signal-quality terms.

Evaluation

Dimension	Dataset / condition	Metric
Speaker privacy	LibriSpeech · lazy-informed and semi-informed attackers	EER ↑
Linguistic utility	LibriSpeech · ASR evaluation	WER ↓
Emotional utility	IEMOCAP · SER evaluation	UAR ↑
System comparison	B5 reference, CTRL, V01–V10, FAR-CAP-DSF	shared scoring

Controlled setup

- Official VoicePrivacy 2026 Track 1 partitions and toolkit.
- Same B5 checkpoint, target inventory, data partitions, and evaluation models.
- Separate semi-informed attacker trained for each anonymization system using the same fixed subset and adaptation budget.
- FAR-CAP-DSF implemented as an inference-time extension.

Interpretation principle

Results are compared at the complete-system level because several variants change multiple processing stages simultaneously.

Implementation: Linux Docker · Python 3.12.11 · PyTorch 2.8.0 · SpeechBrain 1.0.3 · RTX 4090

Metric	B5 Dev.	FAR Dev.	B5 Eval.	FAR Eval.
Generic EER ↑	33.276	34.660	28.574	30.507
Semi-informed EER ↑	23.837	23.070	20.466	23.366
WER ↓	4.962	6.166	4.418	5.765
UAR ↑	40.133	37.519	39.297	36.892

Evaluation partition

- Generic EER: +1.933 percentage points.
- Semi-informed EER: +2.900 percentage points.
- WER increases by +1.347 points; UAR decreases by -2.405 points.
- The improvement is therefore not a Pareto improvement over B5.

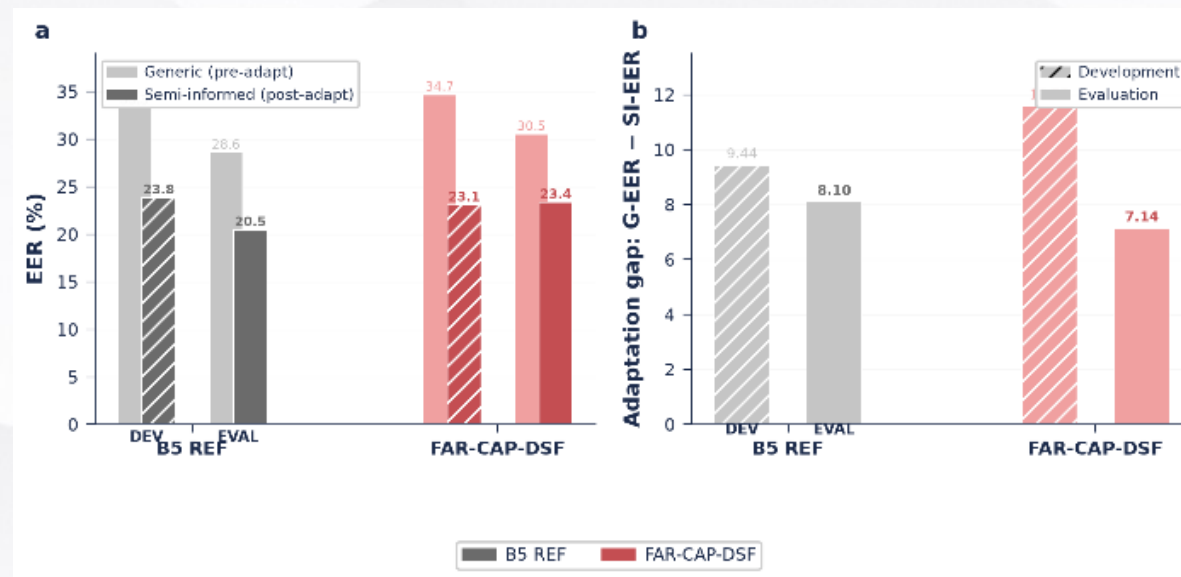
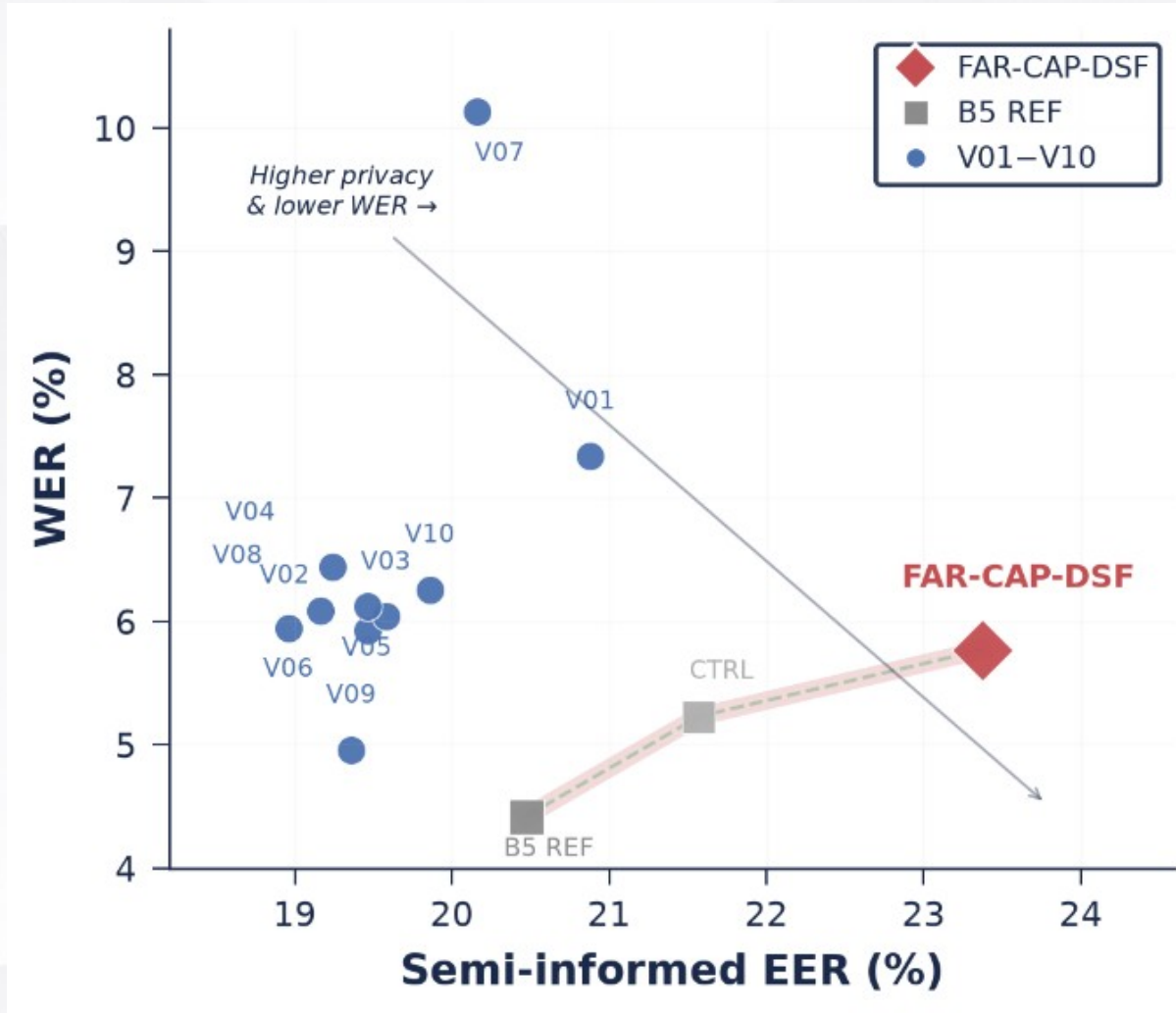


Figure: attacker adaptation gap for B5 and FAR-CAP-DSF



What the design space shows

- FAR-CAP-DSF reaches the highest evaluation semi-informed EER among the evaluated configurations.
- V07 strongly increases distortion, but its adapted-attacker privacy does not improve over B5.
- FAR-CAP-DSF has lower WER than nine of the ten exploratory variants.
- The complete coordinated configuration is supported; individual module contributions are not isolated.

Key observation

High lazy-informed EER alone does not guarantee privacy after system-specific attacker adaptation.

Configuration-level evidence

The study compares complete systems, not factorial ablations. Results cannot be uniquely attributed to a single module.

Partition dependence

Lazy-informed EER improves on both partitions, but semi-informed improvement appears only on the evaluation partition.

Stochastic stability

Reported values come from one complete anonymization and attacker-training run per configuration; run-to-run variance was not measured.

Utility scope

WER and UAR measure ASR and SER performance, but not perceived naturalness, listening effort, or human emotion recognition.

Candidate selection

Mini-batch-level selection and imbalanced profile choices suggest that utterance-level calibrated selection may use candidates more effectively.

Claim scope: FAR-CAP-DSF improves adapted-attacker privacy on the evaluation partition under the evaluated protocol, with moderate utility loss.

Take-home message

On the evaluation partition, coordinated inference-time modifications shift B5 toward stronger adapted-attacker privacy, with a clear privacy–utility trade-off.

Main result

On evaluation, semi-informed EER increases from 20.466% to 23.366%, while WER rises from 4.418% to 5.765% and UAR decreases from 39.297% to 36.892%.

Main lesson

Lazy-informed EER alone is insufficient for characterizing system behavior after system-specific attacker adaptation.

Next steps

Repeated-run stability, controlled component ablations, improved emotion preservation, and utterance-level candidate selection.

Thank you