

FAR-CAP-DSF System for the VoicePrivacy 2026 Challenge

Hanbing Tian¹, Yi Guo², Fuqiang Hu³

¹ Xihua University, China

² Xihua University, China

³ Xihua University, China

2604353676@qq.com, lpngy@vip.163.com, hfqemail@163.com

Abstract

Voice anonymization must conceal speaker identity while preserving linguistic and emotional content, yet privacy gains measured with a lazy-informed attacker may not persist after system-specific adaptation. We present Farthest-Neighbor Aggregation and Reweighting with Content-Adaptive Prosody and Distance-Adaptive Stochastic Fusion (FAR-CAP-DSF), an inference-time extension of the VoicePrivacy 2026 Track 1 B5 baseline. The system combines farthest-neighbor target construction, distance-adaptive stochastic speaker fusion, content-aware F0 modification, mild bottleneck processing, and proxy-guided candidate selection while keeping the pretrained analysis and synthesis networks fixed. We evaluate a uniform-prosody control and ten exploratory system configurations spanning prosody, speaker conditioning, bottleneck processing, and alternative fusion strategies under shared lazy-informed and semi-informed attack conditions. On the evaluation partition, FAR-CAP-DSF increases lazy-informed EER from 28.57% to 30.51% and semi-informed EER from 20.47% to 23.37% relative to B5, achieving the highest semi-informed EER among the evaluated configurations. This privacy improvement is accompanied by a moderate utility trade-off, with WER increasing from 4.42% to 5.77% and UAR decreasing from 39.30% to 36.89%. On the development partition, FAR-CAP-DSF also improves lazy-informed EER, while the semi-informed result follows a different trend. Overall, the results indicate that coordinated inference-time modifications can improve resistance to a system-adapted attacker on the evaluation partition while avoiding the severe linguistic degradation observed with more aggressive perturbation strategies.

Index Terms: voice anonymization, speaker privacy, VoicePrivacy Challenge, lazy-informed attacker, semi-informed attacker, privacy-utility trade-off

1. Introduction

Speech carries not only linguistic content, but also biometric and paralinguistic information that can facilitate speaker re-identification. Voice anonymization therefore aims to transform a speech waveform so that the original speaker identity is more difficult to recover, while retaining sufficient linguistic content, emotional information, and signal quality for downstream use. The VoicePrivacy Challenge has progressively formalized this objective through common datasets, threat models, and privacy-utility metrics. Analyses of the 2022 and 2024 editions indicate that current systems exhibit marked architecture-dependent trade-offs: approaches providing stronger identity protection do not necessarily preserve linguistic and emotional attributes equally well, and no single design consistently dominates across all evaluation dimensions [1]–[3].

Existing voice anonymization systems span several architectural families. ASR-TTS pipelines seek to suppress identity information through phonetic intermediate representations [4], while neural-codec and disentanglement-based systems attempt to separate linguistic, emotional, and speaker-related factors [5]. Voice conversion and synthesis systems have also been combined or selected to balance their complementary privacy and emotion-preservation characteristics [6]. Within the VoicePrivacy baselines, recent

studies have explored stronger speaker encoders, emotion conditioning, pseudo-speaker interpolation, and F0 transformation [7]. In particular, modifications to the F0 trajectory have been investigated as a means of reducing residual speaker information while maintaining compatibility between prosody and the anonymized speaker representation [8]. These studies establish several useful design directions, but their reported gains remain dependent on the underlying synthesis architecture, evaluation protocol, and attacker model.

The attacker model is especially important when interpreting anonymization results. The First VoicePrivacy Attacker Challenge showed that stronger speaker-verification systems can substantially reduce the apparent privacy of previously evaluated anonymization systems [9]. VoicePrivacy 2026 distinguishes between a lazy-informed condition, in which the provided WavLM-ECAPA verifier is applied without system-specific retraining, and a semi-informed condition, in which the same base verifier is adapted to speech generated by the target anonymization system [10]. The protocol also evaluates mixed trials containing same-gender and cross-gender comparisons. Under this setting, an intervention that increases EER against the lazy-informed attacker may provide a smaller benefit, no benefit, or even a degradation after system-specific adaptation.

We investigate this issue using the ASR-BN B5 system of VoicePrivacy 2026 Track 1 as a common backbone. The study includes a uniform-prosody control, ten exploratory variants, and the final FAR-CAP-DSF system, covering target-speaker construction, speaker-condition fusion, F0 transformation, bottleneck perturbation, stochastic processing, and proxy-guided candidate selection. FAR-CAP-DSF combines farthest-neighbor target construction, distance-adaptive stochastic speaker fusion, content-aware F0 modification, mild bottleneck processing, batch-level candidate selection, and waveform-level output protection. Because several variants alter multiple processing stages simultaneously, the experiments are organized as a hypothesis-driven design-space comparison rather than a controlled single-factor ablation study. Accordingly, the comparisons characterize complete system configurations and are not used to assign independent causal contributions to individual components.

On the evaluation partition, FAR-CAP-DSF increases lazy-informed EER from 28.57% to 30.51% and semi-informed EER from 20.47% to 23.37% relative to the B5 reference, while WER rises from 4.42% to 5.77% and UAR falls from 39.30% to 36.89%. On the development partition, semi-informed EER decreases from 23.84% to 23.07%, showing that the adapted-attacker improvement is partition-dependent. The main contributions are a modular inference-time extension of B5, a controlled lazy-informed and semi-informed comparison across the explored configurations under a shared evaluation protocol, and an analysis of when fixed-attacker privacy increases persist after adaptation and how they interact with linguistic and emotional utility.

2. B5 Baseline and Proposed FAR-CAP-DSF

All systems share the same pretrained B5 backbone. FAR-CAP-DSF modifies only inference-time conditioning and candidate selection.

2.1. VoicePrivacy 2026 B5 Baseline

B5 factorizes an utterance into an F0 contour, a vector-quantized linguistic bottleneck, and a target-speaker condition, which are jointly decoded by a HiFi-GAN generator [10]–[14]. YAAPT provides the explicit prosodic contour, while a wav2vec 2.0-TDNN-F encoder followed by vector quantization produces the content representation. This factorized interface exposes prosody, linguistic features, and speaker identity as distinct conditioning streams.

Speaker identity in the reference system is represented by a one-hot vector selected independently for each utterance from the synthesis inventory. The F0 contour, bottleneck sequence, and selected identity are supplied to the pretrained generator to synthesize the anonymized waveform. FAR-CAP-DSF retains the released B5 encoder, speaker inventory, and generator, so the comparison concerns inference-time conditioning rather than retraining or replacement of the backbone.

2.2. System Overview

FAR-CAP-DSF modifies the three B5 conditioning streams before synthesis. Three distant target identities are aggregated into a stochastic soft speaker condition, the original F0 contour is transformed into three content-aware profiles, and the bottleneck representation receives mild perturbation and temporal smoothing.

Since the target identities are fused before synthesis, the modified bottleneck sequence and the three F0 profiles yield three waveform candidates. A proxy-guided selector chooses one candidate for the complete mini-batch, after which an output guard enforces numerical validity and basic waveform constraints. All modifications are applied at inference time while the B5 encoder and generator remain fixed.

2.3. Target Construction and Stochastic Fusion

For each utterance, the source identity is excluded and a pool of up to 100 speakers is sampled from the 247-speaker synthesis inventory.

For candidate i in the sampled pool, the cosine distance to the source embedding is defined as

$$d_i = 1 - \frac{z_s^T z_i}{\|z_s\|_2 \|z_i\|_2}, i \in P.$$

The three candidates with the largest source-to-candidate distances are retained independently for each utterance. Because ranking is restricted to a sampled pool, this is an approximate farthest-neighbor search rather than a globally optimal assignment. The procedure explicitly favors distance from the source speaker but does not optimize the pairwise separation among the three retained targets. FAR should therefore be interpreted as a source-distance-oriented target-construction strategy rather than as a mechanism that guarantees maximal diversity among the selected targets. Distance is used only for target construction and not as a surrogate for the official privacy metrics. The retained distances are converted to normalized fusion weights:

$$w_i = \frac{\exp(d_i/\tau)}{\sum_{j=1}^3 \exp(d_j/\tau)},$$

where the temperature is set to 0.5. Given the corresponding one-hot vectors, the initial fused condition is

$$s^{(0)} = \sum_{i=1}^3 w_i e_i.$$

The fused vector is then reweighted as a function of its distance from the source, using a base coefficient of 0.02 and a distance range of 0.15; a larger source-to-fused distance yields a larger adjustment. Gaussian perturbation with coefficient 0.10

is applied to the full speaker-condition vector, followed by non-negativity clipping and renormalization. Because candidate-pool sampling and speaker-condition perturbation are stochastic, repeated anonymization of the same source utterance is not constrained to produce an identical conditioning vector.

Although initialized as a sparse mixture of three target identities, the perturbed condition is generally dense. We refer to this distance-conditioned reweighting and random perturbation as distance-adaptive stochastic speaker fusion. The present study does not explicitly quantify repeated-run dispersion or recurrence of the resulting effective speaker conditions; these aspects are therefore not used as evidence for the reported privacy claims.

Although initialized as a sparse mixture of three target identities, the perturbed condition is generally dense. We refer to this distance-conditioned reweighting and random perturbation as distance-adaptive stochastic speaker fusion.

2.4. Content-Aware Prosody and Bottleneck Processing

The prosody module applies mean reversion to voiced F0 with a strength controlled by a signal-derived vowel-likeness score. The score is estimated from local energy stability and normalized within-frame variation, without a phoneme recognizer or annotated phone boundaries. After interpolation to the F0 frame rate and temporal smoothing, it controls a continuous coefficient from 0.20 in less vowel-like regions to 0.45 in more vowel-like regions.

Each voiced F0 value is moved toward a local mean computed over a 32-frame window, with three-frame transition smoothing. Three profiles share the same mean-reversion rule but use F0-domain noise levels of 12, 8, and 4 dB, providing strong, medium, and mild alternatives. Unvoiced frames are excluded from the local F0 statistics.

The bottleneck representation is mixed with a channel-wise normalized version using coefficient 0.05, perturbed by zero-mean Gaussian noise with standard deviation 0.002, and smoothed with a three-frame moving average. The same global mixing coefficient is used for all frames, making this intervention deliberately mild relative to the prosodic and speaker-condition transformations.

Both transformations are applied to the conditioning representations before synthesis; the pretrained B5 encoder and HiFi-GAN generator remain unchanged.

2.5. Candidate Selection and Output Guard

The fused speaker condition is paired with the three F0 profiles to generate three waveform candidates. Candidate scoring combines speaker-embedding distance with weight 0.70, bottleneck similarity with weight 0.15, and an emotion-oriented F0/RMS preservation proxy with weight 0.15, together with a signal-quality penalty. The privacy term compares embeddings of the original and anonymized waveforms, while the emotion term summarizes voiced-F0 correlation and energy similarity when available.

Selection is performed at the mini-batch level, with one candidate index applied to every utterance in the batch. The score is a routing heuristic rather than a direct optimization of EER, WER, or UAR. Because all candidates share the same modified bottleneck representation, the content term varies little across profiles; selection is driven mainly by the speaker, prosody, and signal-quality terms.

After selection, the output guard replaces non-finite samples with zero, constrains sample amplitudes, truncates outputs that exceed the input length, and amplifies exceptionally low-RMS waveforms using a bounded gain. These operations protect the evaluation pipeline from invalid outputs and are not an additional privacy transformation.

3. Experimental Setup

All systems were evaluated using the VoicePrivacy 2026 Track 1 partitions and evaluation toolkit [10]. The B5 reference and all intervention configurations used identical data and scoring procedures. Privacy is reported under lazy-informed and semi-informed attacker conditions, while linguistic and emotional utility are evaluated by ASR and SER, respectively.

3.1. Datasets and Metrics

Experiments used the official VoicePrivacy 2026 Track 1 development and evaluation partitions derived from LibriSpeech [15] and IEMOCAP [16]. LibriSpeech was used for speaker-verification and speech-recognition evaluation, whereas IEMOCAP was used for speech-emotion recognition. Speaker-verification trials were scored under the mixed condition, which includes same-gender and cross-gender comparisons [10].

Privacy was measured by anonymized-enrollment versus anonymized-trial equal error rate (EER). Linguistic utility was measured by word error rate (WER), and emotional utility by unweighted average recall (UAR) over the four IEMOCAP emotion classes. Higher EER, lower WER, and higher UAR indicate better performance in the corresponding dimension. Lazy-informed and semi-informed EERs are reported separately.

Table 1: Evaluation datasets and metrics.

Dataset	Partitions	Evaluation target	Metric
LibriSpeech	Development, evaluation	Speaker privacy	Generic and semi-informed EER
LibriSpeech	Development, evaluation	Linguistic utility	WER
IEMOCAP	Development, evaluation	Emotional utility	UAR

3.2. Generic and Semi-Informed Attackers

Privacy was evaluated under lazy-informed and semi-informed attack conditions. The lazy-informed attacker is the fixed WavLM-ECAPA speaker-verification model provided by the challenge. It is pretrained on large-scale voice-converted speech and applied directly to anonymized enrollment and trial utterances without system-specific adaptation [10], providing a common fixed-attacker condition across the evaluated anonymization systems.

The semi-informed attacker is obtained by adapting the same base ASV model to speech generated by the system under evaluation. For controlled comparison across the explored configurations, a separate attacker was trained for each system using an anonymized version of the same fixed 51,773-utterance subset of LibriSpeech train-clean-360. The source utterances, speaker population, attacker initialization, training schedule, and adaptation budget were therefore held constant across configurations; only the anonymization applied to the attacker-training data differed. This common fixed-subset protocol allows the semi-informed results to be interpreted as controlled relative comparisons among the evaluated systems.

Adaptation used all available speakers and utterances in the fixed 51,773-utterance subset for two epochs. Training used three-second random segments, a batch size of 32, AdamW optimization, and a cyclic learning-rate schedule with a maximum learning rate of 0.01. The seed was fixed to 1986, and the second-epoch checkpoint was used for evaluation. The same adaptation procedure was applied to every evaluated configuration.

Lazy-informed and semi-informed EERs are reported as separate conditions. The lazy-informed condition measures resistance to the fixed verifier without system-specific

retraining, whereas the semi-informed condition measures resistance after the same base verifier has been adapted to anonymized speech produced by the target system. In this study, the latter comparison uses the common fixed-subset adaptation protocol described above, so differences among configurations are interpreted under the same attacker-training budget.

3.3. Evaluated Configurations

All evaluated systems share the pretrained B5 content encoder and waveform generator. The comparison includes the original B5 reference, an extended control system, ten exploratory configurations, and FAR-CAP-DSF. The configurations vary prosodic processing, bottleneck transformation, target-speaker construction, speaker-condition fusion, stochastic perturbation, candidate selection, and waveform post-processing.

The configurations are grouped according to their dominant design question. These group labels describe the primary processing emphasis of each configuration but should not be interpreted as controlled component ablations. Because several variants change more than one processing stage, the study compares complete system configurations and does not claim module-wise causal attribution.

Table 2: Summary of the evaluated configurations.

System	Main characteristic
REF	Original B5
CTRL	Uniform F0 control
V01	Strong uniform F0
V02	Content-aware F0
V03	Multi-strength F0
V04	Expanded target pool
V05	Stronger speaker perturbation
V06	Randomized reweighting
V07	Bottleneck-dropout stress
V08	Gumbel speaker fusion
V09	Differential F0
V10	Density-aware target selection
FAR-CAP-DSF	Proposed coordinated system

REF is the unmodified B5 system and provides the primary comparison point. CTRL introduces the extended processing pipeline while retaining a uniform F0 transformation, serving as a control for the content-aware prosody variants.

V01–V03 focus on prosodic processing. V01 strengthens a uniform F0 transformation, V02 replaces uniform treatment with content-aware modulation, and V03 generates multiple F0 profiles with different transformation strengths. This group examines how the structure and intensity of prosodic modification affect the privacy–utility operating point.

V04–V06 focus on target-speaker construction and speaker-condition perturbation. V04 enlarges the candidate target pool, V05 applies stronger reweighting and stochastic perturbation to the fused condition, and V06 introduces randomized reweighting. These configurations examine the effects of candidate-pool size and stochastic speaker-conditioning strategies, without assuming that a larger pool or stronger stochasticity necessarily produces greater pairwise diversity among the resulting target conditions.

V07–V10 examine alternative representation and fusion strategies. V07 emphasizes bottleneck dropout, V08 replaces soft fusion with a Gumbel-based formulation, V09 uses differential-F0 processing instead of mean reversion, and V10 combines density-aware target selection with stronger conditioning transformations.

FAR-CAP-DSF combines farthest-neighbor target construction, content-aware F0 processing, mild bottleneck

transformation, distance-adaptive stochastic speaker fusion, and proxy-guided candidate selection.

3.4. Implementation and Reproducibility

All systems were implemented as extensions of the official VoicePrivacy 2026 B5 pipeline. The pretrained B5 content encoder and HiFi-GAN generator were kept fixed, and the proposed transformations were applied only during inference. The same B5 checkpoint, target-speaker inventory, data partitions, and official evaluation models were used for all configurations.

FAR-CAP-DSF used cached representations for all 247 speakers in the synthesis inventory. For each utterance, three target speakers were selected from a candidate pool of up to 100 speakers and fused into a single speaker condition. Three prosodic profiles were then used to generate three waveform candidates. Anonymization was performed with a batch size of four, and candidate selection was applied at the mini-batch level.

Experiments were conducted in a Linux-based Docker environment using Python 3.12.11, PyTorch 2.8.0, and SpeechBrain 1.0.3 on a single NVIDIA GeForce RTX 4090 GPU.

4. Results and Discussion

This section compares FAR-CAP-DSF with the B5 reference, analyzes the explored design space and attacker adaptation, and discusses the resulting privacy–utility trade-off and limitations.

4.1. Main Results

Table 3 summarizes privacy under generic and semi-informed attackers together with linguistic and emotional utility for B5 and FAR-CAP-DSF.

Table 3: Main results.

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

Against the lazy-informed attacker, FAR-CAP-DSF increases EER by 1.384 percentage points on the development partition and 1.933 points on the evaluation partition, indicating a consistent increase in difficulty for the fixed verifier under this attack condition.

After attacker adaptation, the trend becomes partition-dependent. Semi-informed EER decreases by 0.767 points on development but increases by 2.900 points on evaluation. Thus, under the common fixed-subset adaptation protocol, the semi-informed improvement observed on the evaluation partition does not reproduce on the development partition.

Utility decreases on both partitions. Relative to B5, WER increases by 1.204 points on development and 1.347 points on evaluation, while UAR decreases by 2.615 and 2.405 points, respectively.

FAR-CAP-DSF therefore establishes a more privacy-oriented operating point on the evaluation partition, particularly under the semi-informed condition, rather than a uniform improvement over B5 across privacy and utility measures.

4.2. Design-Space Comparison

Table 4 compares the explored configurations on the evaluation partition. The study asks whether emphasis on a single

anonymization pathway is sufficient for adapted-attacker privacy or whether a coordinated configuration offers a more favorable operating point. The variants are analyzed in the three groups introduced in Section 3.3; because several alter multiple stages, the evidence is interpreted at the configuration level.

Table 4: Evaluation-partition comparison.

System	G-EER ↑	SI-EER ↑	WER ↓	UAR ↑
REF	28.57	20.47	4.42	39.30
CTRL	29.49	21.57	5.23	35.98
V01	28.19	20.88	7.34	37.53
V02	26.17	19.46	5.93	37.19
V03	27.45	19.58	6.04	36.86
V04	27.60	19.24	6.44	38.27
V05	27.69	19.46	6.12	37.58
V06	27.58	18.96	5.95	37.00
V07	31.07	20.16	10.13	36.07
V08	28.10	19.16	6.09	38.06
V09	26.78	19.36	4.96	37.05
V10	28.19	19.86	6.26	37.84
FAR	30.51	23.37	5.77	36.89

V01–V03 test prosody-focused strategies. V01 raises SI-EER from 20.47% for REF to 20.88%, but WER increases from 4.42% to 7.34%. V02 and V03 reduce this linguistic degradation, with WERs of 5.93% and 6.04%, yet their SI-EERs remain below REF. In the tested settings, stronger uniform F0 modification yields only a small privacy gain at substantial ASR cost, while the content-aware and multi-strength alternatives do not retain that gain.

V04–V06 test target diversity and stochastic speaker conditioning. V04 retains a comparatively high UAR of 38.27% but reaches an SI-EER of 19.24%; V05 and V06 obtain 19.46% and 18.96%, respectively. Expanding the target pool, increasing perturbation strength, or randomizing reweighting is therefore not sufficient as a standalone strategy for resisting the corresponding adapted attacker.

V07 illustrates the mismatch between lazy-informed and adapted-attacker results. It attains the highest LI-EER, 31.07%, but its SI-EER is 20.16%, slightly below REF, and its WER reaches 10.13%. The large EER increase under the fixed attacker therefore does not persist after system-specific adaptation and is accompanied by substantial linguistic distortion.

V08 and V10 retain relatively high UAR values of 38.06% and 37.84%, but neither improves SI-EER over REF. V09 provides the lowest WER among V01–V10, at 4.96%, while its SI-EER is 19.36%. These alternatives show that a high lazy-informed EER or a favorable individual utility measure does not by itself imply a favorable operating point after system-specific attacker adaptation.

The REF–CTRL–FAR comparison provides the clearest system-level contrast. CTRL raises SI-EER from 20.47% to 21.57%. FAR-CAP-DSF further increases SI-EER to 23.37% and improves UAR from 35.98% to 36.89% relative to CTRL, while WER rises moderately from 5.23% to 5.77%. Because these systems differ at several processing stages, the comparison supports the complete coordinated configuration rather than attributing the observed difference to any single module.

Across V01–V10, FAR-CAP-DSF exceeds the highest SI-EER by 2.49 points and achieves a lower WER than nine configurations. Within the evaluated design space and under the common adaptation protocol, the complete coordinated configuration provides the highest evaluation-partition SI-EER among the compared systems. This observation does not establish that any individual component, or any specific interaction among components, is independently responsible for the result.

The result should therefore be interpreted at the system-configuration level. The SI-EER advantage on evaluation is accompanied by lower utility than B5 and, as shown next, is not reproduced on the development partition.

4.3. Effect of Attacker Adaptation

Fig. 1 compares the lazy-informed and semi-informed EERs of B5 and FAR-CAP-DSF and summarizes the effect of system-specific attacker adaptation. We define the adaptation gap as lazy-informed EER minus semi-informed EER; a smaller gap indicates a smaller reduction in EER after attacker adaptation.

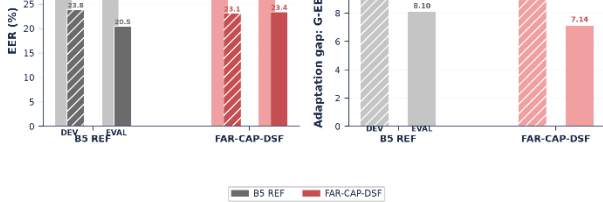


Fig. 1. System-specific attacker adaptation for B5 and FAR-CAP-DSF: (a) lazy-informed and semi-informed EERs; (b) adaptation gap (LI-EER minus SI-EER), where smaller values indicate a smaller EER reduction after adaptation. Units are percent and percentage points, respectively.

Attacker adaptation lowers EER for both systems on both partitions. On development, FAR-CAP-DSF has a larger adaptation gap than B5 (11.59 versus 9.44 percentage points). This reverses the lazy-informed ordering: FAR-CAP-DSF is 1.38 points higher in lazy-informed EER but 0.77 points lower in semi-informed EER.

On evaluation, FAR-CAP-DSF has a smaller adaptation gap than B5 (7.14 versus 8.11 points), and its relative EER difference over B5 increases from 1.93 points under the lazy-informed attacker to 2.90 points under the semi-informed attacker. Thus, the evaluation-partition EER increase remains after system-specific adaptation and is not limited to the fixed lazy-informed attack condition.

The opposite trends across partitions show that lazy-informed EER alone is insufficient for characterizing system behavior after attacker adaptation. Under the same fixed-subset adaptation protocol, the development and evaluation partitions yield different relative outcomes, indicating that the observed effect is partition-dependent rather than uniformly reproduced. The present experiments do not determine whether this difference arises from data composition, interactions with the stochastic transformations, or run-to-run variation. Accordingly, the semi-informed improvement is claimed for the evaluation partition only, while the cross-configuration results are interpreted as comparative trends under the common adaptation setting.

4.4. Privacy–Utility Trade-off

Fig. 2 visualizes the evaluation-partition privacy–utility operating points in terms of semi-informed EER and WER. Movement toward the lower right corresponds to higher semi-informed EER and lower WER under the evaluated protocol.

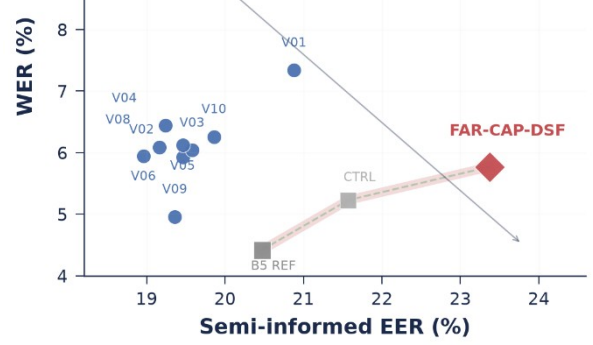


Fig. 2. Privacy–utility operating points on the evaluation partition. Higher semi-informed EER and lower WER are preferred. REF denotes the B5 reference, CTRL denotes the uniform-F0 control, and FAR-CAP-DSF denotes the proposed system.

FAR-CAP-DSF achieves the highest semi-informed EER among the evaluated configurations while maintaining a substantially lower WER than the most aggressive variant, V07. Relative to B5, its operating point shifts toward stronger privacy, but the accompanying WER increase prevents Pareto dominance. CTRL lies between B5 and FAR-CAP-DSF, whereas V09 retains relatively low WER but provides limited semi-informed privacy.

Quantitatively, FAR-CAP-DSF improves semi-informed EER over B5 by 2.90 points, with a 1.35-point increase in WER and a 2.41-point decrease in UAR. Relative to V07, it raises semi-informed EER by 3.21 points while reducing WER by 4.36 points, showing that the privacy gain is not obtained by maximizing distortion.

This trade-off remains partition-dependent. On development, FAR-CAP-DSF incurs comparable utility loss without improving semi-informed EER over B5. The results therefore establish a favorable privacy-oriented operating point on the evaluation partition rather than a universal Pareto improvement.

4.5. Limitations

The configuration study compares complete anonymization strategies rather than factorial or one-variable-at-a-time ablations. Several exploratory variants, as well as the final FAR-CAP-DSF system, modify more than one processing stage. Consequently, the observed differences should be interpreted at the configuration level and cannot be uniquely attributed to an individual component or a particular interaction among components. Controlled component ablations would be required to isolate the independent contributions of target selection, stochastic fusion, F0 transformation, bottleneck processing, candidate selection, and the other processing stages.

FAR is designed to increase source-to-target distance within a randomly sampled candidate pool, but it does not explicitly optimize pairwise diversity among the retained targets. The present study does not quantify pairwise distances among the selected targets, repeated-run variation of the resulting fused speaker conditions, or the frequency with which similar target sets or effective conditioning regions are revisited under repeated sampling. The current results therefore do not establish maximal or guaranteed pseudo-speaker diversity across repeated anonymization runs. FAR should instead be interpreted as a stochastic, source-distance-oriented target-construction strategy.

The reported values are point estimates from one complete anonymization and attacker-training run per configuration. The prosodic and speaker-condition transformations are stochastic, and variance across repeated anonymization or attacker-training

runs was not measured. Consequently, the present experiments do not establish the run-to-run stability of the reported EER, WER, or UAR differences. This limitation is particularly relevant given the opposite semi-informed trends observed on the development and evaluation partitions.

Privacy was assessed with the challenge-provided WavLM-ECAPA attacker and its prescribed system-specific adaptation protocol. Cross-architecture verification, alternative adaptation budgets, linkage attacks, and white-box access may expose other residual identity cues; the reported EERs are therefore specific to the evaluated threat models.

Utility evaluation is limited to the official ASR and SER metrics. WER and UAR do not directly measure perceived naturalness, listening effort, anonymous-speaker consistency, or human emotion recognition. No subjective listening study was conducted, and class-specific SER errors may reflect both information loss and limitations of the fixed classifier.

The candidate selector operates at the mini-batch level, and the shared bottleneck representation provides little content discrimination among the three F0 profiles. The selected-profile distribution is also strongly imbalanced. An utterance-level selector calibrated on held-out privacy and utility outcomes may use the candidate set more effectively.

Stochastic fusion produces a dense speaker condition that differs from the one-hot conditioning used to train the B5 generator. Although the resulting speech is evaluable, the present experiments do not separate the contribution of this distribution shift to privacy from its contribution to utility loss.

5. Conclusion

This paper presented FAR-CAP-DSF, an inference-time extension of the VoicePrivacy 2026 B5 baseline. The system retains the pretrained B5 analysis and synthesis networks while coordinating farthest-neighbor target construction, content-aware prosody, mild bottleneck processing, distance-adaptive stochastic speaker fusion, and proxy-guided candidate selection.

The configuration study examined prosody-focused, speaker-conditioning, and alternative representation or fusion strategies under shared lazy-informed and semi-informed attack conditions. Within the evaluated configurations and the common adaptation setting, FAR-CAP-DSF achieved the highest evaluation-partition SI-EER and a lower WER than nine of the ten exploratory variants. Because the configurations are not controlled component ablations, this result should be interpreted at the complete-system level rather than as evidence that each constituent component is independently beneficial.

Relative to B5 on the evaluation partition, FAR-CAP-DSF increased lazy-informed EER from 28.57% to 30.51% and semi-informed EER from 20.47% to 23.37%, while WER increased from 4.42% to 5.77% and UAR decreased from 39.30% to 36.89%. The result therefore represents a more privacy-oriented operating point under the evaluated conditions rather than a Pareto improvement.

The lazy-informed EER increase appeared on both partitions, whereas the semi-informed improvement was observed only on the evaluation partition under the common fixed-subset adaptation protocol. This partition dependence underscores the importance of system-specific attacker adaptation and limits the scope of the present privacy claim. Future work will assess repeated-run stability and target-condition dispersion, perform controlled component ablations, improve emotion preservation, and develop utterance-level candidate selection.

6. Acknowledgements

Hanbing Tian contributed to methodology, experimentation, and manuscript preparation. Fuqiang Hu contributed to experimentation and visualization. Yí Guo contributed to methodology and supervision and served as the corresponding author.

7. Generative AI Disclosure

Generative AI tools were used solely for language polishing. All methodological content, experimental results, analyses, and conclusions were produced and verified by the authors.

8. References

- [1] M. Panariello, N. Tomashenko, X. Wang, X. Miao, P. Champion, H. Nourtel, M. Todisco, N. Evans, E. Vincent, and J. Yamagishi, “The VoicePrivacy 2022 Challenge: Progress and perspectives in voice anonymisation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3477–3491, 2024, doi: 10.1109/TASLP.2024.3430530.
- [2] N. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent, M. Panariello, N. Evans, J. Yamagishi, and M. Todisco, “The VoicePrivacy 2024 Challenge evaluation plan,” *arXiv preprint arXiv:2404.02677*, 2024.
- [3] N. Tomashenko, X. Miao, P. Champion, S. Meyer, M. Panariello, X. Wang, N. Evans, E. Vincent, J. Yamagishi, and M. Todisco, “The third VoicePrivacy challenge: Preserving emotional expressiveness and linguistic content in voice anonymization,” *Computer Speech & Language*, vol. 100, Art. no. 101988, 2026, doi: 10.1016/j.csl.2026.101988.
- [4] S. Meyer, F. Lux, P. Denisov, J. Koch, P. Tilli, and N. T. Vu, “Speaker anonymization with phonetic intermediate representations,” in *Proc. Interspeech 2022*, 2022, pp. 4925–4929, doi: 10.21437/Interspeech.2022-10703.
- [5] J. Yao, N. Kuzmin, Q. Wang, P. Guo, Z. Ning, D. Guo, K. A. Lee, E.-S. Chng, and L. Xie, “NPU-NTU system for Voice Privacy 2024 Challenge,” in *Proc. 4th Symposium on Security and Privacy in Speech Communication*, 2024, pp. 67–71, doi: 10.21437/SPSC.2024-12.
- [6] H. L. Xinyuan, Z. Cai, A. Garg, K. Duh, L. P. García-Perera, S. Khudanpur, N. Andrews, and M. Wiesner, “HLTCOE JHU Submission to the Voice Privacy Challenge 2024,” in *Proc. 4th Symposium on Security and Privacy in Speech Communication*, 2024, pp. 61–66, doi: 10.21437/SPSC.2024-11.
- [7] N. Kuzmin, H.-T. Luong, J. Yao, L. Xie, K. A. Lee, and E.-S. Chng, “NTU-NPU system for Voice Privacy 2024 Challenge,” in *Proc. 4th Symposium on Security and Privacy in Speech Communication*, 2024, pp. 72–79, doi: 10.21437/SPSC.2024-13.
- [8] U. E. Gaznepoglu, A. Leschanowsky, and N. Peters, “VoicePrivacy 2022 system description: Speaker anonymization with feature-matched f0 trajectories,” in *Proc. 2nd Symposium on Security and Privacy in Speech Communication*, 2022.
- [9] N. Tomashenko, X. Miao, E. Vincent, and J. Yamagishi, “The First VoicePrivacy Attacker Challenge,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–2, doi: 10.1109/ICASSP49660.2025.10888513.
- [10] X. Miao, N. Tomashenko, R. Arefeen, S. Meyer, M. Panariello, X. Wang, E. Vincent, N. Evans, J. Yamagishi, and M. Todisco, “The VoicePrivacy 2026 Challenge Evaluation Plan,” *VoicePrivacy Challenge Technical Report*, Version 1.2, 2026.
- [11] K. Kasi and S. A. Zahorian, “Yet another algorithm for pitch tracking,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2002, pp. 1-361–1-364, doi: 10.1109/ICASSP.2002.5743729.
- [12] A. Baeviski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12449–12460.
- [13] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, “Semi-orthogonal low-rank matrix factorization for deep neural networks,” in *Proc. Interspeech 2018*, 2018, pp. 3743–3747, doi: 10.21437/Interspeech.2018-1417.
- [14] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative adversarial networks for efficient and high-fidelity speech synthesis,” in

Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 17022–17033.

- [15] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “LibriSpeech: An ASR corpus based on public domain audio books,” in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5206–5210, doi: 10.1109/ICASSP.2015.7178964.
- [16] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “IEMOCAP: Interactive emotional dyadic motion capture database,” Language Resources and Evaluation, vol. 42, no. 4, pp. 335–359, 2008, doi: 10.1007/s10579-008-9076-6.