

Discrete Units, Adversarial Vocoding, and the Identity That Remains: The FBK-SpeechTek Systems for VoicePrivacy 2026

Siavosh Sepanta^{1,**}, Marco Matassoni¹, Alessio Brutti¹

¹ Fondazione Bruno Kessler (FBK), Trento, Italy

ssepanta@fbk.eu, matasso@fbk.eu, brutti@fbk.eu

Abstract

We describe the FBK-SpeechTek systems submitted to the VoicePrivacy 2026 Challenge (VPC), together with a diagnostic study of where their privacy comes from and what identity remains. The anonymizer keeps frozen, pre-trained content encoders; voice identity is controlled by a pool-sampled speaker embedding, emotion is preserved with a self-supervised affect embedding, pitch is conditioned explicitly, and a HiFi-GAN generator resynthesizes the waveform. For Track 1, a controlled localization study shows that most baseline privacy arises from neural resynthesis rather than pseudo-speaker selection, and that adversarially fine-tuning the vocoder against a speaker verifier (our submitted system) shows no gain over the unmodified baseline in the reported single-run comparison (9.1% vs. 9.6% semi-informed equal error rate, EER), although attacker fine-tuning variability on this track is too large for that comparison to be conclusive. For Track 2, replacing the continuous content path with discrete mHuBERT units (our submitted system) suppresses a timbre-reconstruction shortcut, yet does not raise privacy over the continuous alternative (2.5% vs. 4.5% four-language average EER, both below the 10% ranking minimum). Dissecting the residual leak, we find that adversarial fine-tuning against an exact, differentiable replica of the challenge’s WavLM attacker and inference-time removal of rhythm and pitch identity each yield privacy gains that persist after attacker retraining—up to a 6.5% four-language average, more than double the baseline’s 2.5%—though neither reaches the ranking minimum and both incur measurable WER costs. The remaining leakage is most consistent with the segmental content itself, with direct consequences for how anonymization systems should be designed and evaluated.

Index Terms: speaker anonymization, voice privacy, voice conversion, discrete units, adversarial training

1. Introduction

The VoicePrivacy Challenge (VPC) promotes speaker anonymization that hides the speaker identity while preserving linguistic content and, increasingly, paralinguistic attributes such as emotion [1, 2]. The 2026 edition introduces a multilingual track (Track 2, English/French/German/Spanish) alongside the established English track (Track 1), and ranks systems that meet a minimum equal error rate (EER) tier by word error rate (WER) and unweighted average recall (UAR) of emotion.

This paper describes the systems submitted by the FBK-SpeechTek team. Our design philosophy is to reuse strong pre-trained models and to add privacy where it is actually

needed. Concretely, we (i) build the content path from a frozen Whisper encoder followed by a small linear projector [3], (ii) anonymize the speaker through a pool of cosine-far reference embeddings, (iii) preserve emotion with a self-supervised affect embedding [4] and pitch with explicit F0 conditioning, and (iv) resynthesize with a HiFi-GAN vocoder [5].

We make four contributions. First, through a controlled “self-speaker” and embedding-swap analysis (Section 3) we show that most baseline privacy arises from neural resynthesis itself rather than from pseudo-speaker selection. Second, we report that adversarially fine-tuning the vocoder against a speaker verifier—our submitted primary Track 1 system—yields no EER gain in the single run we report, a comparison that Track 1’s attacker-training variability leaves inconclusive. Third, motivated by recent work on non-timbral identity leakage [6], our Track 2 submission replaces continuous content with a discrete-unit bottleneck, hypothesizing this would close a timbre-reconstruction shortcut; it does not raise privacy over the continuous system either, and neither clears the challenge’s 10% ranking minimum. Fourth, motivated by these results, we experimentally dissect the identity that remains through two further interventions: adversarial optimization against an exact replica of the deployed attacker, and removal of explicit source-derived rhythm and F0 cues. Both make speaker identity substantially less recoverable after attacker retraining, yet neither defeats the *retrained* semi-informed attacker outright, leaving considerable residual leakage that our remaining evidence points to as primarily segmental/phonotactic. This chain of negative-but-instructive results argues against trusting embedding-space proxies, fixed-attacker objectives, or single-pass evaluation pipelines when designing and reporting anonymization systems.

2. System overview

Both submitted systems share the same core inference pipeline in Figure 1, with different content front-ends and, for Track 1 only, an added training-time adversarial term (both described next). Given a 16 kHz utterance, three disentangled streams are extracted:

- **Content.** Track 1’s submitted system, and Track 2’s continuous candidate (Section 4), use the Whisper-large-v3-turbo encoder, whose frame embeddings a frozen MEUSLI linear projector [3] maps to a compact content representation Z . MEUSLI is a multilingual projector trained to interface Whisper with a speech LLM; we adopt its output as the content representation because it attenuates speaker information relative to the raw Whisper features while preserving linguistic and emotional content, and, like Whisper, it needs no change across languages. Track 2’s *submitted* system instead

^{**} indicates the corresponding author.

replaces this path with discrete mHuBERT-147 units, themselves multilingual by construction (Section 4.1).

- **Speaker.** An ECAPA-TDNN [7] extracts a 192-dimensional speaker embedding s . A pool of per-speaker reference embeddings (VoxCeleb1) is used to draw a pseudo-speaker embedding $\hat{s}$: for each utterance we sample among the N cosine-farthest pool speakers, optionally constrained to the same language for Track 2.
- **Prosody.** Pitch is tracked with a CREPE F0 estimator [8] and passed explicitly to the generator; emotion is captured by an emotion2vec embedding [4] that conditions the generator through adaptive instance normalization (AdaIN).

A content upsampler aligns the content stream (continuous Z or discrete units, depending on system) to the acoustic frame rate, and a HiFi-GAN generator conditioned on [content, F0] with AdaIN injection of $\hat{s}$ and the emotion embedding synthesizes a 24 kHz waveform that is resampled to the 16 kHz challenge format. The vocoder is trained on LibriSpeech (Track 1) and on a balanced multilingual mixture (Track 2) with multi-resolution STFT, mel, feature-matching and adversarial losses; Track 1’s submitted system additionally adds the speaker-adversarial term of Section 3 to this objective (Track 2’s submitted system does not; the same term is used only as a post-submission analysis variant there, Section 5). We refer to the Track 1 system, in either its baseline or adversarially fine-tuned form, as *Projector-Anon*.

3. Track 1: where the privacy comes from

3.1. Localizing the bottleneck

We first ask which component of the pipeline is responsible for the privacy the anonymizer provides. We distinguish two things that could be meant by “the speaker path”: *which* embedding is chosen (the pool-lookup strategy) versus *how* the vocoder uses whatever embedding it is given (the resynthesis mechanism itself, injected via AdaIN). Table 1 reports three lazy-informed (hereafter *lazy*) EER measurements on `libri-test` that separate these. (1) *Self-speaker resynthesis*: we condition the vocoder on the *source* speaker’s own embedding instead of a pool-sampled one [9], so the pool-lookup strategy is removed while the resynthesis mechanism is exercised exactly as normal; this still yields 13.5% EER, versus 2.2% for unprocessed original audio. (2) *Full pool-based system*: restoring the pool lookup on top of self-speaker resynthesis adds only 3.8 points (17.4%). (3) *Content-side subspace removal*: inspired by the disentanglement approach of Champion et al. [10], we instead leave the speaker-conditioning path untouched and remove a rank-20 linear speaker subspace from the content representation Z before resynthesis; on our architecture this drops a linear speaker-ID probe on Z from 14.5% to 2.3%, yet EER moves by only +0.6 points, because the vocoder reconstructs timbre from the residual content under the reconstruction objective—unlike [10], where content-side disentanglement was reported to change EER considerably, on our architecture it does not, most likely because the vocoder is not trained under the same disentanglement objective. Together, (1)–(3) show that most baseline privacy arises from neural resynthesis itself rather than from pseudo-speaker selection, and that this particular linear content-side erasure does not substantially improve it; we refer to the resynthesis mechanism as the *speaker-conditioning path* for the remainder of the paper.

Table 1: *Track 1 bottleneck localization, libri-test lazy EER (%)*; the frozen attacker is used throughout, which is why original speech reads 2.2 here and 3.4 under the semi-informed attacker of Table 2. Full procedural detail in Appendix B.

Configuration	EER
Original speech (no anonymization)	2.2
Self-speaker resynthesis (pool lookup removed)	13.5
Full system (pool lookup restored)	17.4
Content-subspace removal (speaker path untouched)	18.0

Table 2: *Track 1 results on libri-test (%)*; semi-informed attacker. Arrows indicate the favorable direction.

System	EER $\uparrow$	WER $\downarrow$	UAR $\uparrow$
Original speech [†]	3.4	1.8	71.1
Projector-Anon (baseline)	9.6	2.4	35.6
Proj.-Anon + adv. voc.	9.1	2.5	38.6

[†]Organizers’ published figure for original speech under the semi-informed attacker; the frozen-attacker value is 2.2 (Table 1).

3.2. Adversarial vocoder fine-tuning

Motivated by this localization, we add a speaker-adversarial term to the generator objective, in the spirit of gradient-based adversarial training [11]. During training the generator minimizes, in addition to the reconstruction and GAN losses, the cosine similarity between the ECAPA embedding of its output and that of the source speaker, pushing the synthesized voice away from the source in verifier space. The speaker verifier is frozen; gradients flow through it. Starting from the converged vocoder (epoch 155) we fine-tune for 30 epochs at a low learning rate. The resulting checkpoint (suffix `_adv_ep170_anon`) was submitted as our primary Track 1 system.

Its semi-informed EER is 9.1%, comparable to neighbouring checkpoints from the same sweep (ep155: 9.6%, ep175: 9.4%) rather than above them. Table 2 reports the comparison: adversarial fine-tuning against a frozen, cross-verifier (ECAPA, not the challenge’s WavLM) objective produces no measurable additional privacy at this operating point, with essentially unchanged WER (+0.1) and a 3-point UAR improvement. Two things prevent us from reading this as a null result rather than an absent measurement. Re-evaluating one Track 1 configuration three times gives 9.1%, 15.0% and 21.5% (Appendix G2), a span that dwarfs every difference within the sweep, so these single-run scores cannot rank neighbouring checkpoints. This run applied the adversarial term at $w=0.05$, a weight at which the gradient-norm probe of Appendix E shows it is swamped by the reconstruction objective, and its adversarial loss stays flat throughout training (Appendix C). The experiment is therefore uninformative about whether a *different*-architecture proxy transfers: the perturbation was too weak to test it. Section 5 returns to the question with an *exact* replica of the evaluation attacker, at a weight large enough to move the objective.

4. Track 2: continuous vs. discrete content

Because Whisper and the projector are multilingual, only the vocoder and the speaker pool are language-specific. Our first candidate retrained the HiFi-GAN on a balanced mixture of English (LibriSpeech), French, German and Spanish (Vox-

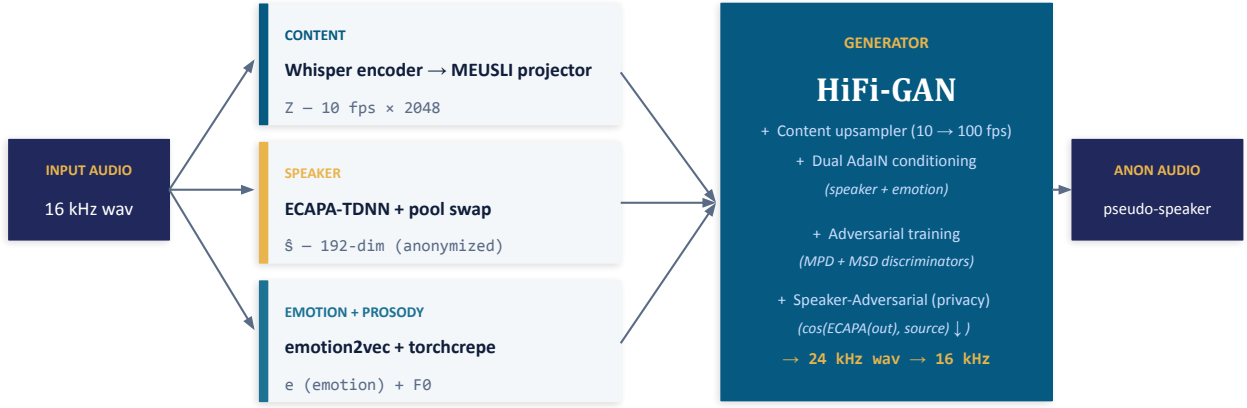


Figure 1: *Anonymization pipeline.* Content is supplied by a frozen Whisper encoder and MEUSLI projector for Track 1’s submitted system and Track 2’s continuous (non-submitted) candidate, or by discrete mHuBERT-147 units for Track 2’s submitted system (Section 4.1); voice identity by a pool-anonymized ECAPA embedding; emotion and pitch by emotion2vec and CREPE. A HiFi-GAN generator resynthesizes the waveform; the speaker-adversarial term shown is part of Track 1’s submitted system (Section 3) and, separately, a post-submission Track 2 analysis variant (Section 5)—it is not part of Track 2’s submitted system.

Populi) speech, keeping emotion and F0 conditioning and a multilingual pseudo-speaker pool. This *continuous-content* system anonymizes all four languages with good intelligibility, but its four-language average semi-informed EER is only 4.5% (Table 3)—below the 10% minimum condition. We attributed this to the timbre-reconstruction shortcut of Section 3.1: even though the speaker-conditioning path is where privacy is produced, a *continuous* content representation can still carry enough residual timbre for the vocoder to partly rebuild the source voice regardless of which embedding conditions it, bypassing that path rather than contradicting its role.

4.1. Discrete units: design and rationale

The fix, also adopted by recent challenge systems [12], is to replace continuous content with *discrete units*: a unit replaces the continuous acoustic representation with a finite category, preventing direct transmission of its continuous frame-level detail, strongly suppressing this reconstruction pathway. We implement this for the submitted system by replacing the projector path with mHuBERT-147 layer-9 features quantized to $K=1000$ multilingual k -means units [13], driving the same generator, while retaining emotion2vec and F0 conditioning so that affect is preserved [14] (we deliberately do *not* deduplicate units or rewrite durations, which would damage emotion).

An embedding-swap diagnostic in ECAPA space (anonymizing one utterance with several random pool targets and measuring output similarity) supports the mechanism: for the continuous system, outputs remain nearly target-invariant and stay closer to the source than to the target regardless of which target is used (within-swap similarity 0.994, similarity to source 0.21, to target 0.10); for the discrete-unit system, outputs become substantially more target-dependent (within-swap similarity drops to 0.62, similarity to source falls to ~ 0.00 , to target rises to 0.27). By this representation-level proxy, the discrete bottleneck strongly suppresses the timbre-reconstruction shortcut. A small-scale English-only pilot (train-clean-100, not the submitted multilingual system) is consistent with this under the *lazy* (frozen) attacker: EER rises from 2.2% (original) to

11.4%, alongside near-original WER (1.8%). We therefore expected the four-language average semi-informed EER to rise substantially above the continuous system’s 4.5%.

4.2. Results

The expectation did not materialize. The four-language average semi-informed EER is **2.5%** (en 3.4%, fr 2.5%, de 2.4%, es 1.8%; Table 3), far below the 10% ranking minimum. The toolkit correction does not affect WER or UAR, which depend only on the anonymized audio.

For Track 2, this is a clear negative result: the discrete-unit bottleneck strongly suppresses the timbre-reconstruction shortcut identified in Section 3.1 (Section 4.1 shows this directly, at the representation level), yet the four-language average EER (2.5%) does not improve over the continuous system’s (4.5%). Both fall well short of BM2 (10.6%, tier-10), though at far better utility (WER 8.3% vs. 27.7%; UAR 44.6% vs. 29.0%); neither of our Track 2 candidates is ranked. Suppressing this timbre-reconstruction shortcut therefore does not measurably improve privacy against the *retrained* attacker, indicating that substantial identity information remains available through other cues. Section 5 investigates what that leak is.

5. Dissecting the residual leak

Sections 3.1–4 leave two questions open: could the residual leak be reduced by training the anonymizer directly against the evaluation attacker, and are there other source-derived signals fed to the generator, besides content, that still carry identity? The generator is conditioned on a speaker embedding (already the target of the pool-swap and adversarial work above), an F0 contour, and an emotion embedding pulled fresh from each source utterance; we check the latter two as candidate carriers below. **Both interventions we tried substantially raise the semi-informed EER over the baseline, though neither reaches the 10% ranking minimum.** Table 4 summarizes.

Emotion. On the English pilot system of Section 4.1, we replaced each utterance’s own emotion embedding with a

Table 3: *Track 2 results (test): semi-informed EER (%) per language and four-language average (the ranked quantity) except Original, average WER (%), and emotion UAR (%). Baselines BM1–BM3 from the official evaluation plan [1].*

System	en	fr	de	es	avg ↑	WER ↓	UAR ↑
Original [†]	4.2	6.5	3.3	3.0	4.3	4.9	78.9
BM1	3.4	2.1	1.4	1.6	2.1	8.5	42.6
BM2	7.5	15.3	10.9	8.7	10.6	27.7	29.0
BM3	20.4	46.3	46.1	45.9	39.7	10.0	24.8
Continuous (ours)	4.5	5.5	4.2	3.7	4.5	—	37.2
Units (submitted)[‡]	3.4	2.5	2.4	1.8	2.5	8.3	44.6

[†]Original EER is a frozen-attacker (lazy) reference, not a semi-informed score. [‡]The organizers’ official figures for this system are 3.6/2.4/1.9/2.6, average 2.6; the small differences from our own re-evaluation reflect a different attacker fine-tuning run (Section 6).

Table 4: *Leak dissection (test): semi-informed EER (%) per language and four-language average (attacker retrained per variant), average lazy EER (%), frozen attacker), average WER (%), and emotion UAR (%). The 10% ranking minimum applies to the four-language semi-informed average.*

Variant	semi-informed EER					WER	UAR
	en	fr	de	es	avg		
Units (submitted)	3.4	2.5	2.4	1.8	2.5	10.4	8.3
+ adversarial (WavLM, $w=15$)	4.2	6.5	3.5	6.1	5.1	24.8	22.8
+ prosody removal (NT)	5.8	7.6	6.3	6.4	6.5	26.5	~14 [†]

[†]NT WER was measured on the English/German pilot only (14.8/14.0); its EERs cover all four languages. UAR was not measured for NT.

single, realistic reference embedding used for every utterance, while keeping the audio clearly intelligible. This left the system’s *lazy* EER unchanged (11.4% before, 11.7% after). We did not re-run the full semi-informed protocol (attacker fine-tuning) for this control, so unlike the other carriers reduced in this section, the explicit emotion-conditioning pathway is tested only under the frozen attacker; given how differently lazy and semi-informed EER can behave elsewhere in this paper, we report this as suggestive, not conclusive.

Adversarial training against the exact attacker. To ask whether the leak could be closed by attacking it directly, we built a *replica* of the challenge’s own WavLM–ECAPA verifier [15]—a second copy of the same architecture, initialized from the evaluation checkpoint’s weights, kept differentiable so that gradients can flow from a similarity score back to the waveform—and verified it reproduces the toolkit’s own embeddings to cosine similarity 1.0000. We then fine-tuned the unit vocoder with an added loss term, $w \cdot \cos(f(\hat{x}), f(x))$, where f is the frozen replica, $\hat{x}$ the anonymized output and x the source; minimizing this term pushes the output away from the source in verifier space. The gradient of this term updates only the vocoder’s trainable parameters (the unit front-end and HiFi-GAN generator), never f itself, exactly mirroring how the reconstruction and GAN losses already update the same parameters. A gradient-norm probe—measuring, on real training batches, the L^2 norm each loss term individually induces on those parameters—showed the reconstruction term dominates the adversarial one by $\sim 53\times$ at the weights ($w \leq 2$) used in prior attempts, explaining why they had no effect; we therefore trained at $w=15$, where the adversarial term reaches $\sim 30\%$

of the reconstruction gradient. The optimization objective was successfully minimized: the cosine similarity between output and source under the *frozen* replica collapsed from 0.86 to 0.07, and the average lazy EER more than doubled (Table 4). After the prescribed attacker fine-tuning, the semi-informed EER also rose substantially in every language, to a four-language average of 5.1%, roughly double the baseline’s 2.5% (Table 4). This gain survives attacker retraining: fine-tuning against an *exact*, differentiable replica of the evaluation attacker makes speaker identity substantially less recoverable even after attacker retraining, not merely an artifact of one frozen scorer. Track 1’s adversarial run does not supply the complementary negative case: it used $w=0.05$, orders of magnitude below the weight the gradient probe (Appendix E) identifies as necessary for the adversarial term to compete with reconstruction, and its objective never moved. Whether a different-architecture proxy would transfer at comparable pressure is untested. The retrained attacker is itself affected: its own EER on *original*, non-anonymized German speech rises to 8.5% (Appendix F), against 3.3% for the frozen verifier on the same audio, indicating some specialization to the heavily-warped adversarial training distribution—yet it still attacks the anonymized speech more effectively (EER 3.5%) than it attacks original speech (EER 8.5%). The gain still falls short of the 10% ranking minimum, and the WER cost is substantial (22.8% vs. 8.3%), so we are not claiming this would have changed our submission’s outcome—only that it is a real, previously-unrecognized privacy gain.

Removing prosodic identity. We then removed two explicit source-derived prosodic carriers at inference: consecutive duplicate units are collapsed and re-expanded to uniform durations (destroying the speaker’s rhythm and duration idiolect while approximately preserving utterance length), and the source F0 is replaced by a synthetic template (random base pitch, declination, slow wobble) carrying no source information—the same general information-removal principle as baseline BM3, which re-estimates prosody from transcriptions. The audio remains clearly intelligible (WER $\sim 14\%$), and the *frozen* attacker’s EER rises substantially (lazy EER, Table 4). The *retrained* attacker’s semi-informed EER rises as well in every language, to a four-language average of 6.5% (Table 4)—the largest gain of any intervention we tested. As with the adversarial intervention above, this is a real privacy gain that survives attacker retraining—though again short of the 10% ranking minimum and at measurable utility cost.

What remains. Direct transmission of continuous frame-level acoustic detail is removed by quantization by construction (Section 4.1): a discrete unit replaces the continuous acoustic frame with a finite category, so it no longer transmits that continuous-valued detail directly. But this representation-level removal only raises EER against the frozen baseline attacker (the swap diagnostic of Section 4.1; the English pilot’s lazy EER, 2.2%→11.4%)—the quantized system’s four-language semi-informed EER is essentially unchanged from the continuous system’s (Table 3), so removing this direct continuous-acoustic pathway alone does not measurably reduce the leak the retrained attacker actually exploits. The combined rhythm/pitch intervention and adversarial training are different: applied on top of the already-quantized system, both yield substantial privacy gains under the full semi-informed protocol (Table 4), showing that identity information remains exploitable beyond the direct continuous-acoustic pathway—but neither closes the gap to the 10% ranking minimum on its own, and the explicit emotion-conditioning pathway remains untested under

the full protocol (probed only under the weaker lazy-attacker control described above). What accounts for the remaining, still-dominant share of the leak is most likely the segmental content itself—which units are selected and in what sequence, potentially reflecting pronunciation, accent and other speaker-correlated segmental patterns—consistent with recent findings that k -means quantization of self-supervised features does not remove speaker information [16] and that accent carries re-identifiable cues [6]. We say “consistent with” rather than “proven”: attributing a residual across several partially-effective, correlated interventions evaluated on overlapping language subsets is inherently imprecise, and we did not control for recording-condition or channel artifacts specific to each language’s source corpus, which could in principle also carry some of the residual signal; a source-matched channel-normalization control is the natural next test. Under either explanation, the practical conclusion is the same: BM3, which discards the acoustic signal entirely in favor of transcription-level representations, is the only baseline family to defeat the 2026 attacker (39.7%)—and even it drops to 20% on English—but pays with the lowest emotion preservation in the table.

This methodological lesson generalizes beyond our own systems, but with one important exception. Three things can make a privacy result look better than it really is. First, checking how similar two things are in embedding space, instead of running the real attack, can be misleading. Second, relying only on a frozen attacker can mislead, because it does not test whether an apparent gain survives attacker adaptation. Third, running an evaluation pipeline only once, without re-checking it, can also give a wrong answer; we encountered all three (Sections 3 and 4). The exception concerns the frozen attacker: training against an exact copy of the real evaluation attacker did produce a real privacy gain (this section). Our Track 1 attempt, against a different architecture, did not—but it also applied far too little adversarial weight to move its own objective, so it does not settle whether the architecture or the weight was the limiting factor.

6. Submitted systems and reproducibility

Table 5 maps each submitted `$anon_data_suffix` to its system variant, as required by the challenge. All systems use frozen pre-trained encoders (Whisper-large-v3-turbo or mHuBERT-147 for content, ECAPA-TDNN for speaker, emotion2vec-large for affect, CREPE for pitch); only the vocoder and the content path differ. The analysis variants of Section 5 (adversarial $w=15$; prosody removal) were evaluated with the full official protocol but not submitted. Vcoders are trained on a single GPU, on which anonymization also runs faster than real time. The submitted audio itself is unchanged.

Recomputed evaluation. Every EER in this paper was recomputed with the official evaluation code after the organizers corrected two bugs in it (an embedding normalization applied over a size-one axis, and Track 1 trial lists missing cross-gender non-target pairs); in the same pass we replaced a non-standard batched embedding extractor in our own copy of the toolkit, which had bypassed the loudness normalization the reference implementation applies before feature extraction. Numbers from our pre-correction pipeline, including those in our submitted system description, are superseded by the values reported here. On the two submitted systems the recomputed figures agree with the organizers’ independent reproduction to within 0.4 (Track 1) and 0.13 (Track 2 average) percentage points, both within the run-to-run variation reported below.

Table 5: Mapping of submitted suffixes to system variants.

<code>\$anon_data_suffix</code>	System variant
<code>_adv_ep170_anon</code>	Track 1 primary: Projector-Anon with adversarial vocoder fine-tuning.
<code>_unit_ml_anon</code>	Track 2 primary: discrete mHuBERT-unit content bottleneck (en/fr/de/es).

Run-to-run variability. The semi-informed EER includes fine-tuning the attacker on the system’s own anonymized data, and the toolkit draws the 90/10 train/validation split of that data with an unseeded `random.shuffle`: repeated evaluations of one system therefore train on different subsets (their held-out sets share only $\sim 5\%$ of their members, a Jaccard overlap consistent with independent 10% draws). We re-evaluated the eight configurations for which we hold two or more independent fine-tuning runs (Appendix G2). Seven agree to within 0.20 points, so the metric is reproducible at the precision reported here. The exception is Track 1, whose three runs span 12.3 points (9.1/15.0/21.5); it is also the only configuration fine-tuned for two epochs rather than four to ten, though with one system at that setting we cannot separate the epoch count from the other differences between the tracks. Comparisons between individual Track 1 checkpoints should be read with that margin, and the organizers’ official scores, like ours, are single runs.

7. Conclusion

We presented anonymization systems built around frozen pre-trained content and speaker models, together with a diagnostic study of where their privacy comes from and what identity remains. A controlled localization study (Section 3.1) shows that most baseline privacy arises from neural resynthesis rather than pseudo-speaker selection, while the tested linear content-side erasure provides little additional benefit. Neither of two further interventions we tried on top of that baseline—adversarially fine-tuning the vocoder against a speaker verifier (Track 1, our submitted primary system) and replacing continuous content with discrete units to close a timbre-reconstruction shortcut (Track 2, our submitted primary system)—raises the semi-informed EER: ep170 does not improve on the unmodified baseline in the single run reported (9.1% vs. 9.6%), a comparison the Track 1 attacker-training variability of Appendix G2 leaves inconclusive, and the discrete-unit system does not improve over the continuous one (2.5% vs. 4.5%, both below the 10% ranking minimum that Track 2 therefore fails to clear). A post-submission dissection then tested further candidate carriers of the residual identity under the full semi-informed protocol: rhythm and pitch removal, and adversarial fine-tuning against an exact replica of the evaluation attacker, both substantially raise the four-language average EER over the baseline (up to 6.5% for prosody removal, versus 2.5%)—a gain that persists after attacker retraining in both cases—though neither reaches the 10% ranking minimum, and both incur WER degradation where measured (NT utility was evaluated on the English/German pilot only, and its UAR was not measured). Track 1’s adversarial run, by contrast, showed no measurable gain—but it applied the adversarial term at a weight far too small to move its own objective, so it does not establish that a *different*-architecture proxy fails to transfer; that comparison

remains open. Segmental content remains the most likely dominant carrier of what is left after these partial gains, though we cannot fully rule out channel or recording-condition confounds specific to each source corpus, nor a semi-informed effect of emotion conditioning that our lazy-only control would have missed. The performance of BM3—the only baseline family that beats the 2026 attacker—suggests that transcription-level re-synthesis may be one effective route to defeating it, though today it sacrifices exactly the paralinguistic information the challenge asks systems to preserve. Characterizing and closing the privacy–affect gap that a text bottleneck imposes is, we still believe, the central open problem the 2026 attacker exposes.

8. Acknowledgments

We thank the VoicePrivacy 2026 organizers. This work is funded by the European Union project ELOQUENCE (grant 101135916). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

9. Generative AI Use Disclosure

Generative AI tools were used to assist with code scaffolding, log analysis, and drafting/editing of this manuscript; all technical claims and results were verified by the authors.

10. References

- [1] 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, Evaluation Plan v2.0, 2026.
- [2] 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,” 2024.
- [3] L. Concina, S. Fong, M. Matassoni, and A. Brutti, “MEUSLI: a multilingual projector for LLM-based ASR and beyond,” 2026.
- [4] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, “emotion2vec: Self-supervised pre-training for speech emotion representation,” in *Findings of the Association for Computational Linguistics (ACL)*, 2024.
- [5] 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 (NeurIPS)*, 2020.
- [6] R. Bakari, O. L. Blouch, N. Gengembre, N. Evans, and M. Panariello, “Identity leakage through accent cues in voice anonymization,” in *Proc. Speaker Odyssey*, 2026.
- [7] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification,” in *Proc. Interspeech*, 2020.
- [8] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, “CREPE: A convolutional representation for pitch estimation,” in *Proc. ICASSP*, 2018, pp. 161–165.
- [9] M. Costante, M. Matassoni, and A. Brutti, “Using Seq2seq voice conversion with pre-trained representations for audio anonymization: Experimental insights,” 2022.
- [10] P. Champion, D. Jouviet, and A. Larcher, “Are disentangled representations all you need to build speaker anonymization systems?” in *Proc. Interspeech*, 2022.
- [11] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in *Proc. International Conference on Machine Learning (ICML)*, 2015.
- [12] 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,” 2024.
- [13] M. Z. Boito, V. Iyer, N. Lagos, L. Besacier, and I. Calapodescu, “mHuBERT-147: A compact multilingual HuBERT model,” in *Proc. Interspeech*, 2024.
- [14] J. Yao, H. Liu, E.-S. Chng, and L. Xie, “EASY: Emotion-aware speaker anonymization via factorized distillation,” *arXiv preprint arXiv:2505.15004*, 2025.
- [15] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, “WavLM: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [16] S.-L. Yeh and H. Tang, “Estimating the completeness of discrete speech units,” in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2024.