

PSST1 System description for VoicePrivacy Challenge 2026

Théo Nguyen^{1,2}, Ben Luks^{3,4,5}, Tom Bäckström¹, Rainer Martin²

¹ Aalto University, Finland

² Ruhr-Universität Bochum, Germany

³ INESC-ID, Portugal

⁴ Instituto Superior Técnico, University of Lisbon, Portugal

⁵ Technische Universität Berlin, Germany

theo.nguyen@aalto.fi, benluks@inesc-id.pt, tom.backstrom@aalto.fi, rainer.martin@rub.de

Abstract

We present the PSST1 system for the VoicePrivacy Challenge 2026. Our system builds upon the SSL-SAS emotion-preserving anonymization framework by replacing its orthogonal Householder neural network (OHNN) speaker anonymizer with a flow-matching-based speaker embedding generator (FM). We further investigate a conditional flow-matching (CFM) variant that incorporates emotion labels during speaker generation and evaluate its synergy with the original downstream emotion compensation module. Replacing the OHNN anonymizer with FM improves both speaker privacy and emotion preservation while maintaining comparable ASR performance. Our results indicate that standalone FM and CFM perform roughly on par. Furthermore, while pairing CFM with emotion compensation achieves the best overall results, this joint approach offers only a minor, non-significant improvement over the unconditional FM system with emotion compensation.

Index Terms: speech anonymization, speaker representation, conditional flow-matching, VoicePrivacy Challenge

1. Introduction

Voice anonymization seeks to conceal speaker identity while preserving the linguistic content and usability of speech. Many contemporary anonymization systems achieve this through representation disentanglement [1, 2, 3]. This technique explicitly separates the source speech into distinct speaker, content, and prosodic components, allowing the system to alter the speaker representation before reconstructing the anonymized waveform.

To mask identity, existing systems handle the generation of the anonymized speaker embedding in a few distinct ways. Some methods select replacement speaker embeddings from an external speaker pool [4, 5, 6] or from a finite set of predefined pseudo-speaker identities [7, 8, 9]. While effective, these approaches are inherently limited to a finite set of target identities and rely on external speaker representations. Other methods instead modulate original speaker embeddings through mapping-based transformation, including adversarially trained autoencoders [10, 11] and orthogonal Householder neural networks (OHNNs) [12]. Finally, generative approaches such as Variational Autoencoders (VAE) [13], Generative Adversarial Networks (GAN) [14], or flow-matching (FM) [15] can directly sample anonymized speaker embeddings from a learned distribution. By generating novel pseudo-speakers on the fly, these methods bypass the constraints of finite identity pools, thereby eliminating the risk of linkability attacks where an attacker correlates multiple anonymized sessions back to the same reused identity.

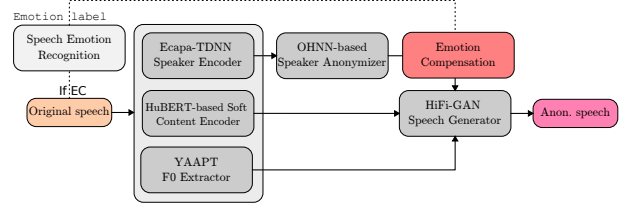


Figure 1: Architecture of SSL-SAS with OHNN, reproduced from [17]

Aligned with this shift toward fully generative modeling, we introduce an FM-based speaker generator into an emotion-preserving voice anonymization framework. While FM has shown promise for generating high-quality synthetic embeddings, its interaction with explicit downstream utility constraints, such as emotion preservation, remains largely unexplored. Specifically, we replace a deterministic, mapping-based OHNN speaker anonymizer with a continuous FM generator. We further propose a conditional FM (CFM) variant that integrates emotion labels directly into the generation path, exploring how generative speaker spaces interact with explicit emotion compensation (EC) modules.

We evaluate our proposed systems on the VoicePrivacy Challenge 2026 [16] benchmark to analyze the trade-offs between speaker privacy, speech utility, and emotion preservation. Our core contributions are as follows:

- **Generative Anonymization:** We integrate an FM speaker generator into an emotion-preserving framework, demonstrating that generative methods outperform mapping-based transformations in both privacy and utility.
- **Emotion Conditioning:** We evaluate a CFM variant to assess the impact of embedding emotional attributes directly into the generated speaker space.

2. Method

2.1. Overview

Our anonymization system builds upon the SSL-SAS voice anonymization framework proposed by [17], presented in Figure 1. We replace the original OHNN speaker anonymizer with an FM-based speaker embedding generator based on [15], as shown in Figure 2. The remainder of the synthesis pipeline is kept unchanged to isolate the effect of the speaker anonymization module.

Given an input speech waveform, three representations are extracted. Linguistic content is represented using 200-

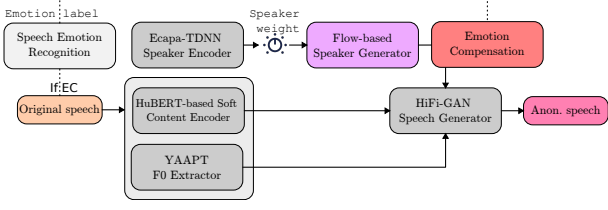


Figure 2: *Speech anonymization architecture with flow-based speaker generator. Note that the speaker encoder is not needed if the speaker weight w is null*

dimensional HuBERT-soft [18] units, speaker identity is represented using a 192-dimensional x-vector extracted from a pre-trained ECAPA-TDNN [19] model, and prosody is represented using the fundamental frequency (F0) estimated using YAAFT [20].

The extracted speaker embedding is anonymized before being combined with the content representation and F0 for waveform synthesis. An optional emotion compensation stage may then be applied to the generated speaker embedding. Finally, the anonymized speaker embedding, together with the HuBERT-soft units and F0 contour, is provided to a HiFi-GAN [21] vocoder to synthesize the anonymized speech waveform. The only component modified between experimental conditions is the speaker anonymization module.

2.2. Baseline OHNN

As a baseline, we evaluate the OHNN anonymizer proposed by Miao et al. [17]. It functions as a deterministic mapping network that acts as a learnable rotation operator. By transforming the source ECAPA-TDNN x-vector through a cascade of layers parameterized by Householder matrices, the model produces an anonymized speaker embedding that is then provided directly to the downstream HiFi-GAN vocoder.

2.3. Flow-matching Speaker Embedding Generation

The speaker anonymization component is responsible for replacing the original speaker embedding with an anonymized embedding while remaining compatible with the downstream HiFi-GAN synthesis model. As an alternative to the mapping-based OHNN anonymizer used in the baseline system [17], we investigate an FM speaker generator based on [15], together with a CFM variant proposed in this work.

FM is a generative paradigm designed to regress a continuous-time vector field $v_t(x)$ that moves samples from a simple noise distribution $x_0 \sim \mathcal{N}(0, I)$ toward a target data distribution $p_1(x_1)$ [22]. To achieve this, a time-dependent conditional probability path $p_t(x|x_1)$ is constructed given a target data point x_1 , which is generated by a target conditional vector field $u_t(x|x_1)$.

To model this vector field, our velocity estimator adopts the architecture proposed by Xiao et al. [15]. As illustrated in Figure 3, the network utilizes a U-Net-like encoder-decoder structure designed to process 192-dimensional speaker embeddings. The scalar timestep t is first mapped to a conditioning vector via a timestep embedder *TimeMLP*. Inspired by the Diffusion Transformer (DiT) framework [23], each processing block is then conditioned on this resulting timestep embedding via adaptive layer normalization (AdaLN). However, because the input consists of fixed-dimensional speaker embeddings rather than

sequential data, the standard self-attention layers are replaced by multilayer perceptrons (MLPs).

The CFM model extends the FM architecture by introducing an emotion conditioning pathway. To achieve this, the categorical emotion label is passed through a dedicated multilayer perceptron, denoted as *EmoMLP*, which projects the label into an embedding space matching the dimensionality of *TimeMLP*. The resulting emotion and timestep embeddings are concatenated to form the conditioning vector, which is provided to every DiT block of the velocity estimator through the AdaLN.

2.3.1. Training

The FM speaker generator is trained to model the distribution of 192-dimensional speaker embeddings extracted from LibriSpeech *train-clean-360* using a frozen, pretrained ECAPA-TDNN model. Prior to training, the extracted x-vectors are standardized using the empirical mean and standard deviation of the training set.

The network is optimized using the conditional flow-matching objective defined by [22], minimizing the mean squared error between the predicted and target velocity fields. For the CFM variant, categorical emotion labels are assumed to be available during both training and inference to guide the trajectory generation.

2.3.2. Inference

During inference, the anonymization pipeline follows an encode-interpolate-decode sequence. First, the source speaker embedding is standardized and mapped to the latent noise space by integrating the learned vector field in the reverse time direction (from $t = 1$ to $t = 0$) using a first-order Euler ODE solver with $N = 10$ steps. Let $\mathbf{z}_{\text{orig}}$ denote this inverted latent representation. To introduce anonymity, a random noise vector $\mathbf{z}_{\text{rand}} \sim \mathcal{N}(0, \mathbf{I})$ is sampled and blended with $\mathbf{z}_{\text{orig}}$ via the variance-preserving interpolation, as in [15]:

$$\mathbf{z}_{\text{anon}} = \frac{(1-w)\mathbf{z}_{\text{rand}} + w \cdot \mathbf{z}_{\text{orig}}}{\sqrt{(1-w)^2 + w^2}}, \quad (1)$$

where $w \in [0, 1]$ denotes the speaker weight. This parameter controls the identity trade-off: approaching 1 preserves the original speaker’s traits, whereas approaching 0 yields a fully synthetic target identity. Note that when $w = 0$, the speaker encoder is not needed in the anonymization process.

The interpolated latent representation $\mathbf{z}_{\text{anon}}$ is subsequently mapped back to the target speaker embedding space by integrating the vector field in the forward time direction (from $t = 0$ to $t = 1$). Finally, this generated embedding is restored to the original feature scale using the stored training set statistics. The CFM model utilizes an identical inference procedure, with the sole exception that the velocity estimator is additionally conditioned on the target emotion label vector throughout both the forward and reverse integration paths.

2.4. Emotion Compensation

To further improve emotion preservation, we evaluate the emotion compensation strategy proposed by Miao et al. [17]. Rather than modifying the speech synthesis pipeline, emotion compensation operates directly on the anonymized speaker embedding prior to waveform generation.

For each emotion category, an emotion-specific support vector machine (SVM) is trained to estimate a decision boundary in the speaker embedding space. Following the setup in

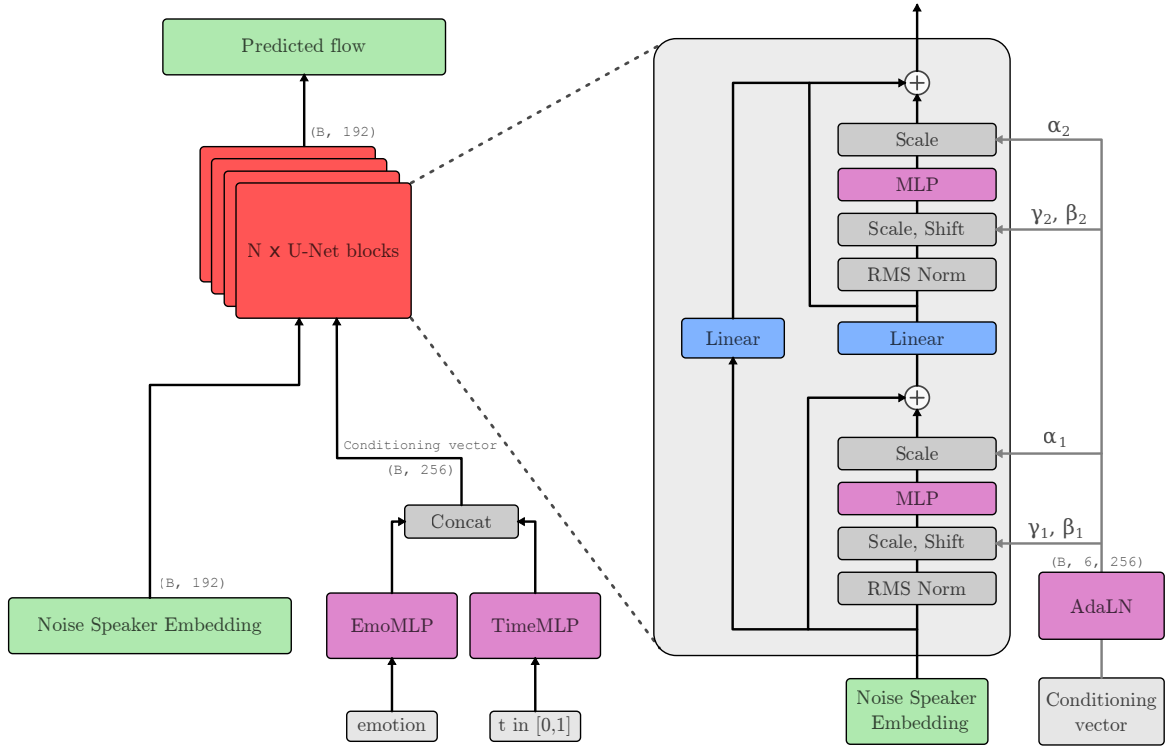


Figure 3: Architecture of the proposed conditional flow-matching (CFM) speaker generator.

Table 1: Overview of the evaluated anonymization systems.

System	Speaker Generator	Emotion Compensation
OHNN	OHNN	✗
OHNN + EC	OHNN	✓
FM	FM	✗
FM + EC	FM	✓
CFM	CFM	✗
CFM + EC	CFM	✓

[17], the emotion compensation module is trained on the MSP-IMPROV and ESD datasets. During inference, the generated speaker embedding is translated along the normal vector of the corresponding emotion boundary, thereby increasing the confidence of the target emotion while preserving the anonymized speaker identity.

Emotion compensation is applied as a post-processing step to the anonymized speaker embeddings and is used in combination with both the OHNN baseline and the proposed FM and CFM speaker generators.

3. Experimental Setup

To evaluate the contribution of the proposed speaker embedding generation strategy, we compare six anonymization configurations, described in Table 1. All remaining components of the anonymization pipeline, including feature extraction, waveform synthesis, and evaluation protocol, are kept unchanged across experiments.

3.1. Datasets and Evaluation Benchmarks

Following Track 1 (English) of the VoicePrivacy Challenge 2026 [16], we utilize standard speech corpora for both model training and utility-privacy benchmarking.

- **LibriSpeech:** The *train-clean-360* subset (~360 hours, 921 speakers) [24] is used for training the speaker generators. Privacy and ASR utility are benchmarked on the official challenge enrollment and trial splits derived from *dev-clean* and *test-clean*.
- **IEMOCAP:** Used exclusively to evaluate emotion preservation utility [25]. The models are evaluated using the official challenge test splits.

Emotion labels for the CFM and emotion compensation pipelines are collected via two ways. Official ground-truth emotion labels are used for IEMOCAP, whereas the LibriSpeech *train-clean-360* subset is annotated with emotion pseudo-labels extracted via a pretrained SpeechBrain [26] Speech Emotion Recognition (SER) model¹. We acknowledge that using an IEMOCAP-fine-tuned SER checkpoint introduces an out-of-rule dataset dependency for pseudo-labeling LibriSpeech. In an ideal setup, a self-supervised foundation model trained, such as *emotion2vec* [27], fine-tuned only on compliant 4-class subsets, should be used to prevent cross-dataset contamination.

3.2. Evaluation Metrics

Following the VPC2026 protocol, we measure performance in three ways:

- **Privacy:** Evaluated with the Equal Error Rate (EER) computed by a pretrained ASV (Automatic Speaker Verification)

¹<https://huggingface.co/speechbrain/emotion-recognition-wav2vec2-IEMOCAP>

Table 2: *VoicePrivacy Challenge Track 1 results on the evaluation set. Higher EER and UAR indicate better privacy and emotion preservation; lower WER indicates better linguistic utility.*

System	EER $\uparrow$	UAR $\uparrow$	WER $\downarrow$
OHNN	6.74	47.47	2.47
OHNN + EC	8.46	55.45	2.45
FM	8.02	53.84	2.40
FM + EC	8.52	58.59	2.45
CFM	8.22	53.62	2.34
CFM + EC	8.53	58.42	2.43

model, where higher values denote better privacy. Evaluations are primarily conducted against a lazy-informed attacker. Due to extreme computational cost, results against a semi-informed attacker are provided only for our highest-scoring model, CFM+EC. Due to substantial training overhead, reported metrics reflect single-run configurations. Due to the computational resources required for multi-seed training, this work focuses on single-run evaluations.

- **Intelligibility:** Measured via the Word Error Rate (WER) of a pretrained Automatic Speech Recognition (ASR) engine, where lower values signify better intelligibility.
- **Emotion Preservation:** Evaluated via the Unweighted Average Recall (UAR) of the pretrained SpeechBrain SER model on the anonymized IEMOCAP evaluation subsets.

3.3. Training Configuration

The flow-matching architectures are optimized to model the 192-dimensional ECAPA-TDNN speaker embedding space. Prior to training, the source x-vectors are standardized using the empirical mean and standard deviation of the training set.

Models are trained for 300 epochs using the AdamW optimizer with a OneCycle learning rate schedule, a batch size of 128, and a peak learning rate of 2×10^{-4} . The conditioning projection dimensionality is set to 256. Separate models are trained for enrollment and trial anonymization using different random seeds.

3.4. Inference Configuration

During the inference phase, ODE integration is performed using a first-order Euler solver with $N = 10$ steps. The speaker weight parameter w was empirically set to 0.25 based on preliminary validation experiments. This value offers an optimal trade-off, as higher weights improve utility at the expense of speaker privacy.

3.5. Computational Complexity

Our trainable CFM generator is lightweight, with 1.64M parameters (6.27 MB). During inference, using $N = 10$ Euler integration steps, processing 2,746 test utterances on a single NVIDIA Tesla V100 GPU took 252 seconds. This corresponds to an average latency of 92 ms per utterance or 11 utterances/sec. This shows that our privacy improvements come with minimal runtime overhead.

4. Results

Overall, the flow-based systems improve the privacy-utility trade-off relative to the OHNN baseline, as shown in Figure 4. The best privacy score is obtained by CFM + EC, while the best emotion preservation is obtained by FM + EC, although the differences between the two are marginal. WER remains similar across all anonymized systems, indicating that the speaker embedding anonymization strategy has little effect on linguistic content preservation.

4.1. Replacing OHNN with flow-matching

Replacing the OHNN anonymizer with the FM speaker generator improves both privacy and emotion preservation. Compared with OHNN, FM increases EER from 6.74% to 8.02% and UAR from 47.47% to 53.84%, while slightly reducing WER from 2.47% to 2.40% as displayed in Table 2. This suggests that flow-based speaker embedding generation is a viable alternative to deterministic speaker transformation in the SSL-SAS framework.

Explicit emotion conditioning provides little benefit over the unconditional baseline. FM and CFM perform similarly regarding privacy and UAR, and while CFM shows a slight advantage in WER, the difference is negligible. These results suggest that, under the current architecture, explicit emotion conditioning does not significantly alter the generated speaker embeddings.

4.2. Effect of Emotion Compensation

Emotion compensation improves emotion preservation for both the OHNN and flow-based systems. For OHNN, adding emotion compensation increases UAR from 47.47% to 55.45% and also improves EER from 6.74% to 8.46%. For FM, emotion compensation increases UAR from 53.84% to 58.59%, giving the best emotion preservation among all evaluated systems. The privacy gain for FM + EC is smaller, increasing EER from 8.02% to 8.52%. The CFM + EC system obtains the highest EER, 8.53%, while maintaining a UAR of 58.42%. Thus, emotion compensation remains beneficial when combined with flow-based speaker generation.

4.3. Discussion

The proposed conditioning mechanism does not produce a measurable improvement over the unconditional FM model. This observation is consistent with the visualization of the generated speaker embeddings shown in Figure 5, where embeddings generated by CFM remain largely intermixed across emotion classes despite explicit conditioning. This suggests that the current conditioning pathway does not substantially alter the geometry of the learned speaker embedding space. A possible explanation is that the conditioning signal is too weak relative to the dominant variation introduced by the latent interpolation and the flow objective. Alternatively, the ECAPA-TDNN speaker embeddings may not encode emotion strongly enough for the conditional generator to exploit the provided labels effectively.

The limited difference between FM and CFM suggests that the current conditioning mechanism does not substantially affect the generated speaker embeddings. Figure 5 visualizes generated speaker embeddings using t-SNE before and after emotion compensation. Without emotion compensation, embeddings remain largely intermixed across emotion classes despite explicit conditioning. After emotion compensation, clear emotion-specific regions emerge, suggesting that the SVM-

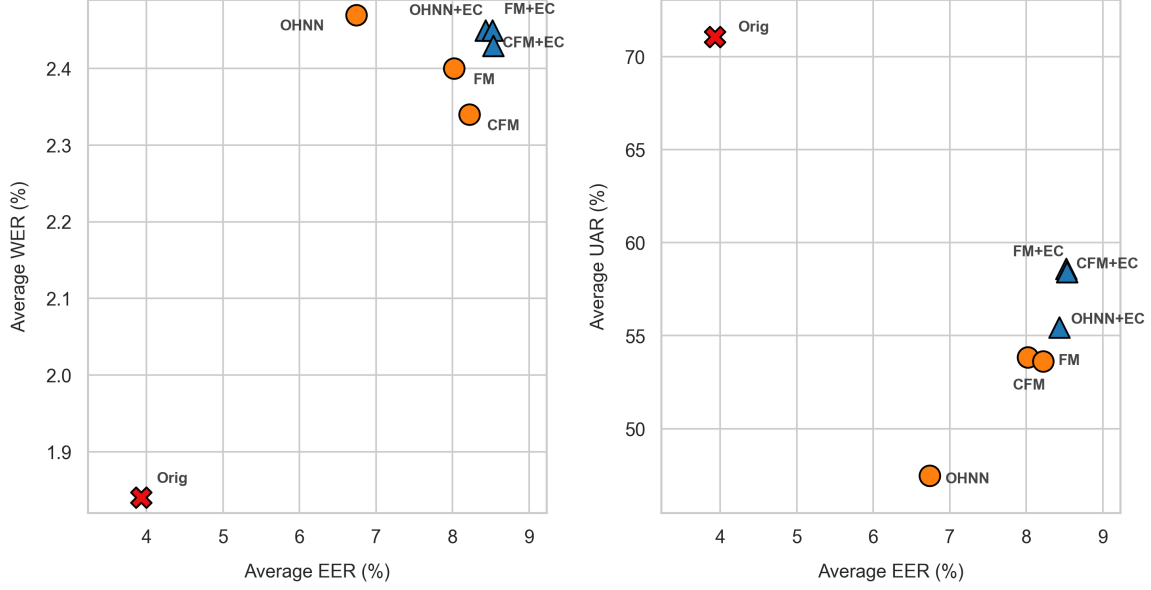


Figure 4: Privacy-utility trade-offs across anonymization systems. Left: WER vs. EER. Right: UAR vs. EER. The red cross represents the results on unprotected data, the orange circles the different systems without EC, and the blue triangles the systems with EC.

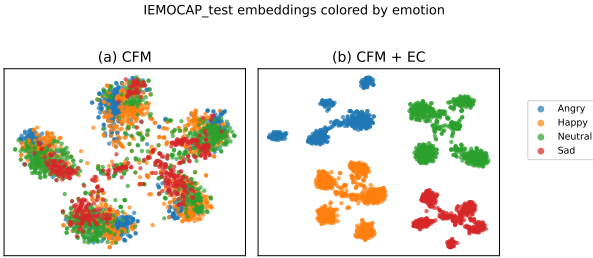


Figure 5: *t*-SNE visualization of speaker embeddings generated by the conditional flow-matching model before (left) and after (right) emotion compensation, colored by emotion labels.

based post-processing has a substantially stronger effect on the embedding geometry than the current conditioning mechanism.

This may indicate that the emotion signal is weakly exploited by the velocity estimator, or that the interpolation with the original latent representation dominates the effect of emotion conditioning. In contrast, emotion compensation applies an explicit displacement in speaker embedding space along emotion-specific SVM directions, which may provide a more direct intervention for preserving emotional information. Future work should investigate stronger conditioning mechanisms and disentangle the relative contributions of latent interpolation and explicit emotion compensation.

Emotion compensation consistently improves both emotion preservation and speaker privacy. A possible explanation is that the latent interpolation used by FM and CFM retains part of the structure of the original speaker embedding through the reverse-mapped latent representation. While this residual information appears beneficial for preserving emotional characteristics, it may also retain speaker-specific information. By explicitly translating the generated embedding along emotion-specific SVM directions, emotion compensation reinforces emotion-discriminative information while simultane-

ously moving the embedding away from regions associated with the original speaker representation. This may explain the simultaneous improvements in UAR and EER observed across the evaluated systems.

4.4. Semi-Informed Scenario

To test our system CFM+EC under the strongest threat conditions, we evaluate it following the VPC2026 evaluation protocol under a semi-informed scenario against DAST [28], an ASV attacker fine-tuned directly on anonymized speech. Under this model, our system achieves an EER of 2.53% on the LibriSpeech test mixed split, down from a baseline of 8.53% under the lazy-informed scenario.

This low result highlights an important architectural boundary condition: our framework only changes the speaker embedding (*timbral feature*) to mask identity. *Non-timbral* features like the F0 contour and overall prosody are kept unaltered to maintain speech utility. We suspect this lower EER comes from DAST exploiting unmapped prosodic cues rather than a breakdown of our timbre anonymizer, offering a potential explanation to be explored in future work.

5. Conclusion

We presented the PSST1 submission to the VoicePrivacy Challenge 2026, in which the mapping-based OHNN speaker anonymizer of the SSL-SAS framework was replaced with a flow-matching-based speaker embedding generator. Experimental results demonstrated that flow-based speaker generation consistently improves speaker privacy and emotion preservation over the deterministic baseline while maintaining comparable ASR performance. Although the proposed conditional flow-matching model did not provide measurable improvements over the unconditional variant, combining flow-based generation with the existing emotion compensation strategy achieved the best overall trade-off between speaker privacy and emotion preservation, with minor degradation to WER.

Future work should investigate conditioning mechanisms that more strongly couple emotion labels to the learned vector field. In the present system, emotion is injected by concatenating an emotion embedding with the timestep embedding, but the resulting embeddings remain weakly organized by emotion. Stronger mechanisms such as emotion-specific adaptive normalization, classifier-free guidance [29], or auxiliary emotion-classification losses may encourage the generator to use the conditioning signal more effectively. Future work should also investigate the influence of acoustic conditions, such as channel characteristics and recording environments, as potential sources of information leakage in speaker embeddings.

6. Acknowledgment

This work was funded by the European Union’s Horizon Europe research and innovation programme grant No 101168193.

7. References

- [1] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. No  , and M. Todisco. (2022) The VoicePrivacy 2020 Challenge Evaluation Plan. [Online]. Available: <http://arxiv.org/abs/2205.07123>
- [2] 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. [Online]. Available: <https://ieeexplore.ieee.org/document/10603395/>
- [3] N. Tomashenko, X. Miao, P. Champion, S. Meyer, M. Panariello, X. Wang, N. Evans, E. Vincent, J. Yamagishi, and M. Todisco. (2026) The Third VoicePrivacy Challenge: Preserving Emotional Expressiveness and Linguistic Content in Voice Anonymization. [Online]. Available: <http://arxiv.org/abs/2601.11846>
- [4] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, “Speaker Anonymization Using X-vector and Neural Waveform Models,” in *10th ISCA Workshop on Speech Synthesis (SSW 10)*. ISCA, 2019, pp. 155–160. [Online]. Available: https://www.isca-archive.org/ssw.2019/fang19_ssw.html
- [5] P. Champion, A. Larcher, and D. Jouv  t, “Are disentangled representations all you need to build speaker anonymization systems?” in *Interspeech 2022*. ISCA, 2022, pp. 2793–2797. [Online]. Available: https://www.isca-archive.org/interspeech.2022/pierre22_interspeech.html
- [6] M. Panariello, F. Nespoli, M. Todisco, and N. Evans, “Speaker Anonymization Using Neural Audio Codec Language Models,” in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 4725–4729. [Online]. Available: <https://ieeexplore.ieee.org/document/10447871/>
- [7] B. M. L. Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, “Evaluating Voice Conversion-Based Privacy Protection against Informed Attackers,” in *ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2802–2806. [Online]. Available: <https://ieeexplore.ieee.org/document/9053868/>
- [8] I.-C. Yoo, K. Lee, S. Leem, H. Oh, B. Ko, and D. Yook, “Speaker Anonymization for Personal Information Protection Using Voice Conversion Techniques,” *IEEE Access*, vol. 8, pp. 198 637–198 645, 2020. [Online]. Available: <https://ieeexplore.ieee.org/document/9247219/>
- [9] P. Champion, “Anonymizing Speech: Evaluating and Designing Speaker Anonymization Techniques,” Ph.D. dissertation, Universit   de Lorraine (LORIA), Le Mans Universit   (LIUM), 2023. [Online]. Available: <http://arxiv.org/abs/2308.04455>
- [10] J. M. Perero-Codosero, F. M. Espinoza-Cuadros, and L. A. Hern  ndez-G  mez, “X-vector anonymization using autoencoders and adversarial training for preserving speech privacy,” *Computer Speech & Language*, vol. 74, p. 101351, 2022. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S088523082200002X>
- [11] H. Turner, G. Lovisotto, and I. Martinovic, “Generating identities with mixture models for speaker anonymization,” *Computer Speech & Language*, vol. 72, p. 101318, 2022. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0885230821001133>
- [12] X. Miao, X. Wang, E. Cooper, J. Yamagishi, and N. Tomashenko, “Speaker Anonymization Using Orthogonal Householder Neural Network,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3681–3695, 2023. [Online]. Available: <https://ieeexplore.ieee.org/document/10244064/>
- [13] F. Huang, K. Zeng, and W. Zhu, “DiffVC+: Improving Diffusion-based Voice Conversion for Speaker Anonymization,” in *Interspeech 2024*. ISCA, 2024, pp. 4453–4457. [Online]. Available: https://www.isca-archive.org/interspeech.2024/huang24_interspeech.html
- [14] S. Meyer, P. Tilli, P. Denisov, F. Lux, J. Koch, and N. T. Vu, “Anonymizing Speech with Generative Adversarial Networks to Preserve Speaker Privacy,” in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 912–919. [Online]. Available: <https://ieeexplore.ieee.org/document/10022601/>
- [15] Y. Xiao, Y. Zhao, Z. Ma, S. Wang, K. Yu, J. Liao, and X. Chen. (2026) Anonymization, Not Elimination: Utility-Preserved Speech Anonymization. [Online]. Available: <http://arxiv.org/abs/2604.17000>
- [16] X. Miao, N. Tomashenko, R. Arefeen, S. Meyer, X. Wang, E. Vincent, N. Evans, J. Yamagishi, and M. Todisco. (2026) VoicePrivacy 2026 Challenge Evaluation Plan v0.0. [Online]. Available: https://www.voiceprivacychallenge.org/vp2026/docs/VPC.2026_march15.pdf
- [17] X. Miao, Y. Zhang, X. Wang, N. Tomashenko, D. Soh, and I. McLoughlin. (2024) Adapting General Disentanglement-Based Speaker Anonymization for Enhanced Emotion Preservation. [Online]. Available: <https://www.ssrn.com/abstract=4937233>
- [18] B. Van Niekerk, M.-A. Carbonneau, J. Za  idi, M. Baas, Hugo Seut  , and H. Kamper, “A Comparison of

- Discrete and Soft Speech Units for Improved Voice Conversion,” in *ICASSP 2022 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6562–6566. [Online]. Available: <https://ieeexplore.ieee.org/document/9746484>
- [19] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification,” in *Interspeech 2020*. ISCA, 2020, pp. 3830–3834. [Online]. Available: https://www.isca-archive.org/interspeech_2020/desplanques20_interspeech.html
- [20] K. Kasi and S. A. Zahorian, “Yet Another Algorithm for Pitch Tracking,” in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2002, pp. I–361–I–364. [Online]. Available: <https://ieeexplore.ieee.org/document/5743729/>
- [21] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/hash/c5d736809766d46260d816d8dbc9eb44-Abstract.html
- [22] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, “Flow Matching for Generative Modeling,” in *The Eleventh International Conference on Learning Representations - ICLR 2023*, 2022. [Online]. Available: <https://openreview.net/forum?id=PqvMRDCJT9t>
- [23] W. Peebles and S. Xie, “Scalable Diffusion Models with Transformers,” in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4172–4182. [Online]. Available: <https://ieeexplore.ieee.org/document/10377858>
- [24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210. [Online]. Available: <http://ieeexplore.ieee.org/document/7178964/>
- [25] 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. [Online]. Available: <https://doi.org/10.1007/s10579-008-9076-6>
- [26] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio. (2021) SpeechBrain: A General-Purpose Speech Toolkit. [Online]. Available: <http://arxiv.org/abs/2106.04624>
- [27] 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*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 15 747–15 760. [Online]. Available: <https://aclanthology.org/2024.findings-acl.931/>
- [28] R. Arefeen, X. Miao, R. Tong, A. B. Ng, S. See, and T. Liu. (2026) DAST: A Dual-Stream Voice Anonymization Attacker with Staged Training. [Online]. Available: <http://arxiv.org/abs/2603.12840>
- [29] J. Ho and T. Salimans, “Classifier-Free Diffusion Guidance,” in *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. [Online]. Available: <https://openreview.net/forum?id=qw8AKxfYbI>