



A Training-Free Projection-Guided Conditional Flow-Matching System for Voice Privacy 2026

Xiang Shi, Han Zhu, Ming Li, Xiaoxiao Miao T23 DKU



Seed VC v2: the backbone

Pretrained voice conversion model with two generation stages.

Stage 1: Condition Construction

The AR model constructs tokens for the source content.

Stage 2: Speech Generation

A flow-matching model generates the output speech.

Built-in anonymization removes target conditioning and uses token guidance.



Proposed method: two training-free interventions

Stage 1: Random-target-conditioned AR tokens

Regenerate the source-content token representation under an independently sampled speaker context.

Stage 2: Source-conditioning-aware projected CFG

Use source prompt/style only as a reference direction, then remove the component of token guidance aligned with it.

No new model, no retraining, no parameter update — only inference-time conditioning and guidance are changed.

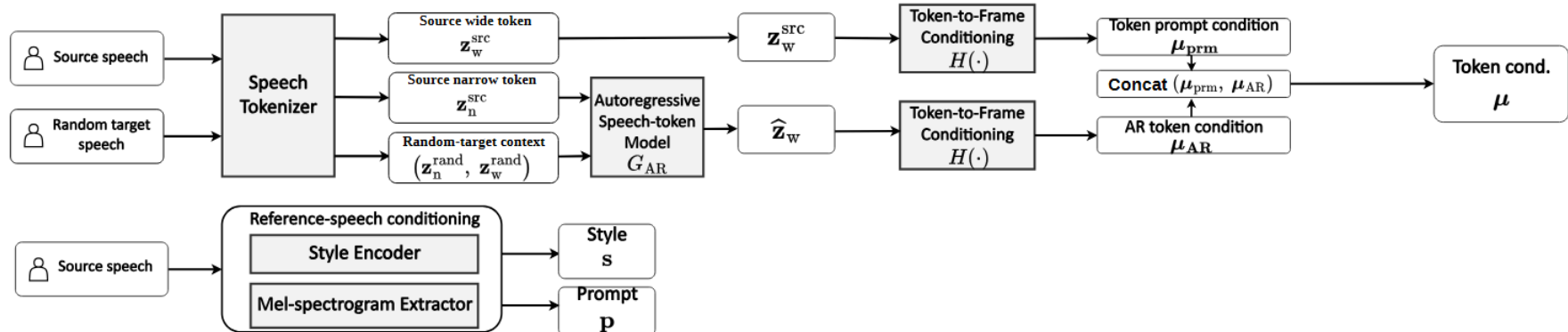


Stage 1: Random-target-conditioned token construction

Key Modification:

- Random target utterance supplies the AR context and prefix.
- The source still supplies content tokens and the aligned prompt/style.

Stage 1: Random-Target-Conditioned Token and Source-Acoustic Condition Construction





Stage 2: projected classifier-free guidance

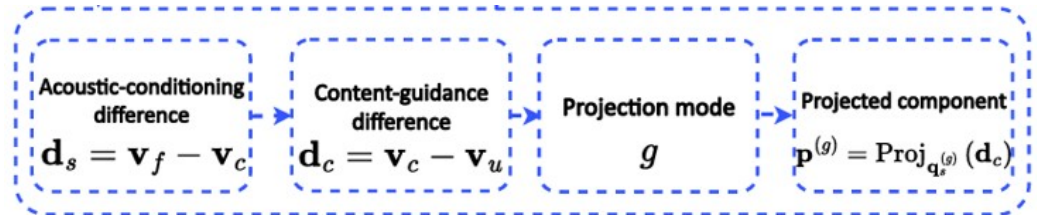
Key Modification:

Construct two differences:

$\mathbf{d}_s = \mathbf{v}_f - \mathbf{v}_c$ captures the source-acoustic conditioning direction.

$\mathbf{d}_c = \mathbf{v}_c - \mathbf{v}_u$ captures the token/content guidance direction.

We project $\mathbf{d}_c$ onto $\mathbf{d}_s$ and suppress the aligned component, then reinforce the remaining token guidance.



Proposed anonymization guidance rule

$$\mathbf{v}_{\text{anon}} = \mathbf{v}_c - \lambda_s \mathbf{p}^{(g)} + \lambda_c \mathbf{d}_c$$

λ_s

Controls source-aligned suppression:
stronger privacy, but too large may harm
WER.

λ_c

Controls token/content guidance: better
intelligibility, but too large may reduce
anonymization.



Three projection mode

Token

Frame-level source-conditioning vector at each time step.

Global

One utterance-level mean direction over the whole sequence.

Smooth21

Local sliding-window mean with $K = 21$ frames.

All three variants use the same basic idea: remove token-guidance components aligned with source conditioning.



Development ablation: both stages help

TABLE I

DEVELOPMENT-SET ABLATION RESULTS. RT MEANS THAT μ_{AR} USES RANDOM-TARGET CONTEXT. HIGHER EER AND UAR ARE BETTER, WHILE LOWER WER IS BETTER. THE HIGHEST EER IS SHADED.

System	RT	λ_s	λ_c	EER	WER	UAR
Built-in		0.7	0.7	33.15	4.18	41.85
Global		0.9	2.8	33.90	4.20	39.21
Smooth-21		1.0	3.0	34.20	4.39	40.35
Token		0.2	2.4	33.48	4.22	40.30
Global	✓	0.9	2.8	34.44	4.72	38.62
Smooth-21	✓	1.0	3.0	34.79	4.47	39.06
Token	✓	0.2	2.4	33.80	4.60	39.43
Global	✓	1.2	2.9	34.61	4.42	40.13
Smooth-21	✓	1.1	2.8	35.09	4.52	40.38
Token	✓	0.1	1.3	34.72	4.36	39.95

Finding 1

All projection variants improve EER over built-in anonymization.

Finding 2

Random-target conditioning improves EER across all projection modes.

Selected

Smooth-21 + RT: best development EER with acceptable WER/UAR.



Track 1 comparison: selected Smooth-21 system

Table 2: Track 1 comparison with the official baselines under lazy-informed and semi-informed attacks. Privacy is measured by EER (%), and utility by WER (%) and UAR (%). A checkmark indicates that μ_{AR} uses random-target context. A slash denotes a non-applicable scale. The maximum value in each EER column is shaded.

System	Random-target for μ_{AR}	λ_s	λ_c	Development				Evaluation			
				Lazy-EER	Semi-EER	WER	UAR	Lazy-EER	Semi-EER	WER	UAR
Orig.		/	/	7.34	–	1.80	69.08	3.91	–	1.84	71.06
B2		/	/	8.15	4.23	10.44	55.66	5.62	3.72	9.96	53.49
B3		/	/	26.03	16.32	4.31	38.08	19.06	14.24	4.31	35.24
B4		/	/	19.73	15.28	6.16	41.97	14.64	11.13	5.90	42.78
B5		/	/	30.94	20.05	4.91	40.08	25.65	17.05	4.44	38.25
Built-in anon.		0.7	0.7	33.15	24.64	4.18	41.85	28.19	17.78	4.58	40.38
Smooth-21	✓	1.1	2.8	35.09	25.14	4.53	40.38	29.39	21.51	4.30	40.16

Privacy gain

Smooth-21 improves both lazy-informed and semi-informed EER over built-in anonymization and other baselines on both development and evaluation set.

Utility

WER remains comparable to the strongest official baseline B5, while UAR is slightly better and is acceptable compared to other UAR performances.



Track 2 comparison: selected Global system

Lan.	Metric	Development						Evaluation					
		Orig.	BM1	BM2	BM3	Global	Built-in	Orig.	BM1	BM2	BM3	Global	Built-in
fr	Lazy-EER	6.15	8.02	25.43	47.43	21.37	17.70	7.51	9.49	28.48	48.49	25.80	20.29
	Semi-EER	6.15	1.57	15.45	46.11	9.28	8.32	7.51	2.08	15.30	46.25	9.93	9.10
	WER	5.80	12.42	52.17	13.77	21.38	14.93	5.13	9.93	42.24	11.99	18.66	12.97
es	Lazy-EER	8.97	9.31	28.78	48.78	24.71	21.13	4.86	5.76	25.03	48.19	18.89	16.01
	Semi-EER	8.97	0.93	10.74	46.23	8.78	8.87	4.86	1.58	8.74	45.92	7.50	7.29
	WER	3.79	6.15	20.37	10.56	12.27	10.21	3.75	6.21	28.92	10.12	12.22	9.72
en	Lazy-EER	2.24	2.31	9.73	36.68	10.09	8.88	4.75	4.27	11.55	36.87	11.86	10.42
	Semi-EER	2.24	1.19	5.25	22.02	7.02	7.86	4.75	3.40	7.53	20.40	7.88	9.01
	WER	4.57	7.40	14.29	6.71	10.95	11.36	5.46	8.54	15.41	7.73	12.08	11.43
de	Lazy-EER	3.00	5.86	25.25	48.25	20.41	17.92	5.02	8.75	25.72	48.24	22.76	18.90
	Semi-EER	3.00	1.12	10.05	46.54	8.96	8.92	5.02	1.36	10.90	46.05	10.21	10.57
	WER	5.25	7.81	19.28	9.74	15.12	12.21	5.38	9.38	24.40	10.10	18.58	14.21
Avg.	Lazy-EER	5.09	6.38	22.30	45.29	19.15	16.41	5.54	7.07	22.70	45.45	19.83	16.41
	Semi-EER	5.09	1.20	10.37	40.23	8.51	8.50	5.54	2.11	10.62	39.66	8.88	8.99
	WER	4.85	8.45	26.53	10.20	14.93	12.18	4.93	8.52	27.74	9.98	15.38	12.08
emo	UAR	72.09	41.30	27.13	26.24	28.07	27.72	78.90	42.63	29.04	24.84	26.62	27.64

**Better performances
than BM1**

**Substantial room for
improvement in privacy
performance and
linguistic and emotion
preservation**



Limitations and future work

Limitations

This anonymous method currently performs well only in English. However, it does not demonstrate strong anonymization capabilities in multilingual contexts.

Future work

Adapt target-token construction and projection guidance to multilingual speech.

Improve the privacy, intelligibility and emotion trade-off.



Thank you for your listening!

Questions?