

Acoustic token admixture for joint speaker and content anonymization

Suppressing the biometric and the linguistic channel inside a single acoustic token space, with span-level edits

Ali Golmakani · Seyed Ahmad Hosseini · Omar Manil Bendali · Emmanuel Vincent · Brij Mohan Lal Srivastava

Nijta SAS, France · Université de Lorraine, CNRS, Inria, LORIA, Nancy



Ali Golmakani



Seyed Ahmad
Hosseini



Omar Manil
Bendali



Emmanuel
Vincent

Presented by Brij Mohan Lal Srivastava.
Code, weights and audio samples are public;
links on the final slide.

MOTIVATION

Two re-identification channels in one recording

“However, Jaime Alguersuari failed to score any points in the season.”

the utterance used throughout this talk

Biometric

The voice itself: vocal tract, prosody and energy distribution, matched by speaker verification [3].

addressed by voice conversion

Linguistic

The name sits in the transcript and survives any change to the voice; transcripts alone support authorship attribution [11].

untouched by voice conversion

Recordings are kept for how they sound: channel, accent, timing. Removing the identity and that character together protects the speaker and destroys the reason to keep the data.

Signal processing: McAdams coefficient shifting



No learned model, and no path for the transcript: the content channel does not exist here.

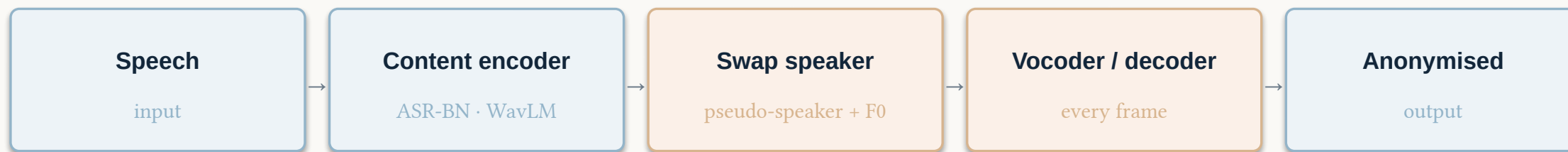
Achieves

Formant shifting without training [4], used as a challenge baseline.

Differs from ours

We keep the original tokens outside edited spans, and we act on the content channel at all.

Voice conversion and codec models: the strong baselines



The content representation passes through untouched, so the name is preserved by design.

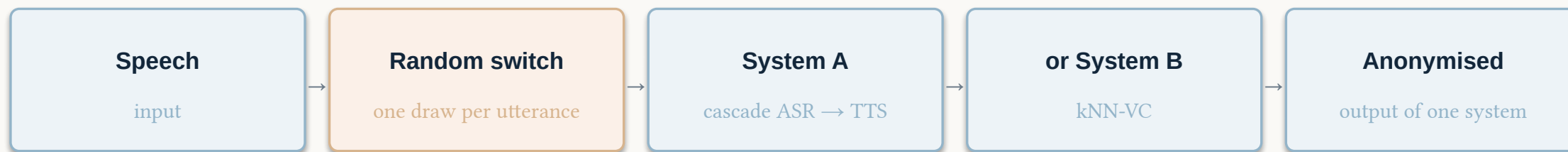
Achieves

The strongest equal error rates in the challenge: kNN-VC [5], the baselines [1], the top 2024 system [6].

Differs from ours

Same conditioning idea, but our content path can be rewritten frame by frame instead of copied.

Utterance-level routing: mixing two systems



The switch sits before two complete systems; the output belongs to one of them.

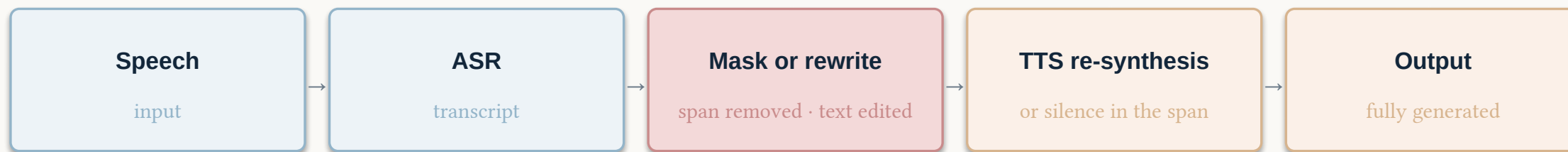
Achieves

Routing across two systems widens the attacker's problem [7]; the closest antecedent to this work.

Differs from ours

We move the switch inside one system, to every frame, which is what makes span editing possible.

The content side, developed separately



Either the span becomes unusable, or the whole utterance is generated from text.

Achieves

Masking removes a sensitive span [8]; substitution and paraphrasing rewrite the transcript [9, 10].

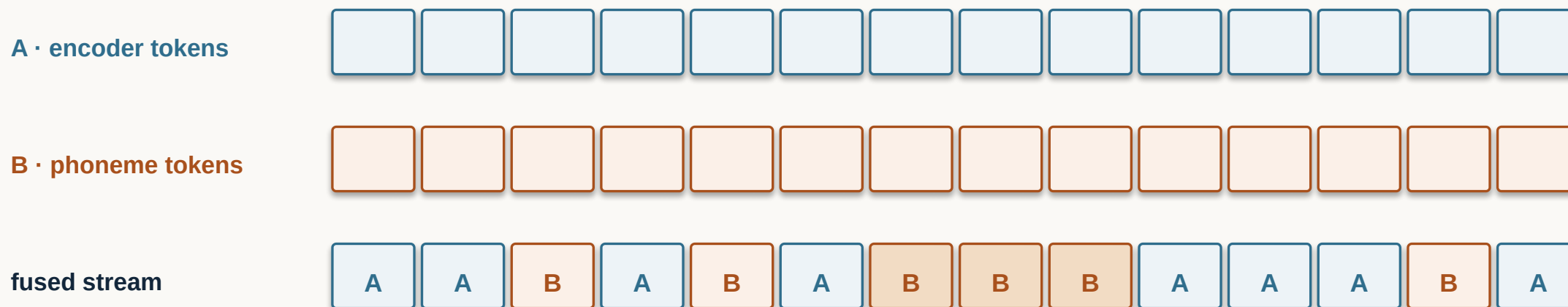
Differs from ours

We regenerate only the entity frames and keep the surrounding tokens and timing.

METHOD

Frame-level selection between two token streams

Each column is one 20 ms frame, 50 per second. Both streams emit eight RVQ indices per frame.



$\beta = 0.7$

probability a frame is taken from Stream B

NER span

those frames always take B, overriding β and the gate

$\tau = 0.6$

outside spans, reject a B frame that disagrees with A

DESIGN DECISIONS

Three decisions and the alternatives considered

01

Discrete tokens

Adopted: selection between two index sequences

Alternative: interpolation of continuous features

A swap needs a shared alphabet; interpolation produces a vector neither stream would emit.

02

Per-frame selection

Adopted: a Bernoulli draw at every frame

Alternative: routing whole utterances

At a matched rate privacy is comparable. It removes subsystem attributability and enables span editing.

03

Similarity gate

Adopted: reject B frames that disagree with A

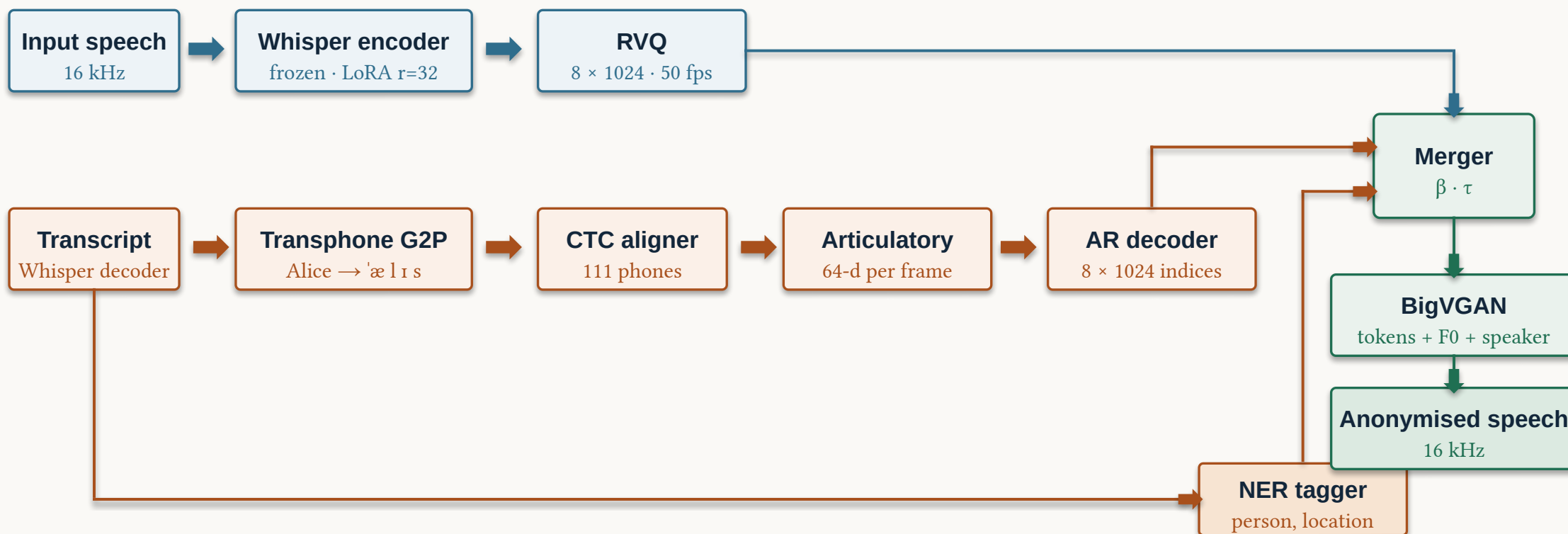
Alternative: trusting the generator at high β

Alignment errors otherwise become confident wrong tokens: WER degrades, privacy does not improve.

System-level finding. The gate decouples privacy control from intelligibility, and is disabled inside NER spans, where the streams are expected to disagree.

SYSTEM

Architecture: two streams, one merger, one synthesiser



Both streams end in the same eight-stage index format. The vocoder input is 1473-dimensional: 1280 token dimensions, one F0 value, a 192-dimensional speaker embedding. The dashed path from the tagger is the span override.

NER-triggered span replacement



the stream is re-indexed to absorb a length change; one pseudospeaker conditions the whole utterance

Duration is predicted, not copied

The replacement may be shorter or longer than the entity it replaces.

The gate is bypassed inside the span

There the two streams are expected to disagree, so the override suppresses both β and τ .

AUDIO EXAMPLES

One utterance under three conditions

“However, Jaime Alguersuari failed to score any points in the season.”

1 • ORIGINAL

source recording



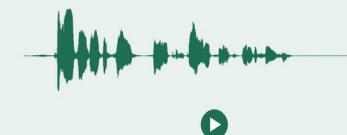
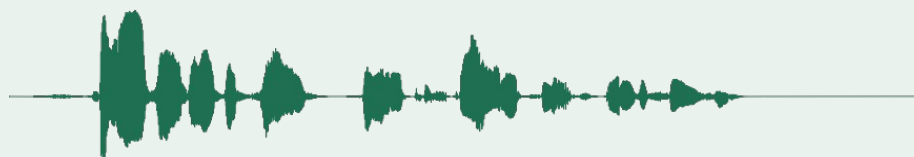
2 • ANONYMISED, $\beta = 0.7$

same words and timing, different speaker



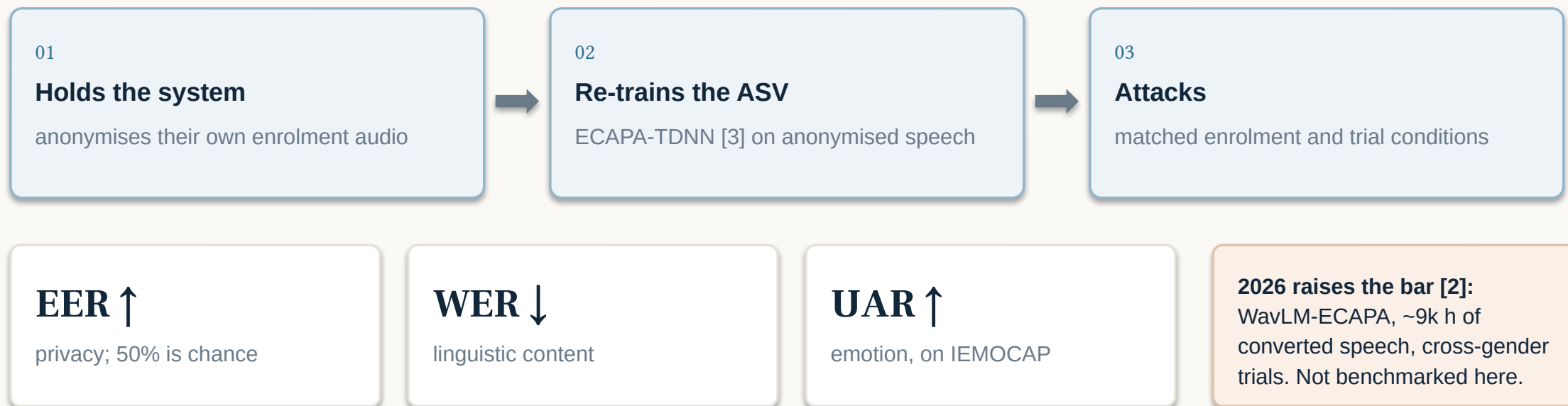
3 • ENTITY REPLACED

“... David Jones failed to score ...”,
utterance shorter



EVALUATION PROTOCOL

The semi-informed attacker [1]



RESULTS

VoicePrivacy 2024 benchmark, semi-informed condition

42.54%

EER ↑ at $\beta = 0.7$

3.73%

WER ↓, Whisper medium.en

40.11%

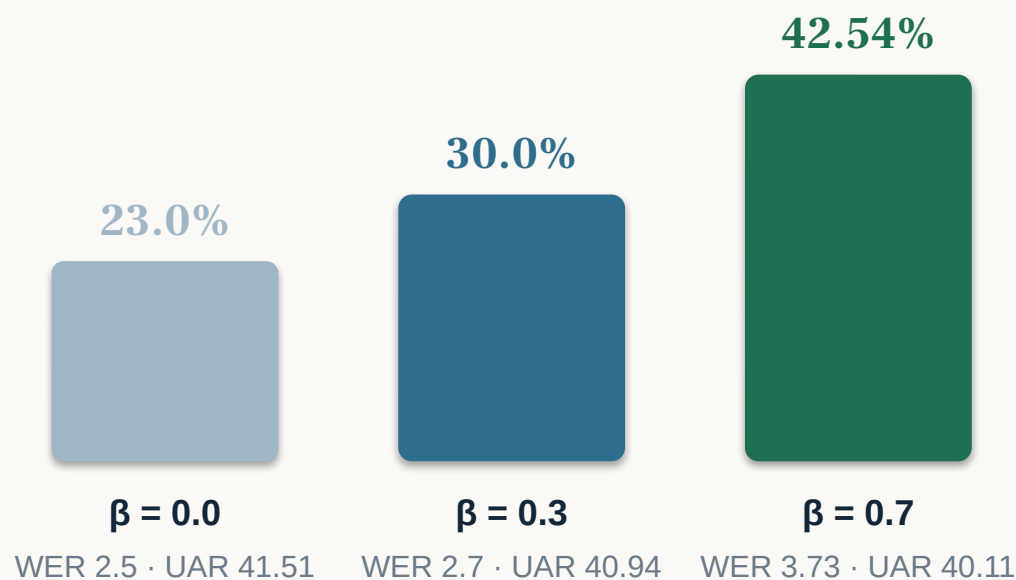
UAR ↑, weakest result

System	EER ↑	WER ↓	UAR ↑
Original speech	10.2	1.8	70.1
T12-5, top-ranked VPC 2024 submission	43.23	4.56	37.83
HLTCOE utterance-level admixture, $p = 0.4$	40.81	3.33	47.09
This work, $\beta = 0.7$	42.54	3.73	40.11

WER uses Whisper medium.en, not the challenge recogniser, so that column is indicative. All figures follow the 2024 protocol.

RESULTS

β controls the privacy–intelligibility trade-off



EER increases monotonically with β : more phoneme-derived frames remove more speaker-correlated evidence.

The cost across the whole range is 1.2 points of WER, from alignment errors the gate suppresses only partially.

The two lower points are reported without the F0 transformation, so this is not a single controlled ablation.

Two cosine metrics: what each establishes, and what neither can

0.123

Anonymisation similarity ↓

x-vector cosine to the original speaker.

Establishes: the voice moved far from the source.

Cannot: show that a trained attacker fails. EER does that.

0.959

Edit consistency ↑

cosine between the anonymised source and its edited version.

Establishes: the pseudospeaker does not drift across the splice.

Cannot: show the edit is inaudible, or the entity unrecoverable.

—

What is missing

no privacy metric for content.

No attacker recovers the entity; no detector looks for the splice.

Linkability and singling out [19] are the direction we intend to take.

Separate set of 213 utterances with person or location entities. WER 10.8 → 11.9 and predicted MOS 3.91 → 3.84 on that set; MOS is wv-mos [18], not a listening test.

LIMITATIONS

What the paper does not establish

1 • Content anonymisation is not evaluated as privacy

No attacker is applied to the transcript. The channel that motivates the work carries no privacy number in our results.

2 • Entity detection bounds the method, and its errors propagate both ways

A missed entity is unprotected; a false positive destroys material that mattered, and that cost falls on the content. Test boundaries were verified by forced alignment; a deployment cannot assume that.

3 • Emotion

UAR 40.11%. Disentangled representations discard emotional information [20].

4 • Scope

English only • 213-utterance editing set • predicted MOS • 2024 attacker.

CONCLUSIONS

Contributions, results and open directions

Contributions

A single token space in which per-frame admixture and entity-triggered span replacement act on the same representation. One synthesis stage; a gate that decouples privacy control from intelligibility.

Results

42.54% EER at 3.73% WER under the 2024 semi-informed protocol, within 0.69 EER points of the top submission. 0.959 speaker consistency across replaced spans.

Open directions

Prosody conditioned on the source; content coverage beyond named entities. Evaluation under the 2026 attacker and under legally grounded metrics.

Code (MIT) · weights and pools (CC BY-NC 4.0) · samples · verification script

github.com/Nijta/acoustic-token-admixture



Constructing the shared token space

A codebook index, not a sound

A 20 ms frame is 1280 numbers. A frozen codebook of 1024 reference vectors approximates it; the token is the index of the nearest entry.

Eight residual stages

Each stage quantises what the previous one left over, so eight indices refine one frame. Reconstruction is their sum.

Stream B is trained on Stream A's indices

Indices extracted from real audio by Stream A are the training targets for Stream B, which sees only text-derived articulatory features. The codebooks are frozen, so Stream B does not choose the alphabet; it learns to address it.

Interpolation of continuous features would produce a vector neither stream emits. Selection between index sequences keeps every frame inside the distribution the vocoder was trained on.

BACKUP

Released configuration

Encoder	Whisper large-v2, frozen; LoRA rank 32, α 64 on query and value projections
Quantiser	8 stages $\times$ 1024 entries over 1280-d features, 50 frames per second
Aligner	CTC head, 111 phoneme classes, Viterbi decoding
Phoneme features	64-d articulatory vectors (IMS-Toucan inventory)
Gate	threshold 0.6 on stream agreement; falls back to the encoder frame
Span replacement	graft with $C = 8$ frames (~ 160 ms) crossfaded at each seam
F0	$\alpha = 0.75$ toward a 32-frame voiced moving average, with additive noise
Vocoder	BigVGAN, 1473-d input (1280 + F0 + 192), 16 kHz output
Pseudospeaker pool	1,465 English speakers, ECAPA 192-d, Affinity Propagation clustering

All reported figures

VoicePrivacy 2024	EER $\uparrow$	WER $\downarrow$	UAR $\uparrow$
Original speech	10.2	1.8	70.1
T12-5, top submission	43.23	4.56	37.83
HLTCOE admixture, $p = 0.4$	40.81	3.33	47.09
This work, $\beta = 0.0$ (no F0 transform)	23.0	2.5	41.51
This work, $\beta = 0.3$ (no F0 transform)	30.0	2.7	40.94
This work, $\beta = 0.7$	42.54	3.73	40.11

Editing set, 213 utterances: admixture only, anonymisation similarity 0.123, WER 10.8, MOS 3.91. Full system, edit consistency 0.959, WER 11.9, MOS 3.84.

References

- [1] Tomashenko et al., VoicePrivacy 2024 evaluation plan
- [2] Miao et al., VoicePrivacy 2026 evaluation plan
- [3] Desplanques et al., ECAPA-TDNN, Interspeech 2020
- [4] Patino et al., McAdams, Interspeech 2021
- [5] Baas et al., kNN-VC, Interspeech 2023
- [6] NPU-NTU VPC 2024 system, arXiv:2409.04173
- [7] Xinyuan et al., HLTCOE, ICASSP 2025
- [8] Williams et al., content masking, ITG 2023
- [9] Turan et al., privacy-transformed data, LREC 2022
- [10] Aggazzotti et al., CALFA, ICASSP 2026
- [11] Aggazzotti et al., authorship attribution, TACL 2024
- [12] Radford et al., Whisper, ICML 2023
- [13] Li et al., Transphone, 2022
- [14] Lux and Vu, IMS-Toucan features, 2022
- [15] Lee et al., BigVGAN, ICLR 2023
- [16] van den Oord et al., VQ-VAE, NeurIPS 2017
- [17] Zaratiana et al., GLiNER, NAACL 2024
- [18] Andreev et al., wv-mos, 2023
- [19] Vauquier et al., legally validated evaluation, Interspeech 2025
- [20] Gaznepoglu and Peters, emotion loss, ICASSP 2025
- [21] Bakari et al., accent cues, ICASSP 2026