

Natural Language instructions for Voice Anonymization

Orange Research submission for the VoicePrivacy 2026 Challenge – Track 1& Track 2 (T37)

Rachid Benhsina, Olivier Le Blouch, **Rayane Bakari**, Nicolas Gengembre

VoicePrivacy 2026

Evaluation protocol

EER

Privacy

Semi-informed ASV attacker · higher is better · 50% optimal

WER

Linguistic utility

Automatic Speech Recognition · lower is better

UAR

Emotion utility

Speech emotion recognition · higher is better

What changed since 2024

Multilingual Track 2

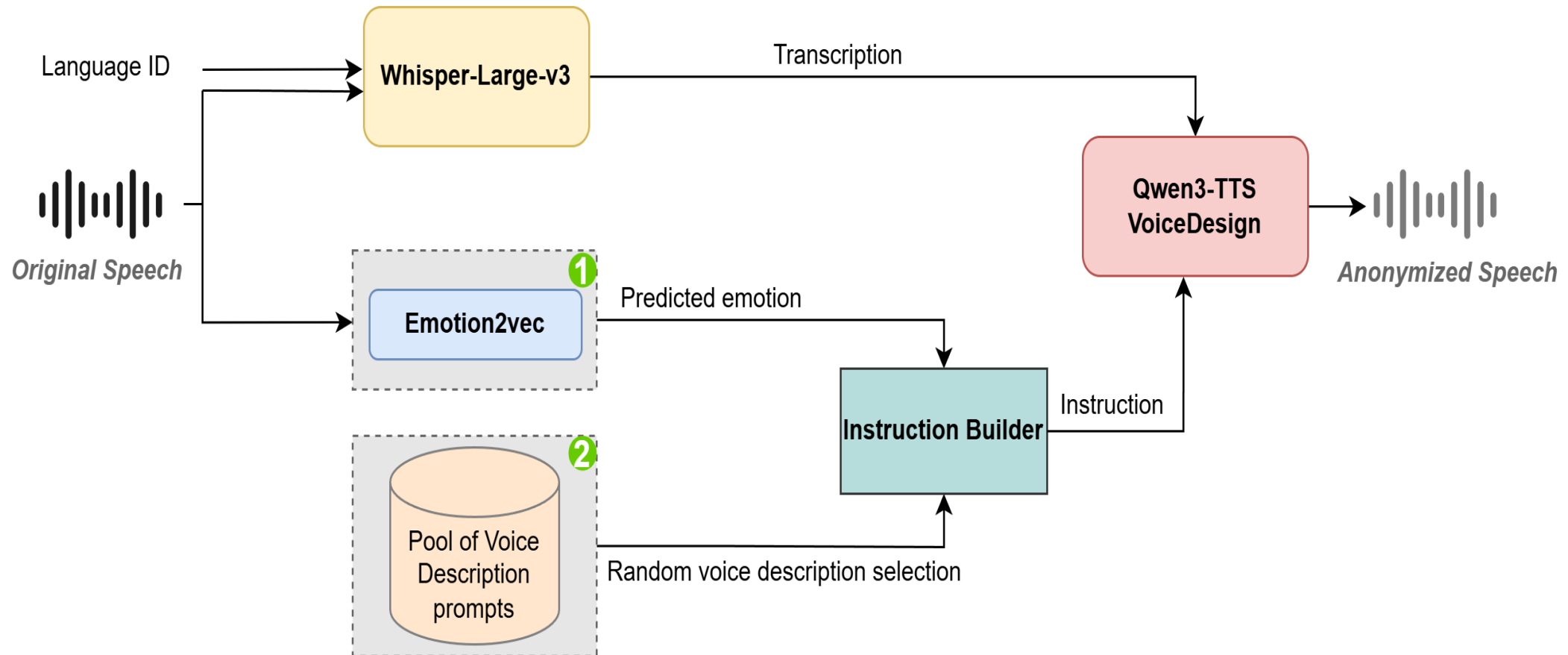
French, Spanish, English, German — alongside the original English-only Track 1.

Stronger attacker

WavLM-ECAPA, pre-trained on 94k hours incl. voice-converted data, fine-tuned on anonymized speech.

Result: fooling the attacker while keeping speech useful is significantly harder than in 2024.

A cascaded ASR-TTS pipeline, instruction-conditioned



Two ways to write the instruction

VD-Pool

Track 2

Voice description only

Voice description prompt, randomly drawn per utterance from a pool of 100 — no emotion conditioning.

EXAMPLE PROMPT

“A middle-aged female voice, deep and rich, steady pace, professor.”

VD-E2V

Track 1 & 2

Emotion instruction only

Emotion-conditioned instruction derived from Emotion2vec prediction— no voice description.

EXAMPLE PROMPT (emotion: angry)

“Speaking with anger, intense and serious tone.”

Track 1: High EER and minimal linguistic loss

English evaluation

System	EER	WER	UAR
Original	3.41	1.84	71.06
B2	3.72	9.11	53.52
B3	14.24	4.26	37.70
B4	11.13	5.90	42.78
B5	17.05	4.53	38.09
VD-E2V	31.60	2.45	42.42

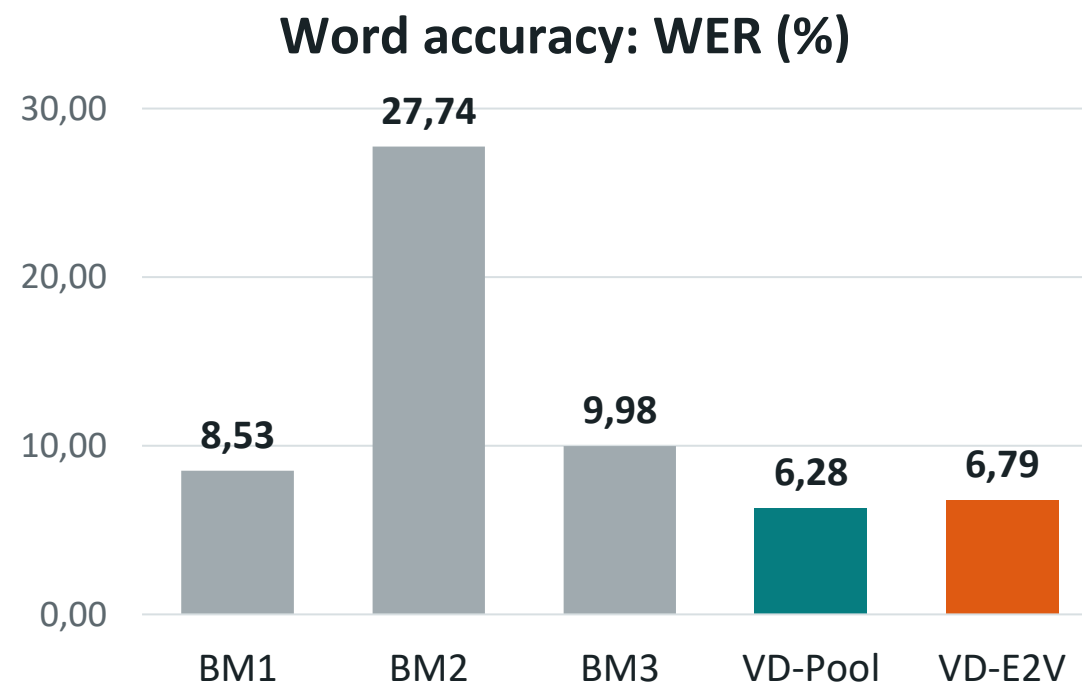
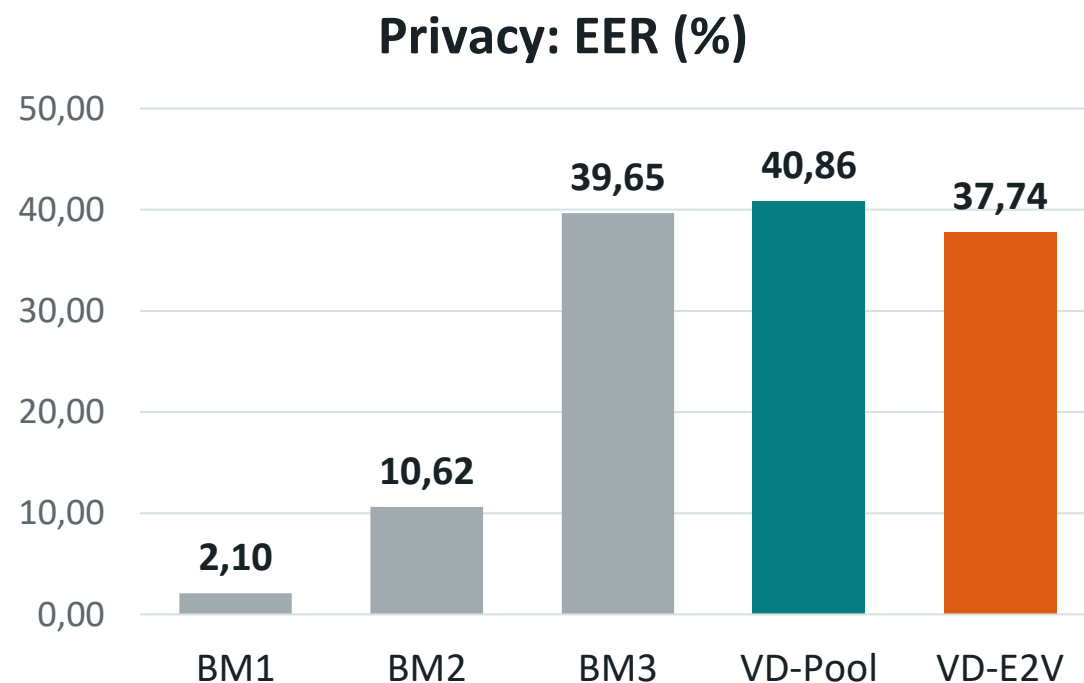
+ 14.55 EER points vs B5

+0,61 WER points vs original

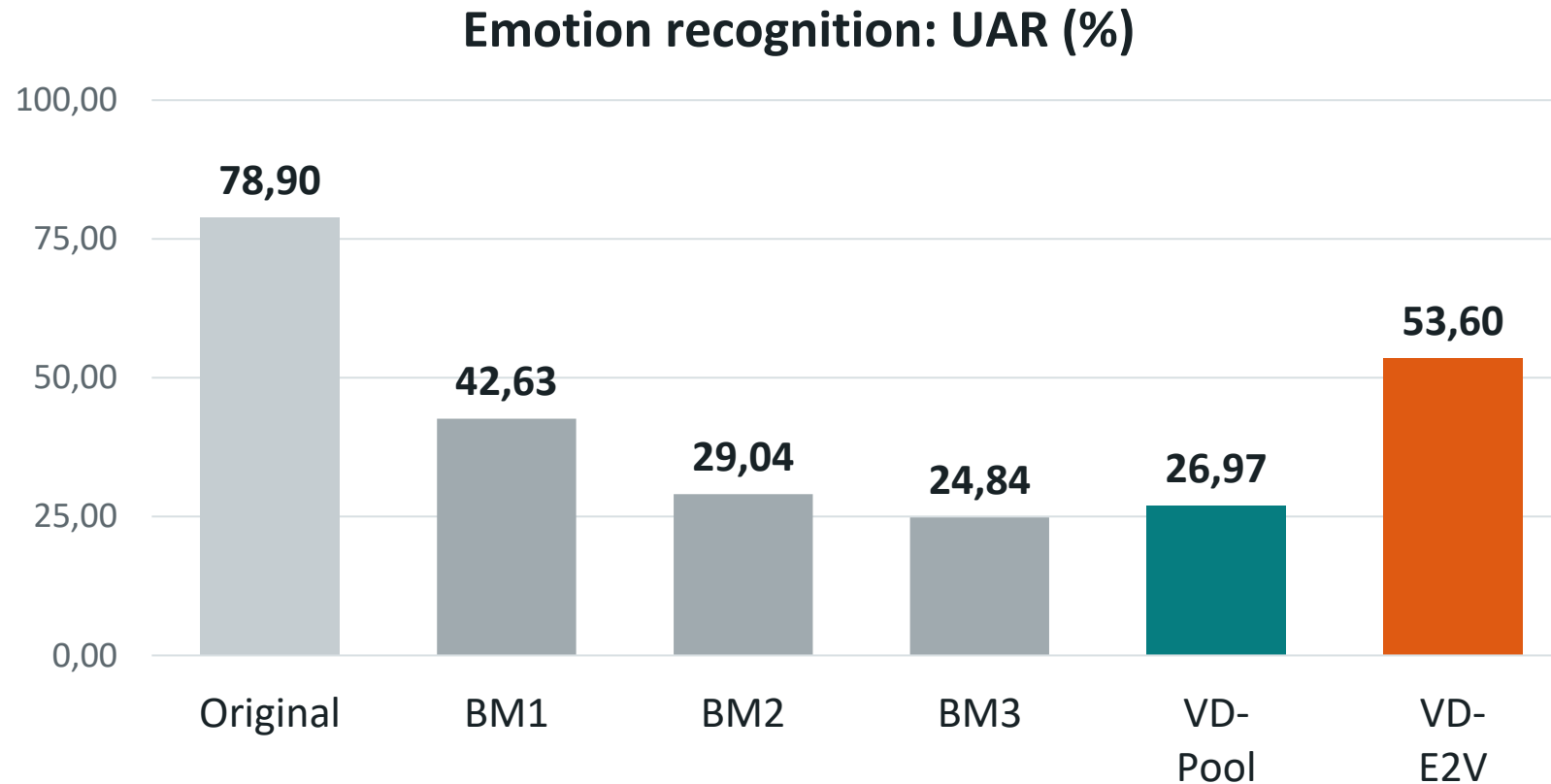
42.42% UAR on par with B4 (42.78%)

Track 2: Two systems with complementary trade-offs

French, Spanish, English and German



Track 2: Emotion conditioning raises UAR



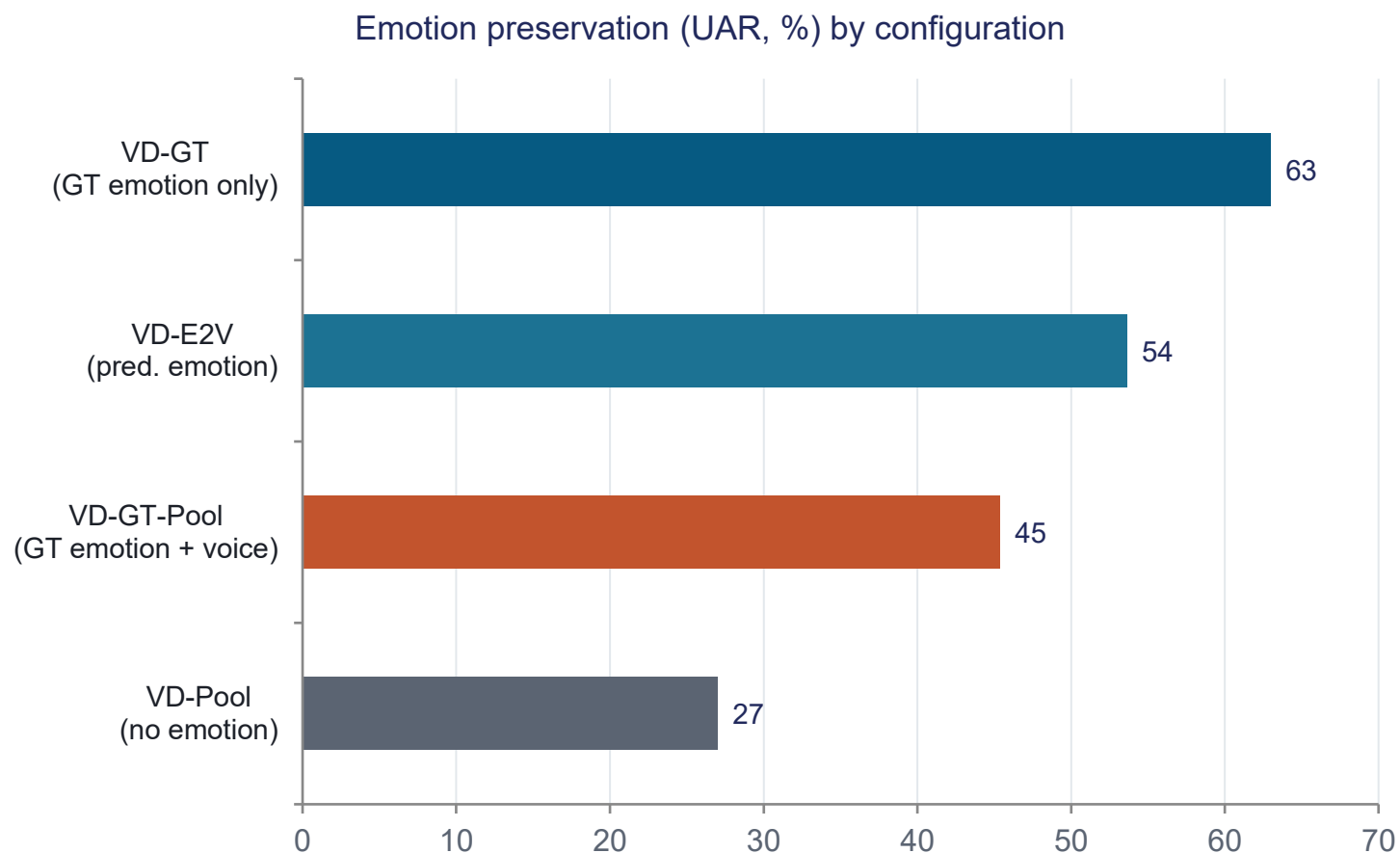
Cost vs VD-Pool

-3.12 EER points

+0,51 WER points

+26,93 UAR points

Voice descriptors can reduce the emotion preservation



+36 pts

VD-GT vs VD-Pool: perfect emotion labels show the achievable upper bound.

-17.6 pts

VD-GT-Pool vs VD-GT: adding a voice description drops UAR sharply.

Why

The model prioritizes matching the described timbre over expressing the emotion.

Conclusion and Discussion

1

Natural language prompts

Natural language instructions offer lightweight, effective control lever. VD-Pool best privacy & linguistic utility for Track 2 with no emotion control. VD-E2V Strong privacy on both tracks, with UAR far exceeds baseline.

2

Goodhart's Law

UAR measures emotion-label preservation, not authentic expressiveness. The system recreates a label-matching expression, not the speaker's own way of conveying it.

3

Open question

English consistently shows lower EER than French, Spanish and German — across both our systems and the baselines. Left for future work.

**Thanks
for your attention**

