

VoxCPM2 Voice-Design Anonymization for VoicePrivacy 2026

English and Multilingual Systems

Yibo Duan^{1*}, Xuyang Xing^{1*}, Zihan Chai^{1,3}, Maquan Yang¹,
Shengchen Li³, Sheng Zhao^{1,2**}

¹ VOICECODE TECHNOLOGY PTE. LTD., Singapore

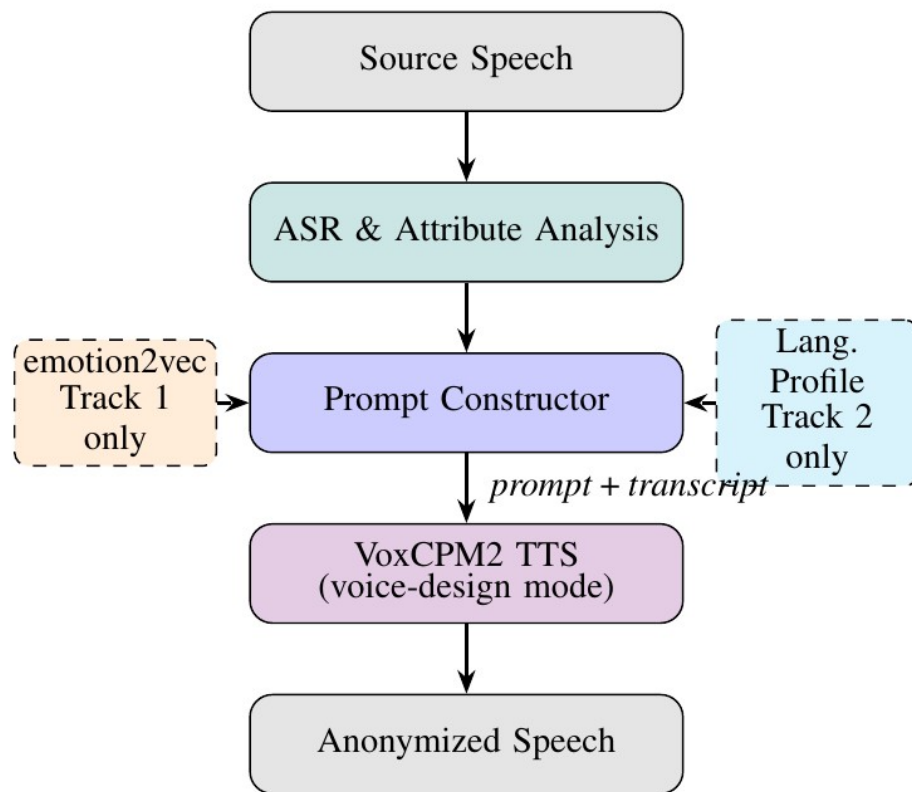
² SHANGHAI LONGYUAN TECHNOLOGY CO., LTD., China

³ XI'AN JIAOTONG-LIVERPOOL UNIVERSITY, Suzhou, China

Sydney, Australia / 26 September 2026

Voice design with text and coarse attributes

Shared architecture for English and multilingual anonymization



Text and coarse attributes

Whisper supplies the transcript [8].
Discrete delivery categories guide the prompt.

VoxCPM2 voice-design mode [6]

Fixed pretrained weights.
No source-speaker embedding or
reference audio enters synthesis.

Figure 1. Shared system architecture

Control labels come from the input speech

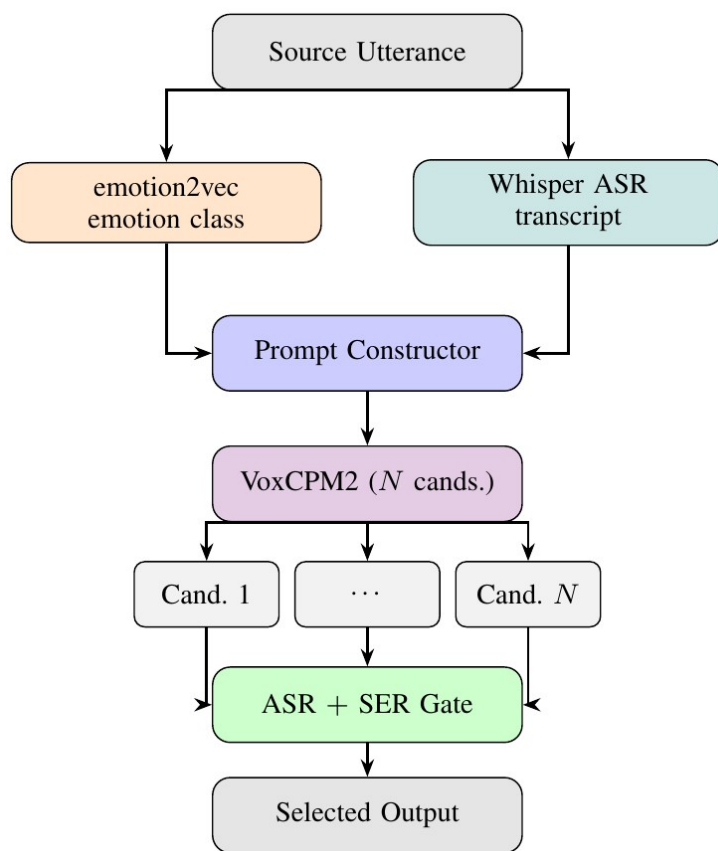
Predictions supply synthesis and candidate-selection targets

Control signal	Source	Use
Reference text	Whisper ASR [8]	TTS input and ASR gate
Target emotion	emotion2vec [12]	Prompt and SER gate
Source gender	ECAPA-TDNN classifier [10]	Opposite-gender prompt
Pace and energy	Word rate and waveform RMS	Coarse text categories

No dataset transcript, emotion label or gender metadata supplies these targets.

Track 1: emotion-aware candidate selection

General emotion prediction, with a dedicated IEMOCAP selection branch



Predicted targets

Whisper transcript [8]
emotion2vec emotion label [12]

Candidate checks

ASR: WER / CER against stored text
SER: match the predicted emotion

Emotion prediction covers all English inputs.
The IEMOCAP branch adds ASR/SER selection.

Figure 2. Emotion-aware candidate selection

Track 1: strong privacy and intelligibility



Updated official test evaluation

System	EER ↑ (%)	WER ↓ (%)	UAR ↑ (%)
Original speech	3.41	1.84	71.06
B5: ASR-BN	17.05	4.53	38.09
T15-2	31.39	3.72	64.69
T3: Voicecode	43.43	1.29	59.35

Highest EER and lowest WER among ranked Track 1 systems

Emotion utility remains below original speech and the strongest ranked UAR.

Sources: paper Table 5; official results Table I. Selected comparisons.

Track 2: language-aware selection and fallback



German, Spanish and French use multiple language-prompt profiles

Simple instruction / news reading / dictation

Candidate ASR selects the lowest CER, then compares with the standard output.

Language	EER ↑ (%)	WER ↓ (%)
English	17.67	4.51
German	36.03	12.40
Spanish	34.95	10.79
French	35.93	14.74
Average	31.15	10.61

Average UAR: 41.53% across four emotion classes

Sources: paper Tables 7–8; official results Table II.

Remaining limitations



Emotion fidelity, multilingual robustness and evaluation scope

Emotion loss

Track 1 neutral recall falls from 71.66% to 45.88%.
Emotion preservation remains incomplete.

Multilingual gaps

Some submitted systems outperform ours on all three
Track 2 averages. Language-specific errors remain.

Evaluation scope

Privacy is attack-specific. IEMOCAP-specific selection
also leaves a deployment-generalization question.

No ablation isolates the contribution of each design choice.

Conclusions



Text-driven voice design supports strong English anonymization

Track 1: 43.43% EER, 1.29% WER and 59.35% UAR

Predicted labels guide content and delivery control

The system uses source ASR, SER and gender predictions.

Multilingual robustness and emotion fidelity remain open

Thank you. Questions?

Backup: label provenance and branch scope

Candidate-selection clarification

Record	Documented contents
prompts.jsonl	emotion_source: "predicted_ser"
*.asr_gate.json	Reference, hypothesis, WER, CER, SER and selection
prompt_thresholds.json	Thresholds used during selection

Emotion prediction covers all inputs. Candidate selection includes an IEMOCAP-specific branch.

Label provenance and dataset-dependent routing are separate questions.

Backup: Track 2 comparisons



Updated official test averages (%)

System	EER ↑	WER ↓	UAR ↑
Original	4.25	4.93	78.90
BM3	39.65	9.98	24.84
T7	36.12	9.38	62.22
T24	43.90	7.85	56.34
T3: Voicecode	31.15	10.61	41.53

EER and WER average four languages. UAR averages four emotion classes.

Backup: emotion preservation by class

Track 1 IEMOCAP test recall (%)

Emotion	Original	Anonymous
Sad	72.57	66.40
Angry	72.83	76.56
Happy	67.19	48.56
Neutral	71.66	45.88
UAR	71.06	59.35

Happy and neutral show the largest losses in class recall.

Source: paper Table 6. Values rounded to two decimals.

Backup: perspectives for future challenges



Suggestions from our participation

Uniform inference policies

Clarify permitted dataset and language metadata, and evaluate one selection policy across all utterances.

Reproducible adaptive attacks

Provide versioned attack recipes and test leakage through text, coarse attributes and generated speech.

Per-language and per-emotion reporting

Retain breakdowns alongside averages to expose uneven utility.

Selected references



Numbering follows the camera-ready paper

[6] VoxCPM Team. VoxCPM2: Tokenizer-free TTS for multilingual speech generation, creative voice design, and true-to-life cloning. arXiv:2606.06928, 2026.

[8] A. Radford et al. Robust speech recognition via large-scale weak supervision. ICML, 2023.

[10] J. Huh et al. Voice gender classification using ECAPA-TDNN architecture. Interspeech, 2023.

[12] Z. Ma et al. emotion2vec: Self-supervised pre-training for speech emotion representation. Findings of ACL, 2024.

Figures 1–2 and table references refer to our camera-ready paper.