



The VoicePrivacy 2026 Challenge

6th Symposium on Security and Privacy in
Speech Communication

Xiaoxiao Miao, Natalia Tomashenko, Ridwan Arefeen, Sarina Meyer, Michele Panariello,
Xin Wang, Emmanuel Vincent, Nicholas Evans, Junichi Yamagishi, and Massimiliano Todisco

September 2026, Sydney, Australia

Organizers



Xiaoxiao Miao

Duke Kunshan University,
China



Natalia Tomashenko

Université de Lorraine,
CNRS, Inria, LORIA, France



Ridwan Arefeen

Singapore Institute of
Technology, Singapore



Sarina Meyer

University of Stuttgart,
Germany



Michele Panariello

EURECOM, France



Xin Wang

NII, Japan



Emmanuel Vincent

Université de Lorraine,
CNRS, Inria, LORIA, France



Nicholas Evans

EURECOM, France



Junichi Yamagishi

NII, Japan



Massimiliano Todisco

EURECOM, France

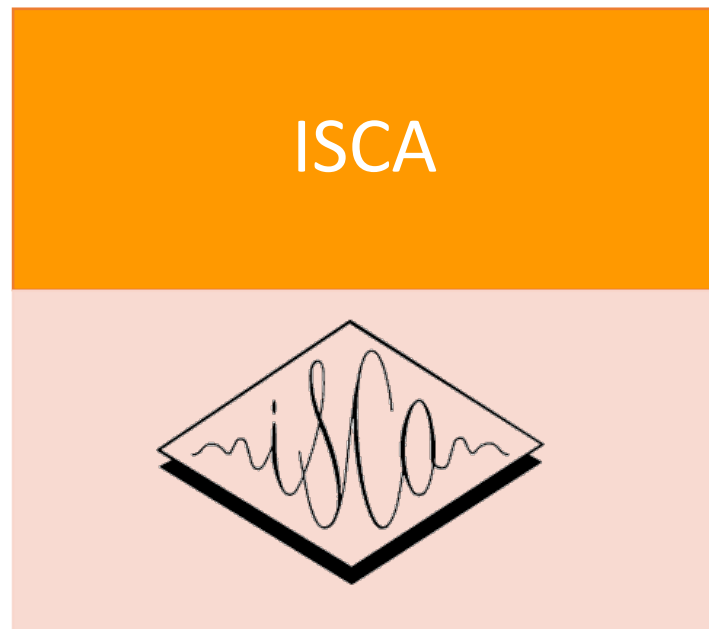


In Memory of Emmanuel Vincent (1979–2026)

VoicePrivacy 2026 is dedicated to the memory of Emmanuel Vincent, whose pioneering work and leadership were central to the creation of the VoicePrivacy initiative and to advances in voice anonymization and privacy in speech technology



Acknowledgement



Special thanks to everyone who reported issues and helped improve the VPC evaluation resources

- **Abdelkrim Mahmoudy**
 - Track 2 WER normalization
- **Martha Larson, Jiusi Zheng, and colleagues**, Radboud University
 - speaker-embedding normalization
- **Victor Ménéstrel**, Technische Universität Berlin:
 - Track 1 trial-list issue

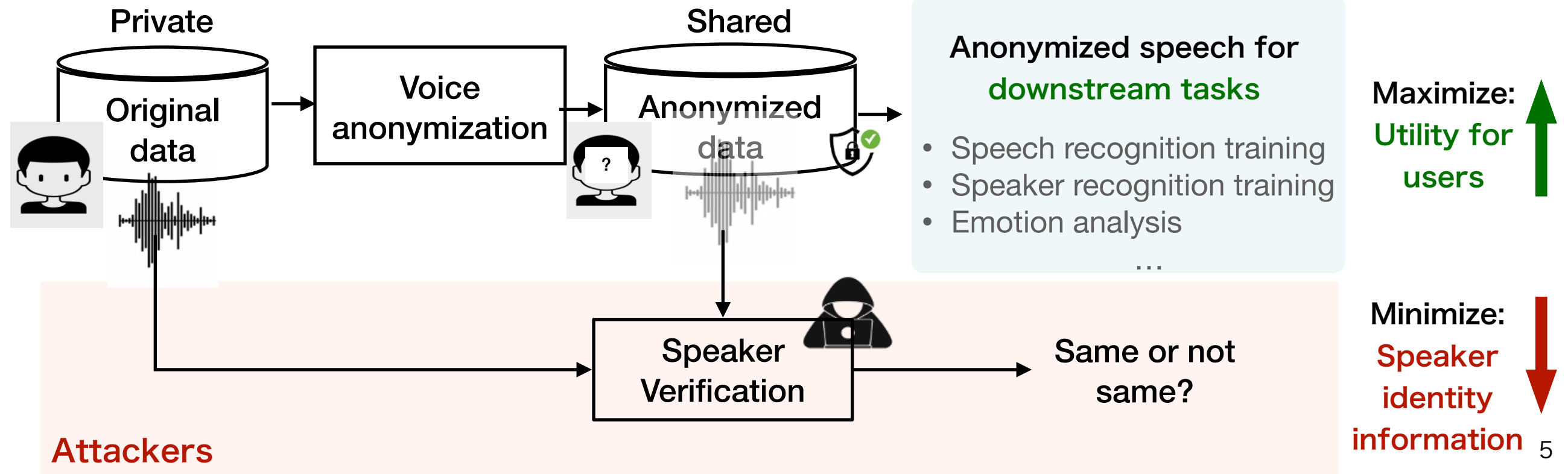
VPC: where we started and where we are

<https://www.voiceprivacychallenge.org/>

- ✗ Formal definition
- ✗ Benchmark
- ✗ Database
- ✓ Task definition
- ✓ Metrics
- ✓ Database



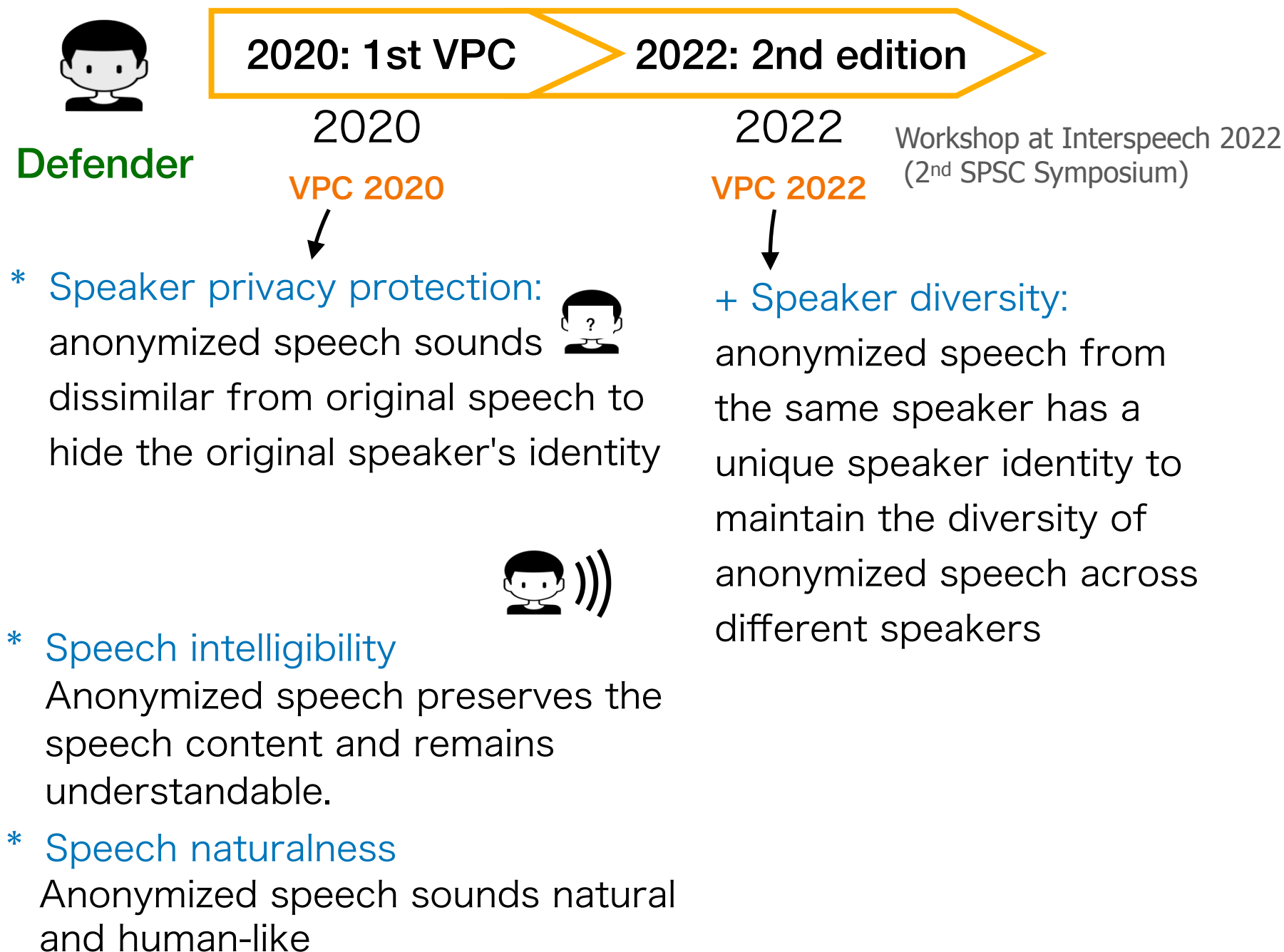
- remove personally identifiable information
- keep other characteristics unchanged



VPC: where we started and where we are

<https://www.voiceprivacychallenge.org/>

- ✗ Formal definition ✓ Task definition
- ✗ Benchmark ✓ Metrics
- ✗ Database ✓ Database



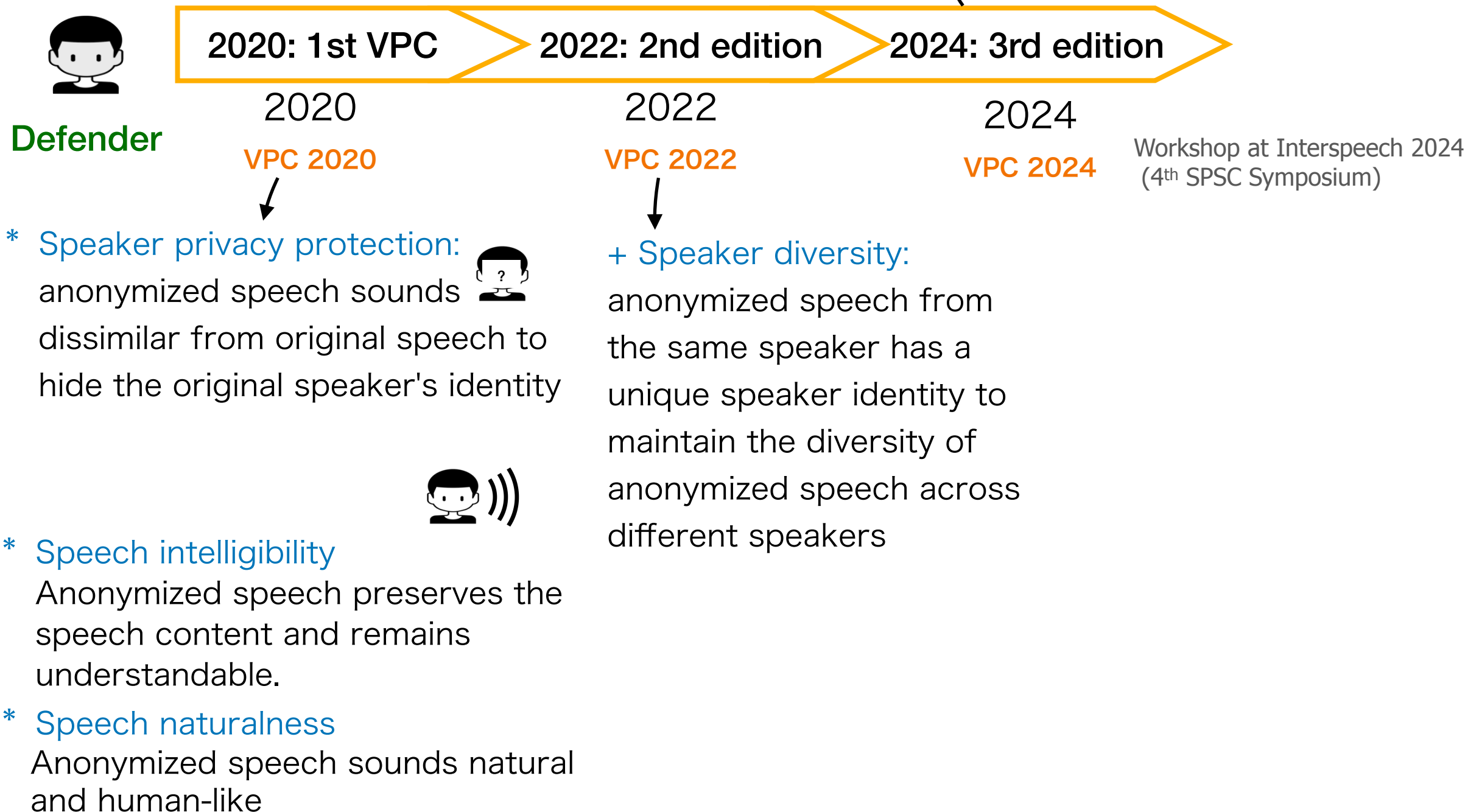
VPC: where we started and where we are

<https://www.voiceprivacychallenge.org/>



- ✗ Formal definition
- ✗ Benchmark
- ✗ Database
- ✓ Task definition
- ✓ Metrics
- ✓ Database

+ Emotional content: anonymized speech maintains the emotional attributes of original speech



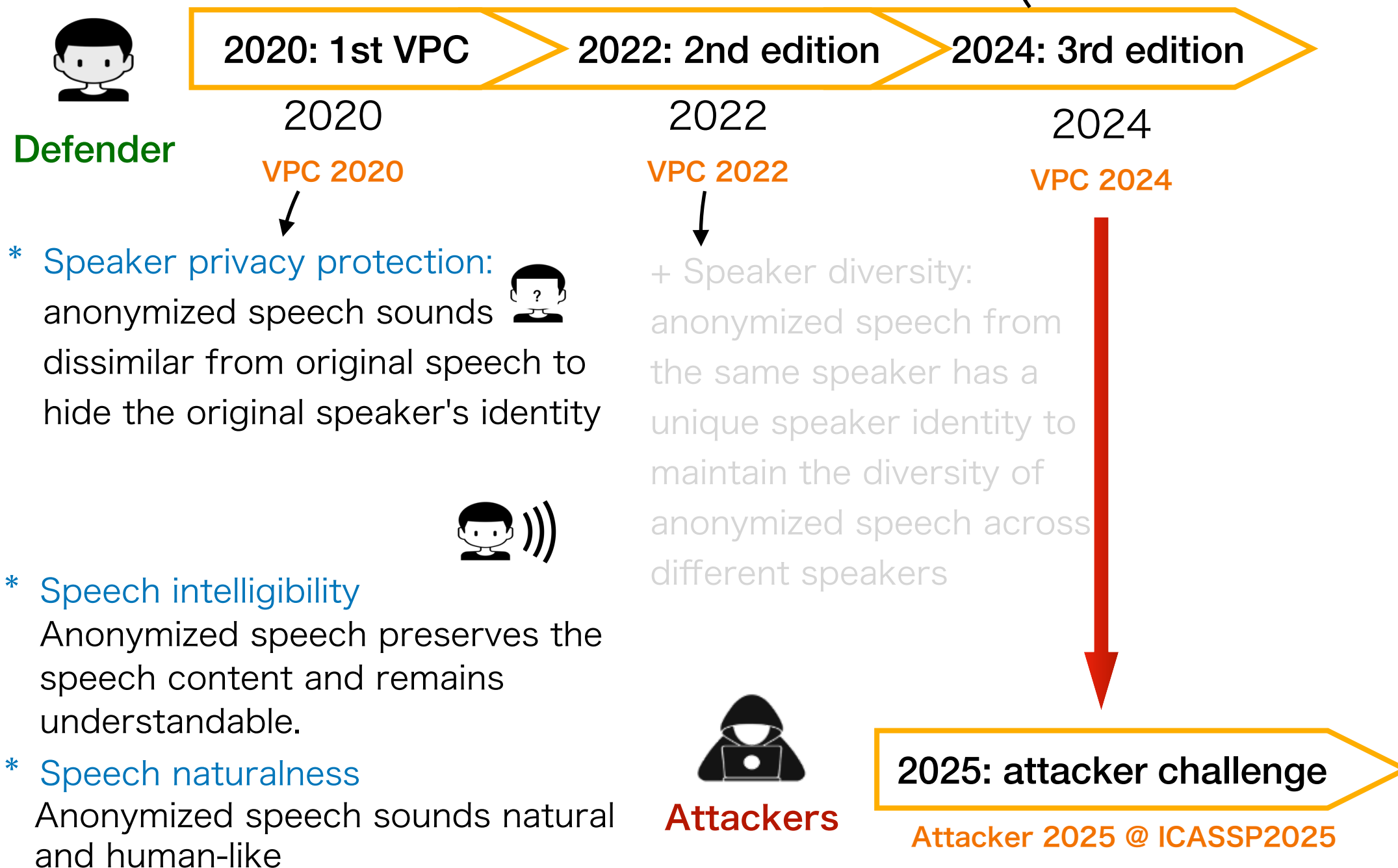
VPC: where we started and where we are

<https://www.voiceprivacychallenge.org/>



- ✗ Formal definition
- ✗ Benchmark
- ✗ Database
- ✓ Task definition
- ✓ Metrics
- ✓ Database

+ Emotional content: anonymized speech maintains the emotional attributes of original speech



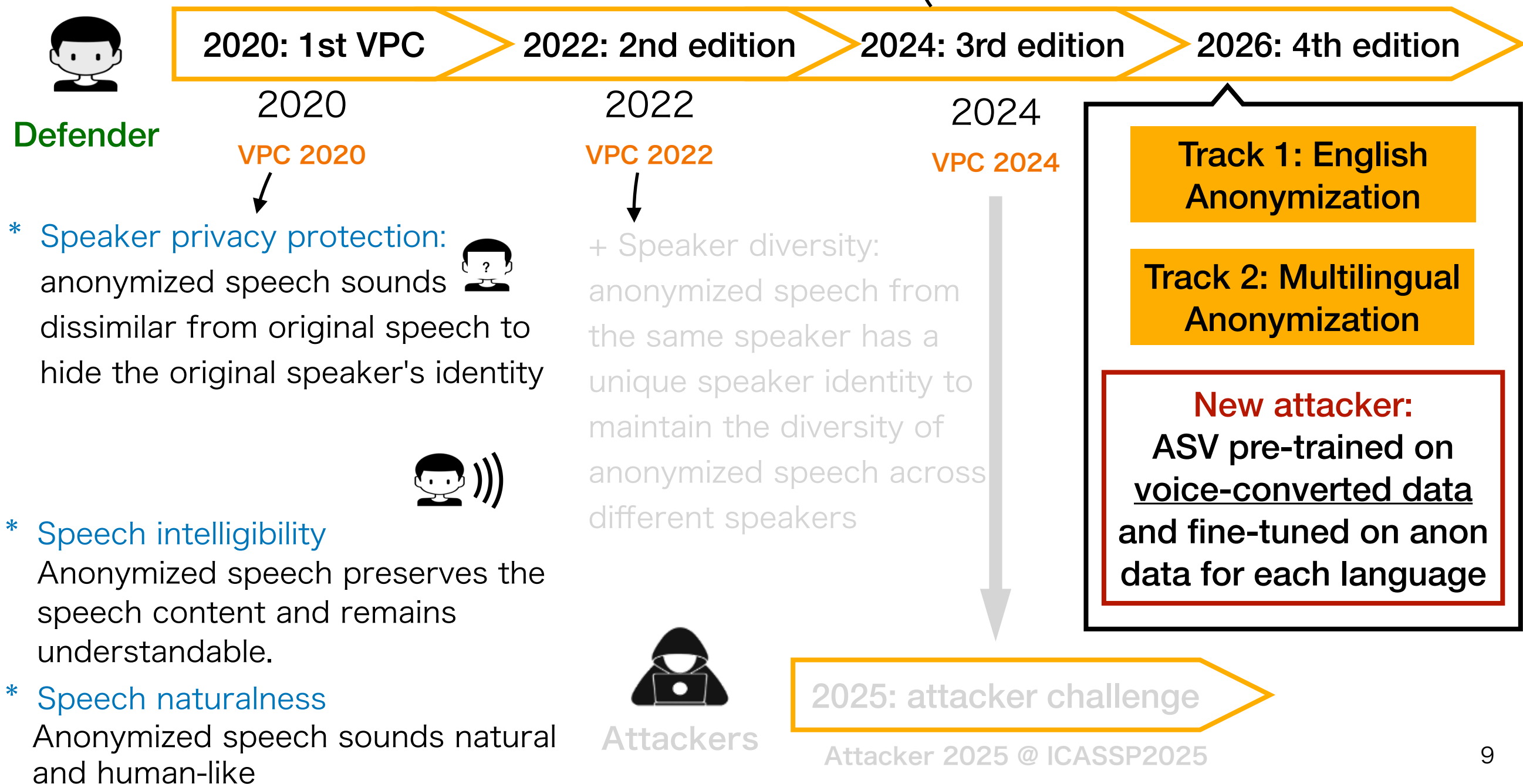
VPC: where we started and where we are

<https://www.voiceprivacychallenge.org/>



- ✗ Formal definition
- ✗ Benchmark
- ✗ Database
- ✓ Task definition
- ✓ Metrics
- ✓ Database

+ Emotional content: anonymized speech maintains the emotional attributes of original speech



VPC: where we started and where we are

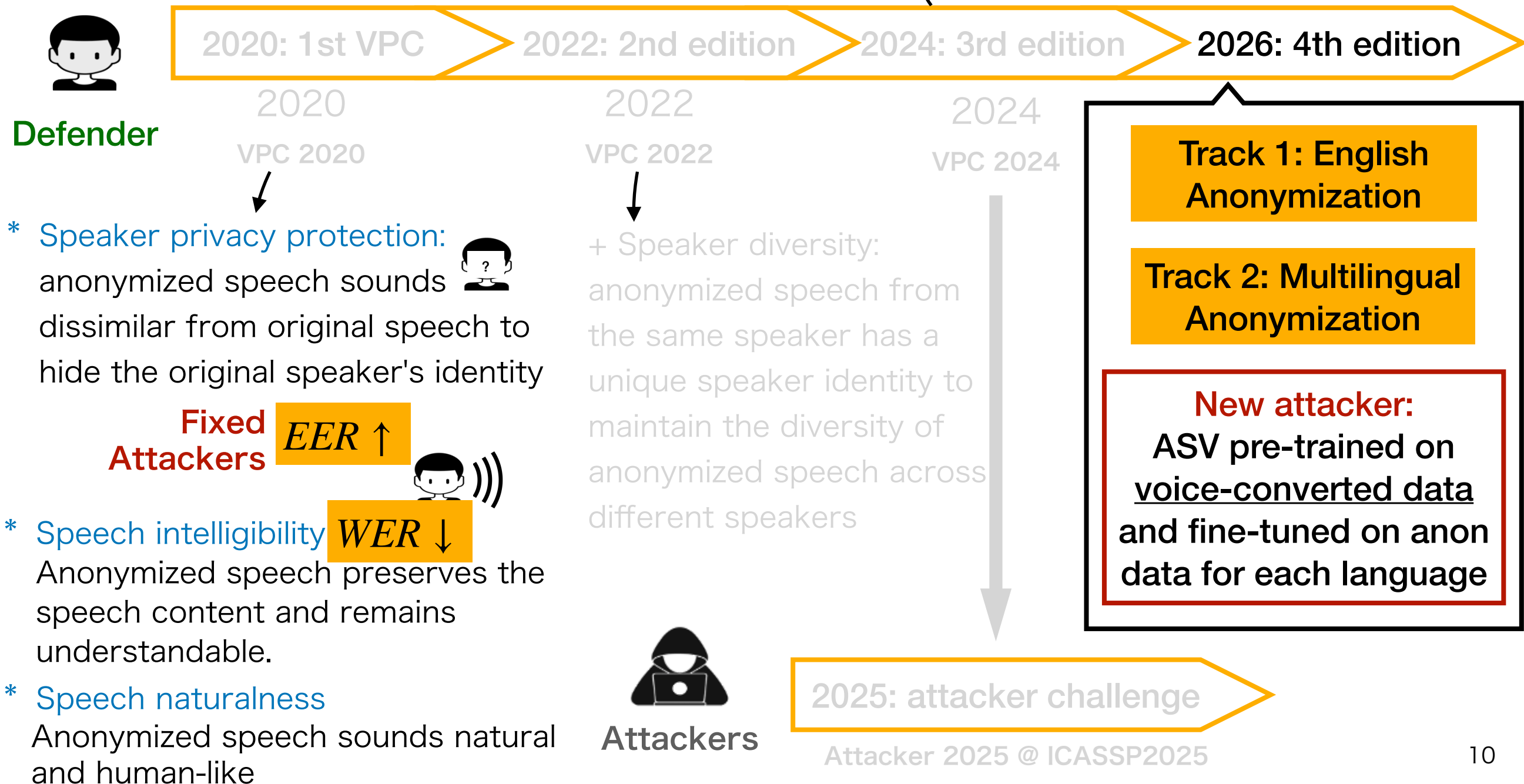
<https://www.voiceprivacychallenge.org/>



UAR ↑

- ✗ Formal definition
- ✗ Benchmark
- ✗ Database
- ✓ Task definition
- ✓ Metrics
- ✓ Database

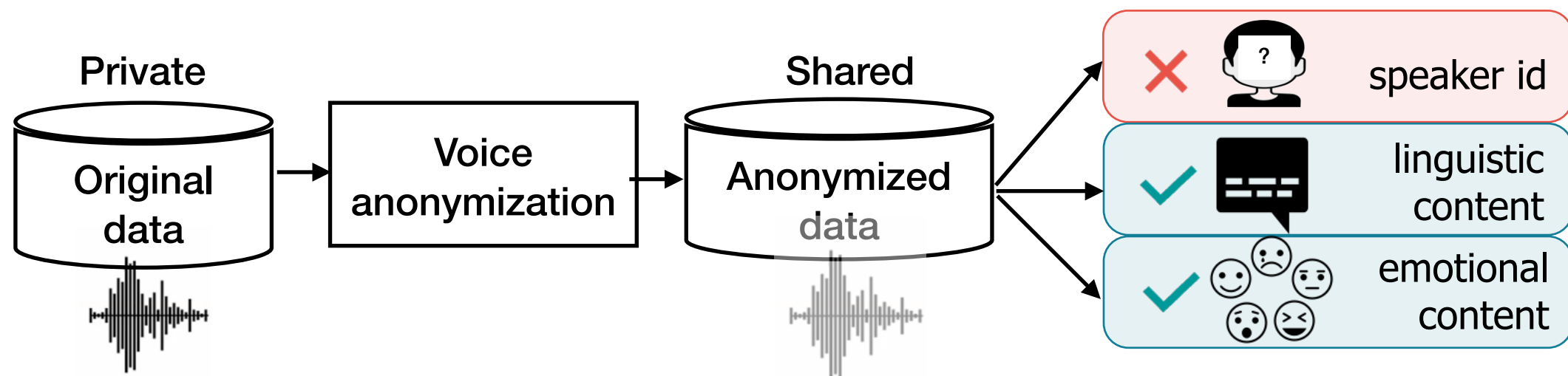
+ **Emotional content:** anonymized speech maintains the emotional attributes of original speech



Challenge task and requirements

- **Task:** Develop a voice anonymization system

Track 1: English Anonymization	<i>English</i> <i>en</i>
Track 2: Multilingual Anonymization	<i>English</i> <i>Spanish</i> <i>German</i> <i>French</i> <i>en</i> <i>es</i> <i>de</i> <i>fr</i>



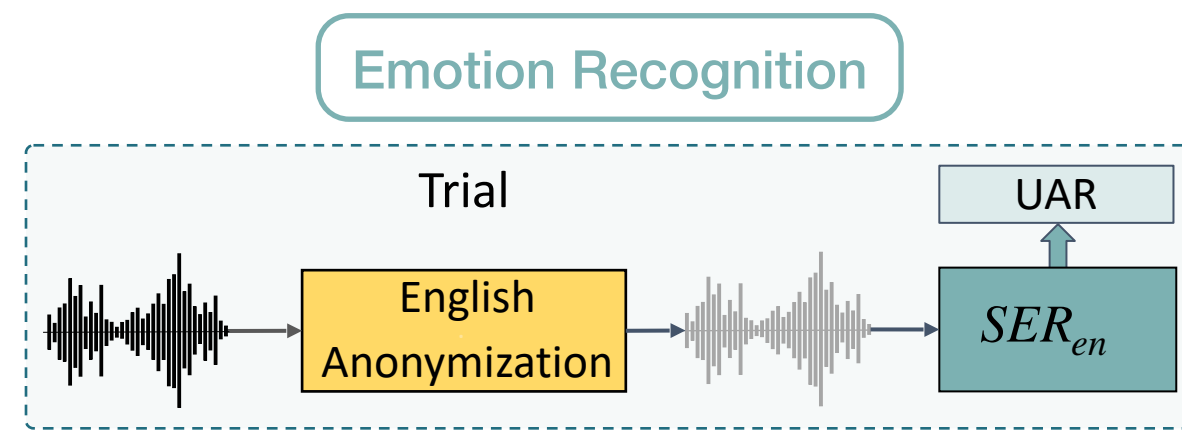
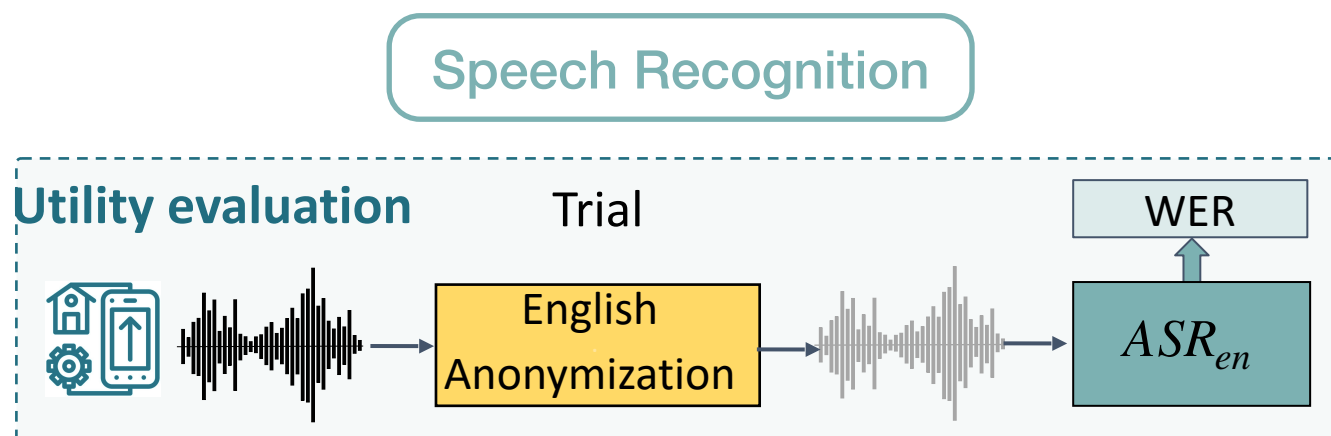
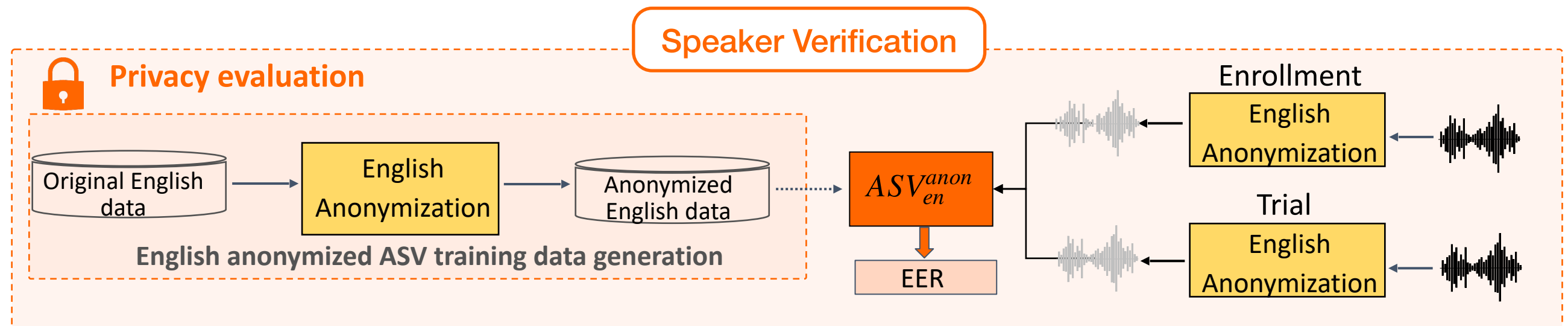
We provide:

1. training, development & evaluation data
2. baseline anonymization systems
3. evaluation scripts and metrics

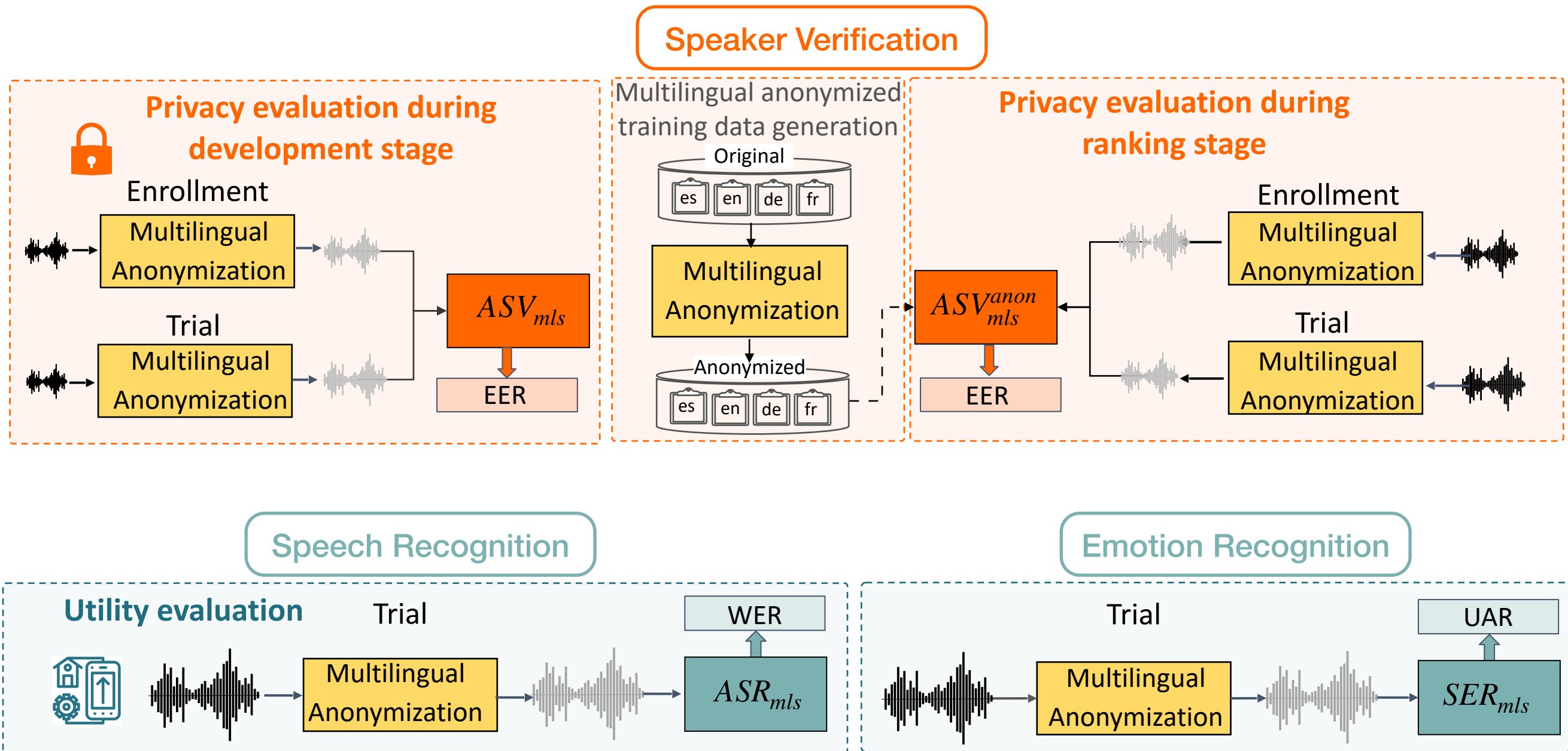
Participants:

1. apply their anonymization systems
2. run evaluation scripts
3. submit results & anonymized speech data

Track 1: English voice anonymization evaluation



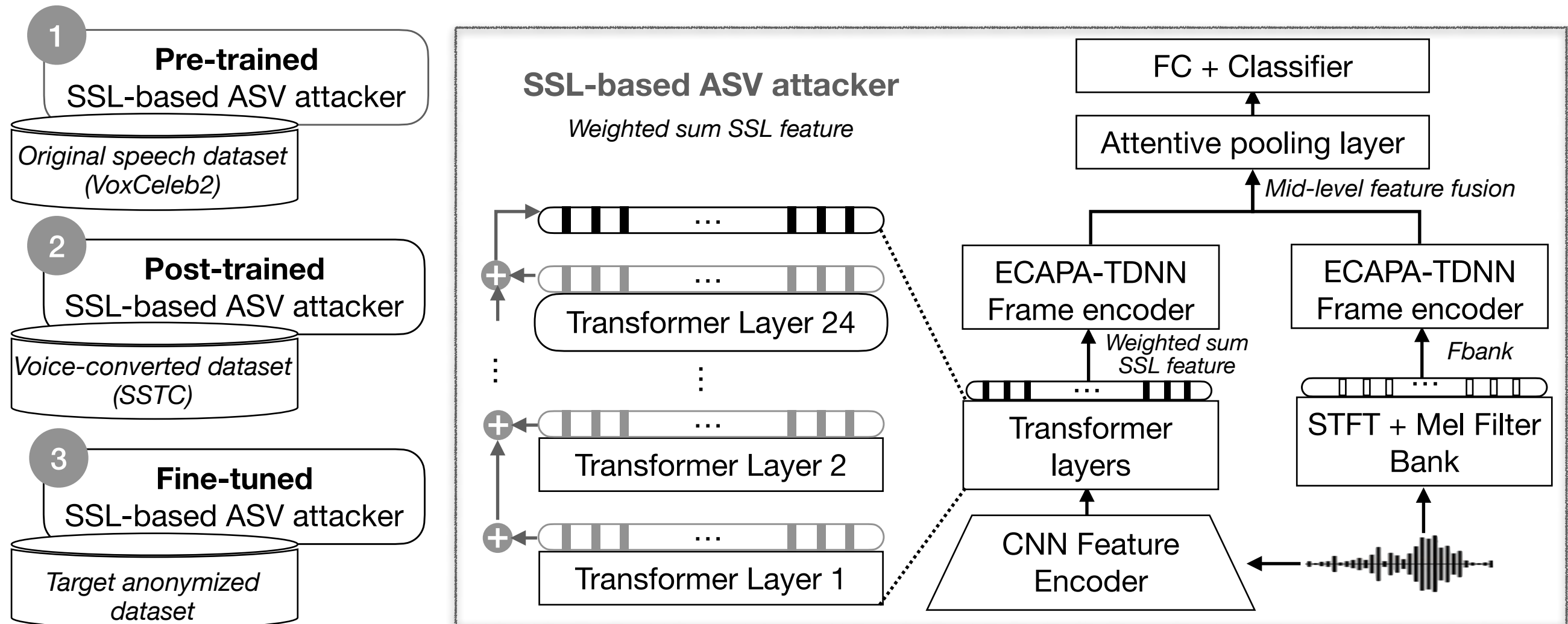
Track 2: Multilingual voice anonymization evaluation



Attacker model for privacy evaluation

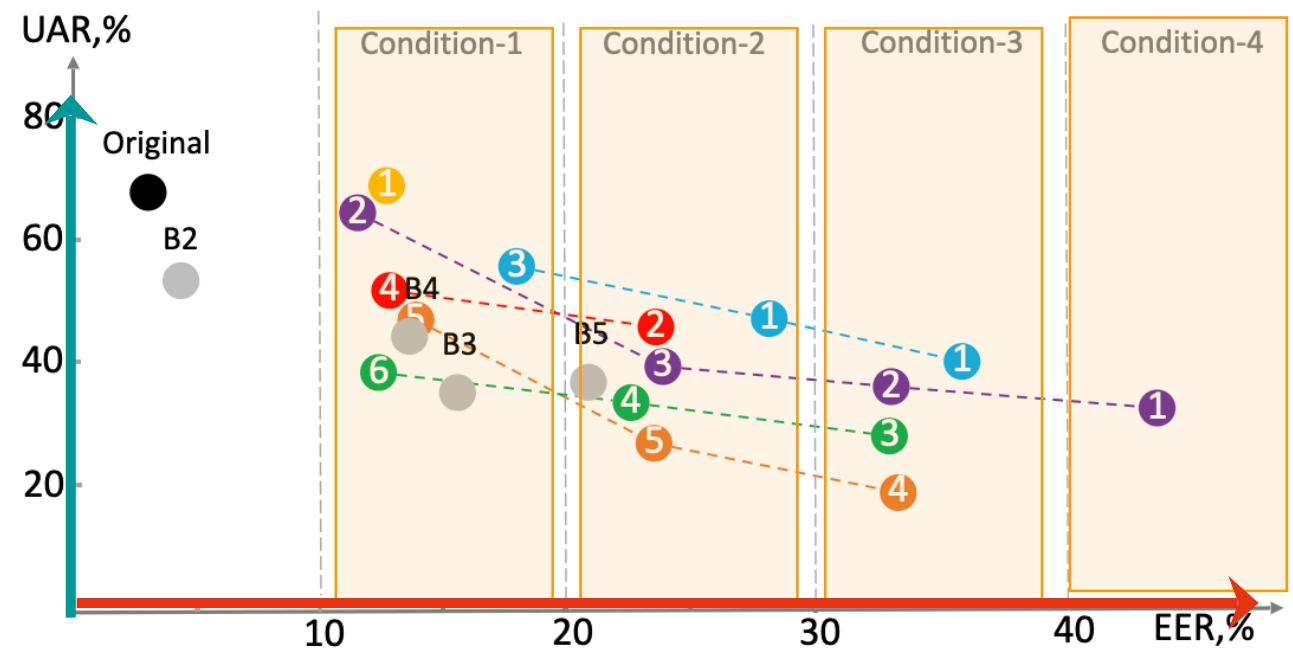
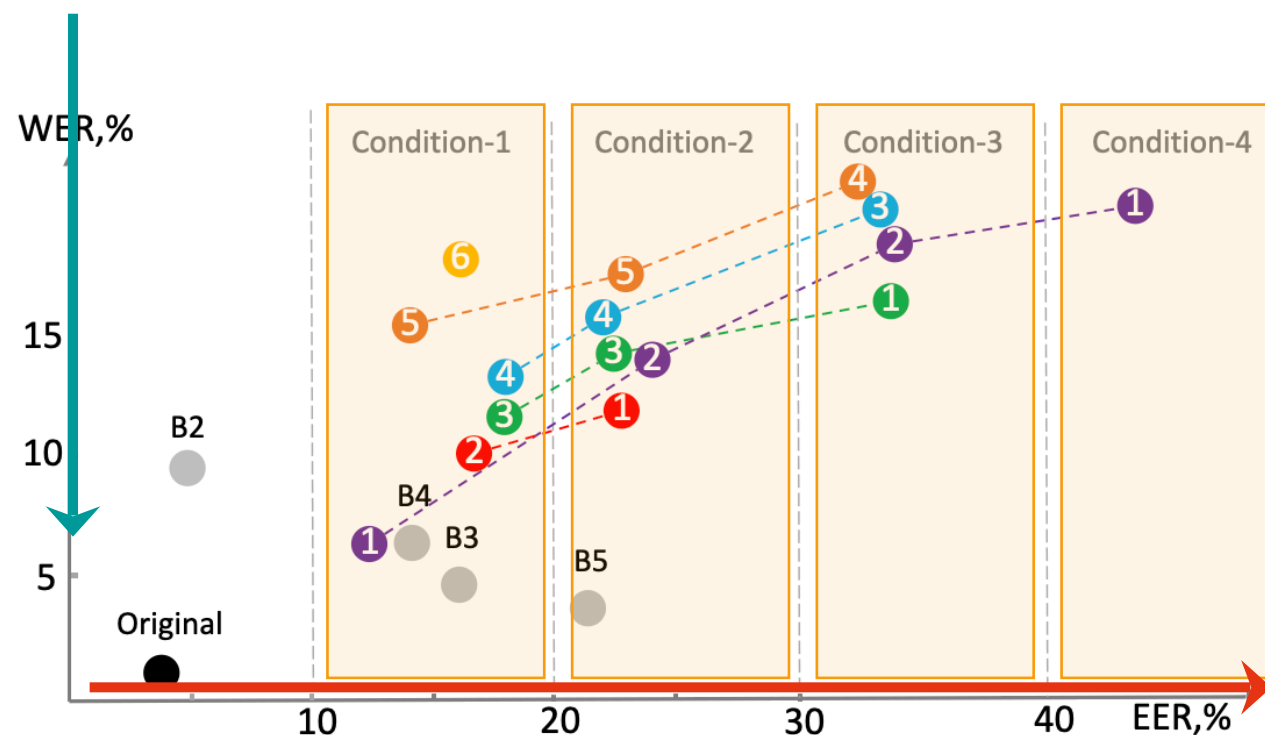
- Semi-informed attacker:
 - has access to the anonymization system
 - anonymizes enrollment & training data on the utterance level

New attacker[1]:
ASV pre-trained on voice-converted data and fine-tuned on anon data for each language



Ranking Policy

- **4** evaluation **conditions** with minimum target **privacy** EERs: 10%, 20%, 30%, and 40%
- Within each EER threshold condition, eligible systems are ranked separately by WER (lower is better) and UAR (higher is better).
- Track 2: **averaged** EER and WER **across all languages** are used for ranking



Data and Pretrained Models

Open training resources

Most used Training resources

Data



CAMEO
CREMA-D
MSP-Podcast
MEAD
EMO-DB
RAVDESS
LibriTTS
VoxCeleb-1,2
VoxPopuli

Models



Whisper-large-v3 ★★★★★
Whisper med.en ★★★★★
emotion2vec ★★★★★
VoxCPM2 ★★★★★
Qwen3-TTS ★★★★★
VITS ★★★★★
CosyVoice ★★★★★
IMSToucan ★★★★★
UUVC ★★★★★
SeedVC ★★★★★
wav2vec2 ★★★★★
HuBERT ★★★★★
ECAPA-TDNN ★★★★★
HiFi-GAN ★★★★★
BigVGAN ★★★★★

Fixed development and evaluation data

Development & Evaluation

Track 1

LibriSpeech ✓ Speaker verification
✓ Speech recognition
IEMOCAP
with 4 classes: ✓ Emotion recognition
neutral, sadness, anger & happiness

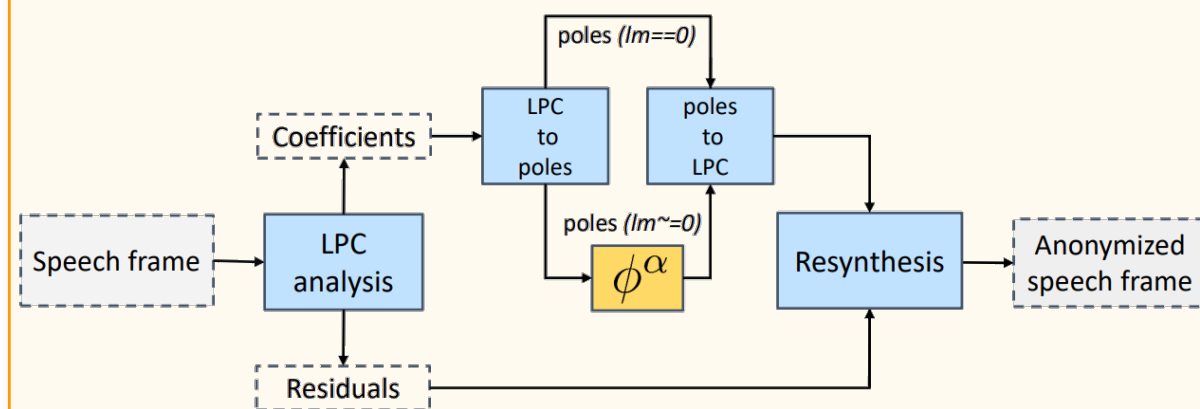
Track 2

Multilingual LibriSpeech
(English, French, German, Spanish)
✓ Speaker verification
✓ Speech recognition
Oreau French, MESD Spanish,
EMNS English, EmoDB German
with 4 classes:
neutral, sadness, anger & happiness
✓ Emotion recognition

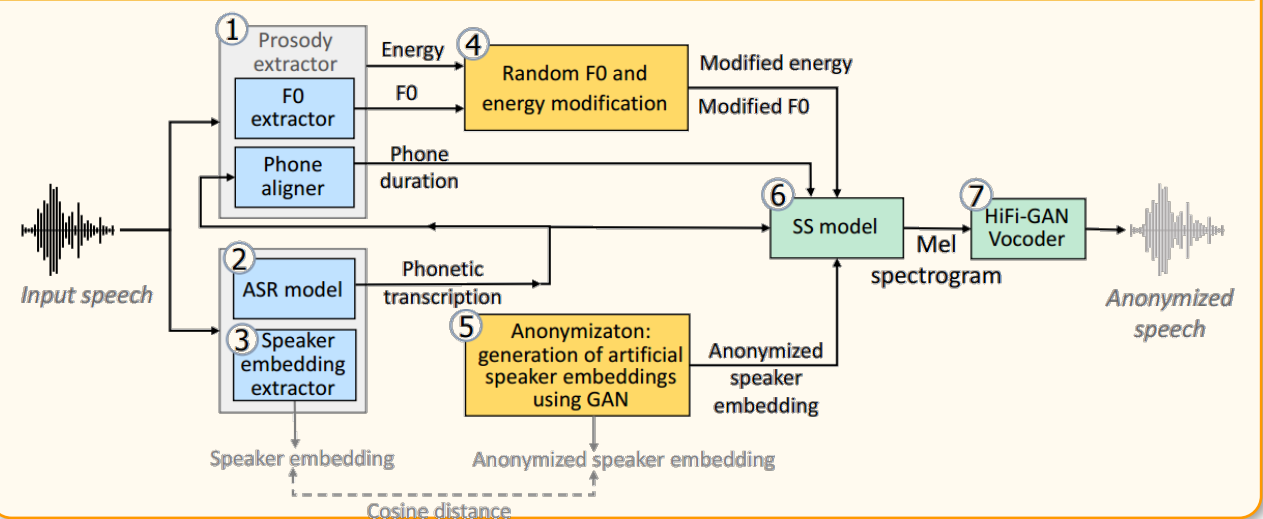
Baselines - Track 1

- Inherited from the 2024 challenge edition, but remove the legacy system B1 and excluded B6, as it is similar to B5 but not better

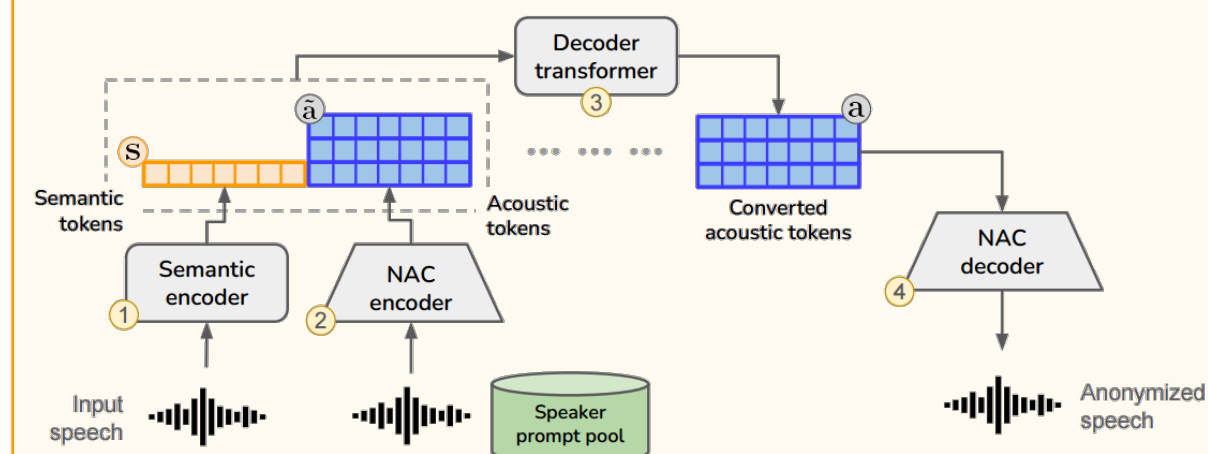
B2 McAdams coefficient



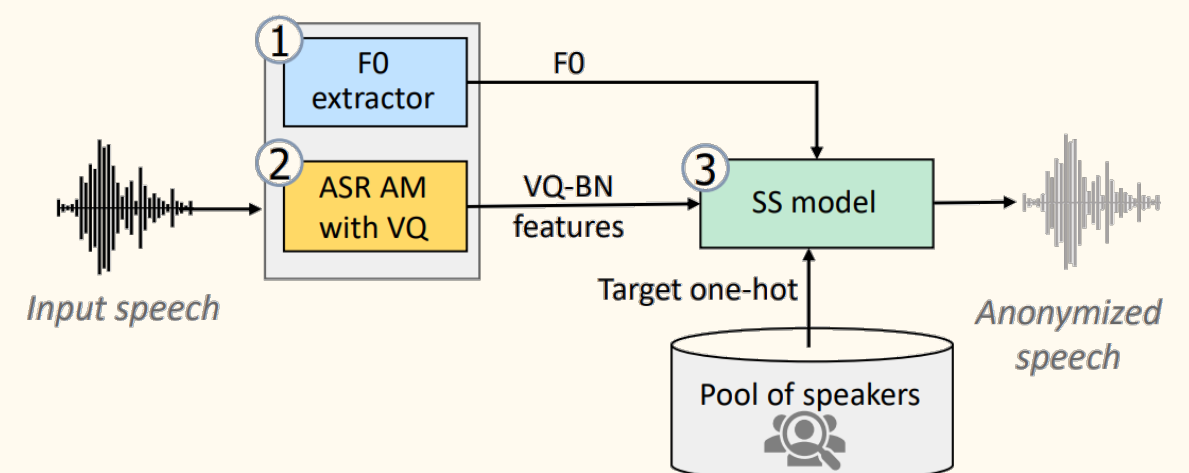
B3 phonetic transcriptions and a generative adversarial network



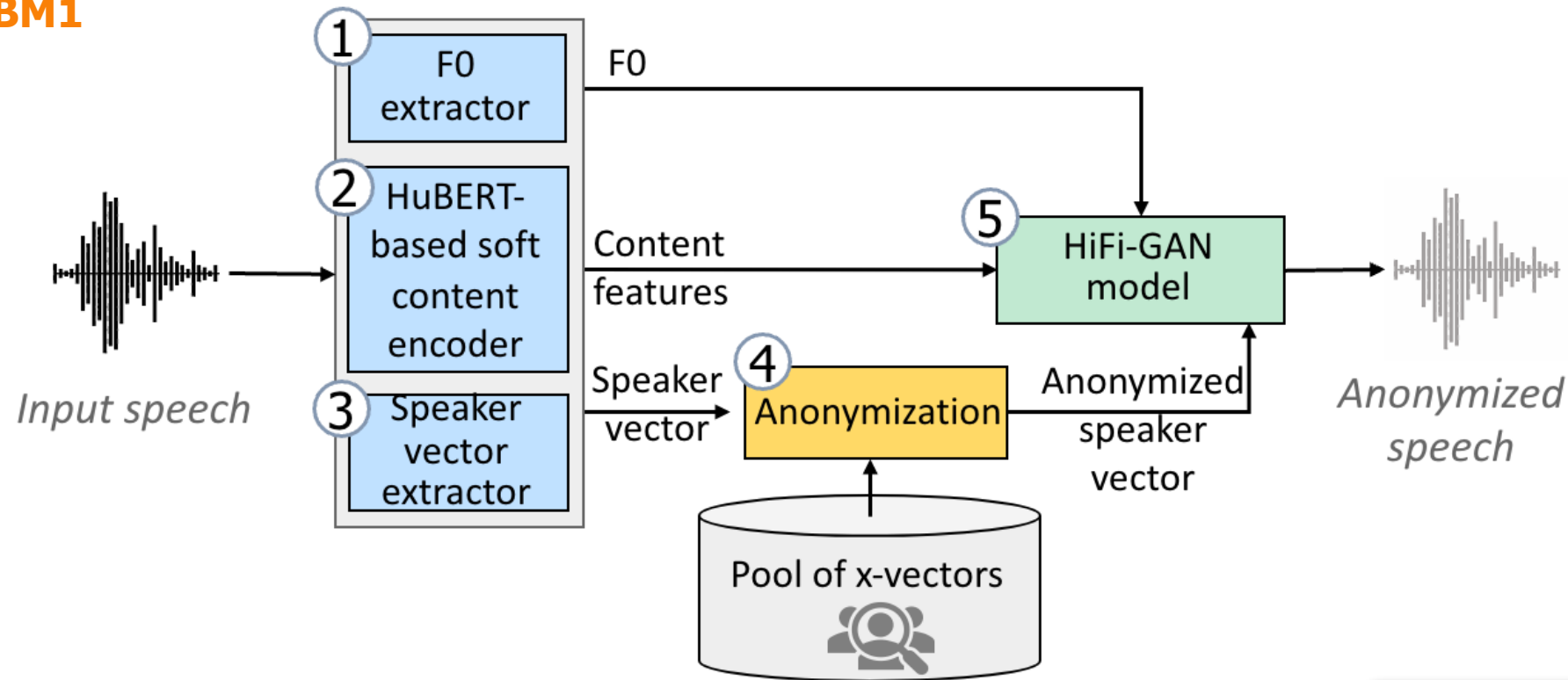
B4 neural audio codec language modeling



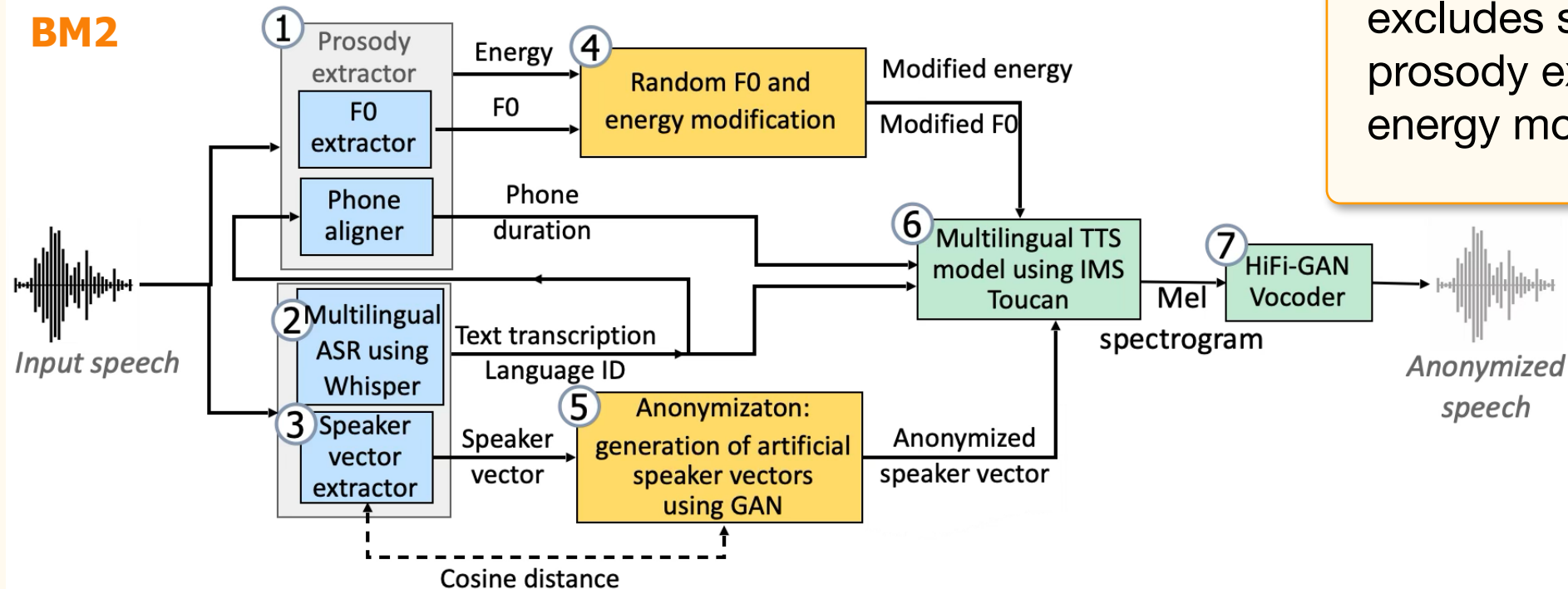
B5 ASR-BN with VQ and a pretrained wav2vec2 model



BM1



BM2



BM3

excludes steps 1 and 4, i.e., no prosody extraction or F0/energy modification is applied.

Participants

Registered teams: 40 (101 participants)
from 17 countries

Team submitted results: 13

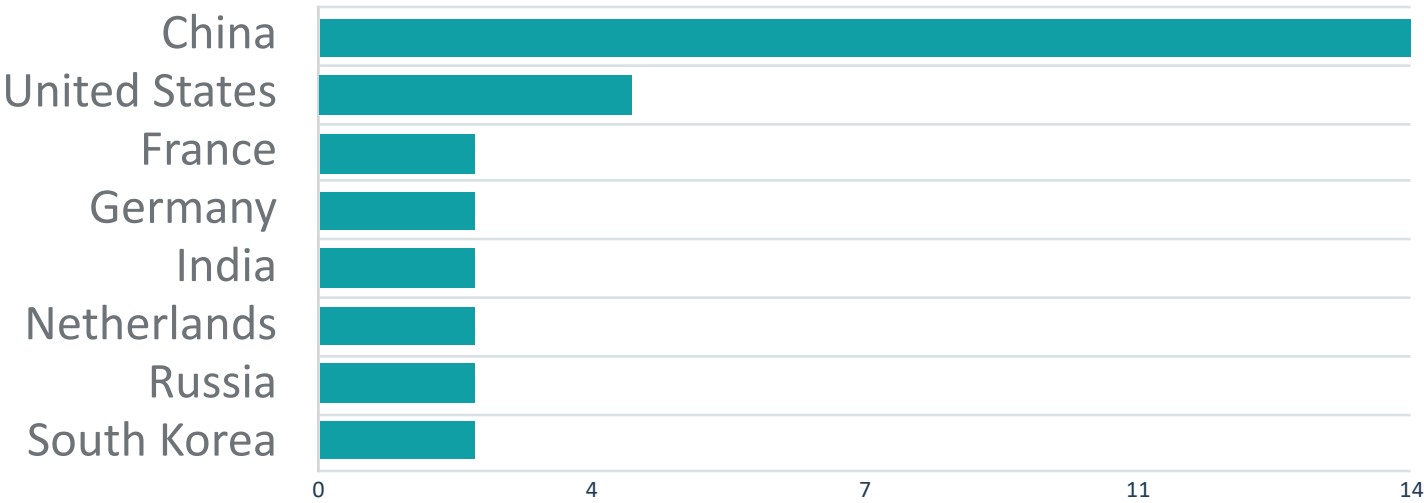
- 12 in Track 1
- 8 in Track 2

Submitted anonymization systems: 23

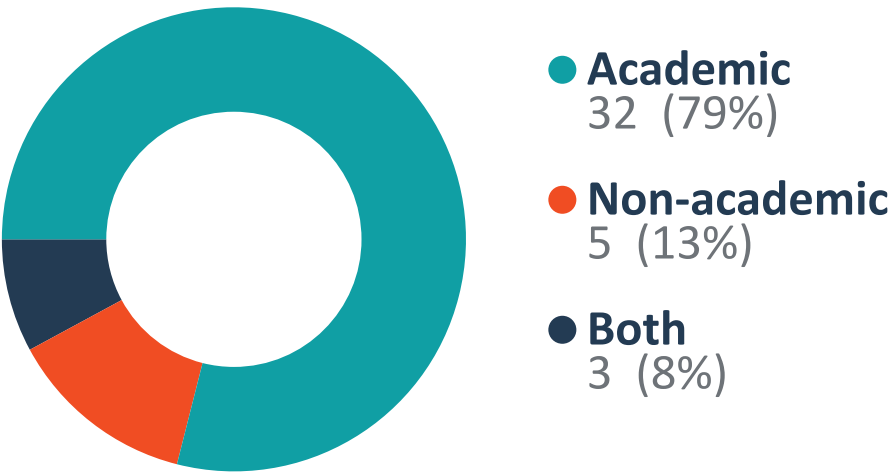
- 13 in Track 1
- 10 in Track 2

	2020	2022	2024	2026
Registered teams	25	43	40	40
Participants	> 45	>79	107	101
Countries	13	17	16	17
Submitted systems	16	16	36	23
Described systems			>48	
Teams	7	6	13	13

Registered teams by country / region



Organization type



Teams and systems

Team ID	Team	Affiliation	Track 1	Track 2
ASR + SER + TTS cascades				
T3	Voicecode	Voicecode Technology, Singapore, China	1	1
T7	AISRC	Soongsil University, South Korea	1	1
T15	Radboud-Aalto (ZZZ)	Radboud University (NL); Aalto University (FI)	2	0
T18	RocknRiver	JAIST, Japan	1	2
T24	Caramelo	UFG; Unicamp; UFMT; Unesp, Brazil	1	1
T32	Chameleos	ShanghaiTech University, China	1	0
T37	SHIVA	Orange Research, France	1	2
Disentanglement VC				
T1	Xihua-Voice-Processing	Xihua University, China	1	0
T2	PSST1	Aalto (FI); Ruhr Bochum & TU Berlin (DE); INESC-ID/IST (PT)	1	0
T11	AI4AI	Beijing University of Posts and Telecommunications, China	0	1
T12	M-PSIvacy	UGA; CNRS; Grenoble INP; LIG; LJK, France	1	0
T13	SpeechTek	Fondazione Bruno Kessler, Italy	1	1
Generative VC (hors concours)				
T23†	DKU	Duke Kunshan University; Xiaomi; CUHK Shenzhen, China	1	1
Total	13 teams 23 systems		12 teams 13 systems	8 teams 10 systems

See teams ids also at: <https://spsc-symposium.de/#program>

Submitted systems

Track 1

13 systems:
8 cascades + 4
disentanglement VC
+ 1 generative VC

- 1 ASR + SER + TTS cascades**
T3, T7
whisper + emotion2vec + VoxCPM2
T15-1, T15-2
Emotion-gated processing: Whisper + VITS for neutral speech.
Non-neutral: UUVc reconstruction (T15-1) or unchanged speech (T15-2)
T18
Whisper + emotion2vec + IMSToucan + HiFi-GAN
T24, T32, T37
Whisper + emotion2vec + Qwen3-TTS, VoxCPM2 or CosyVoice2

- 2 Disentanglement VC**
T1
wav2vec2 bottleneck + VQ + HiFi-GAN
T2*
HuBERT-soft + ECAPA-TDNN + flow-matching speaker anonymization + HiFi-GAN
T12
wav2vec2 + content quantization + HiFi-GAN
T13
Whisper features + ECAPA-TDNN + adversarially fine-tuned HiFi-GAN
- 3 Generative VC**
T23 not ranked
Seed-VC v2 with projected classifier-free guidance

Track 2

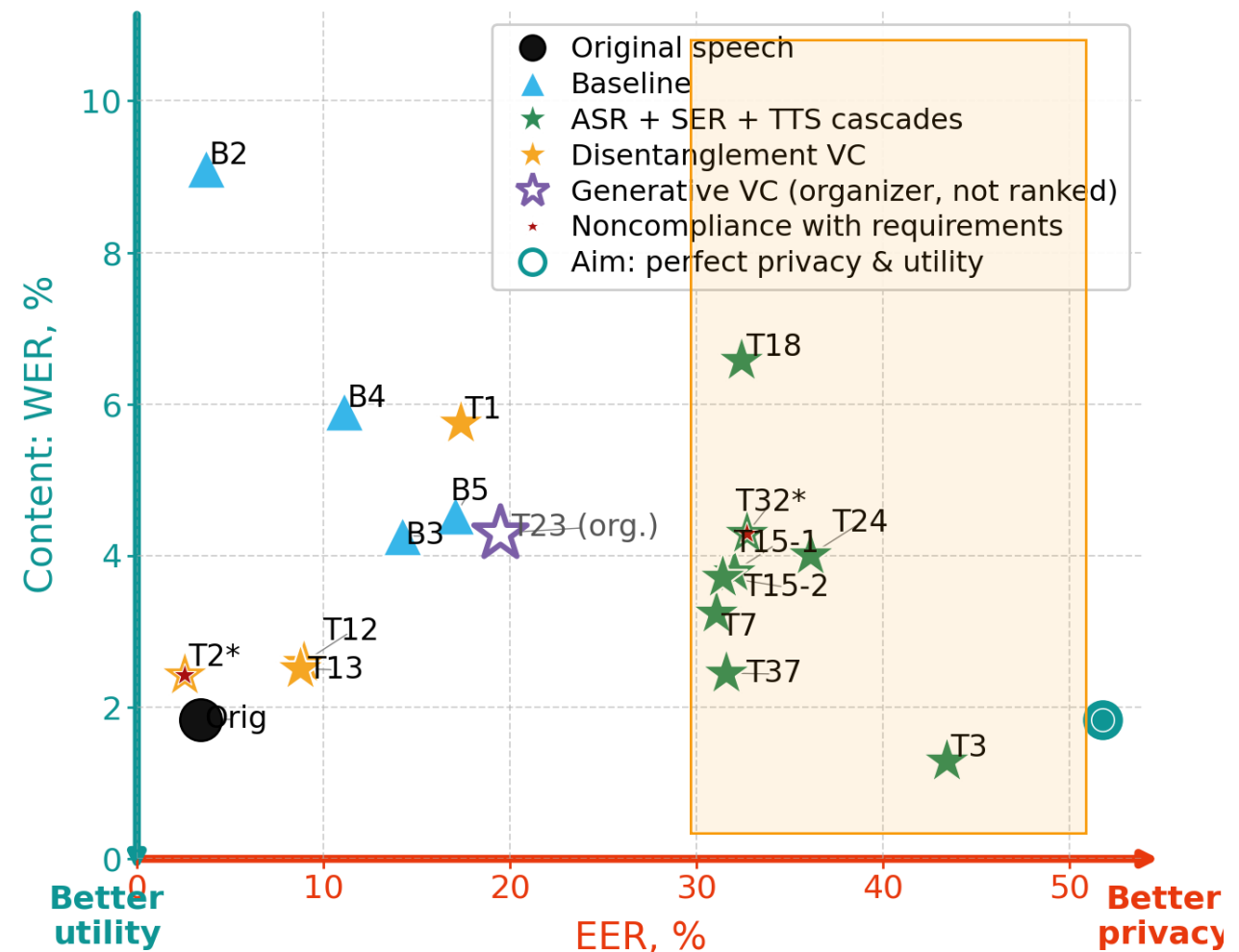
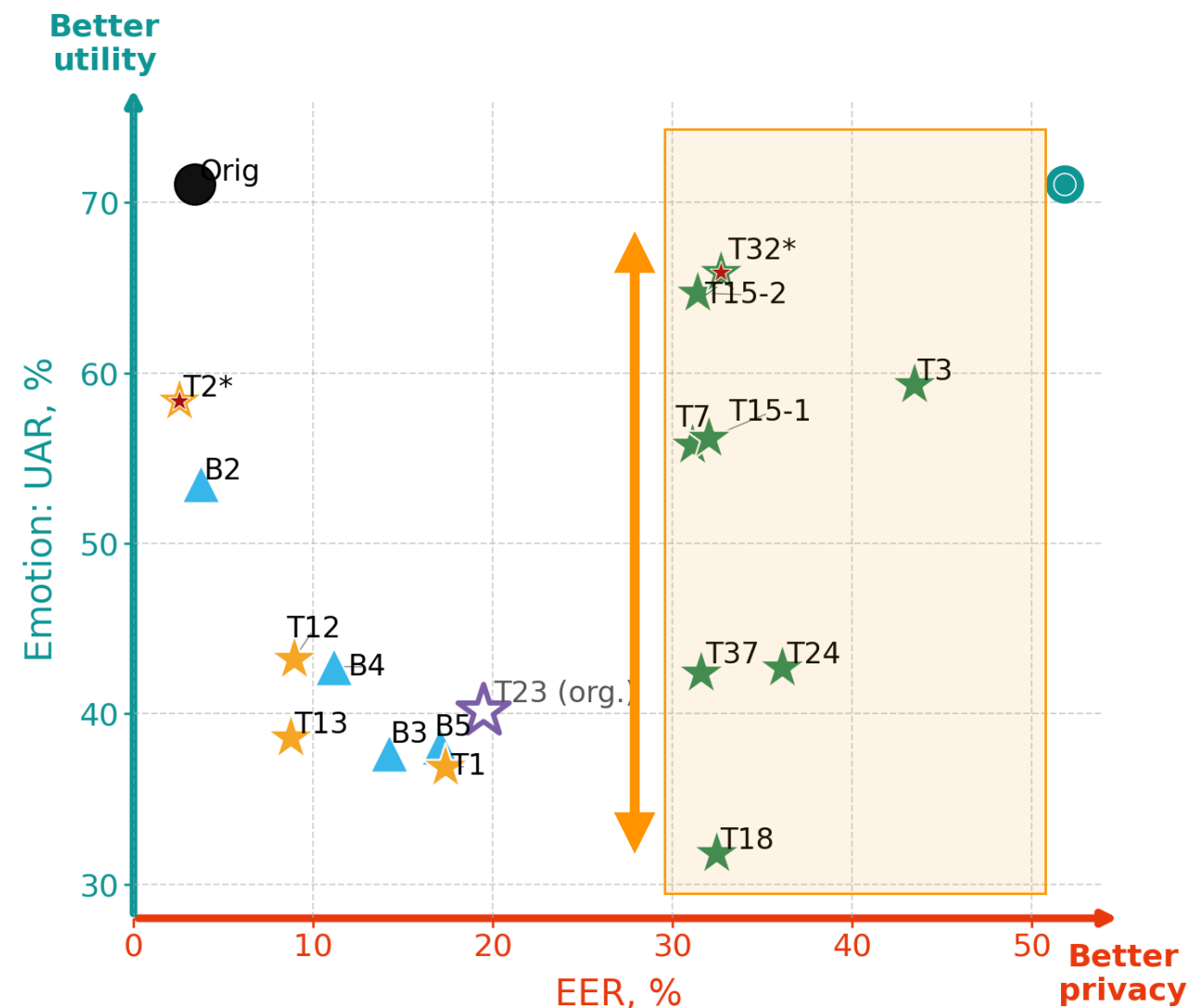
10 systems:
7 cascades + 2
disentanglement VC
+ 1 generative VC

- 1 ASR + SER + TTS cascades**
T3
Whisper + VoxCPM2 voice design
T7
Whisper + emotion2vec + VoxCPM2 with CAMEO references
T18-1, T18-2
Whisper + emotion2vec + IMSToucan + HiFi-GAN
T24, T37-1, T37-2
Whisper + VoxCPM2 or Qwen3-TTS, language-aware prompts or references

- 2 Disentanglement VC**
T11
Soft-HuBERT + ECAPA + HiFi-GAN
T13
mHuBERT-147 + ECAPA + emotion2vec + HiFi-GAN,
Both systems preserve multilingual content through VC features.
- 3 Generative VC**
T23 not ranked
Seed-VC v2 with projected classifier-free guidance

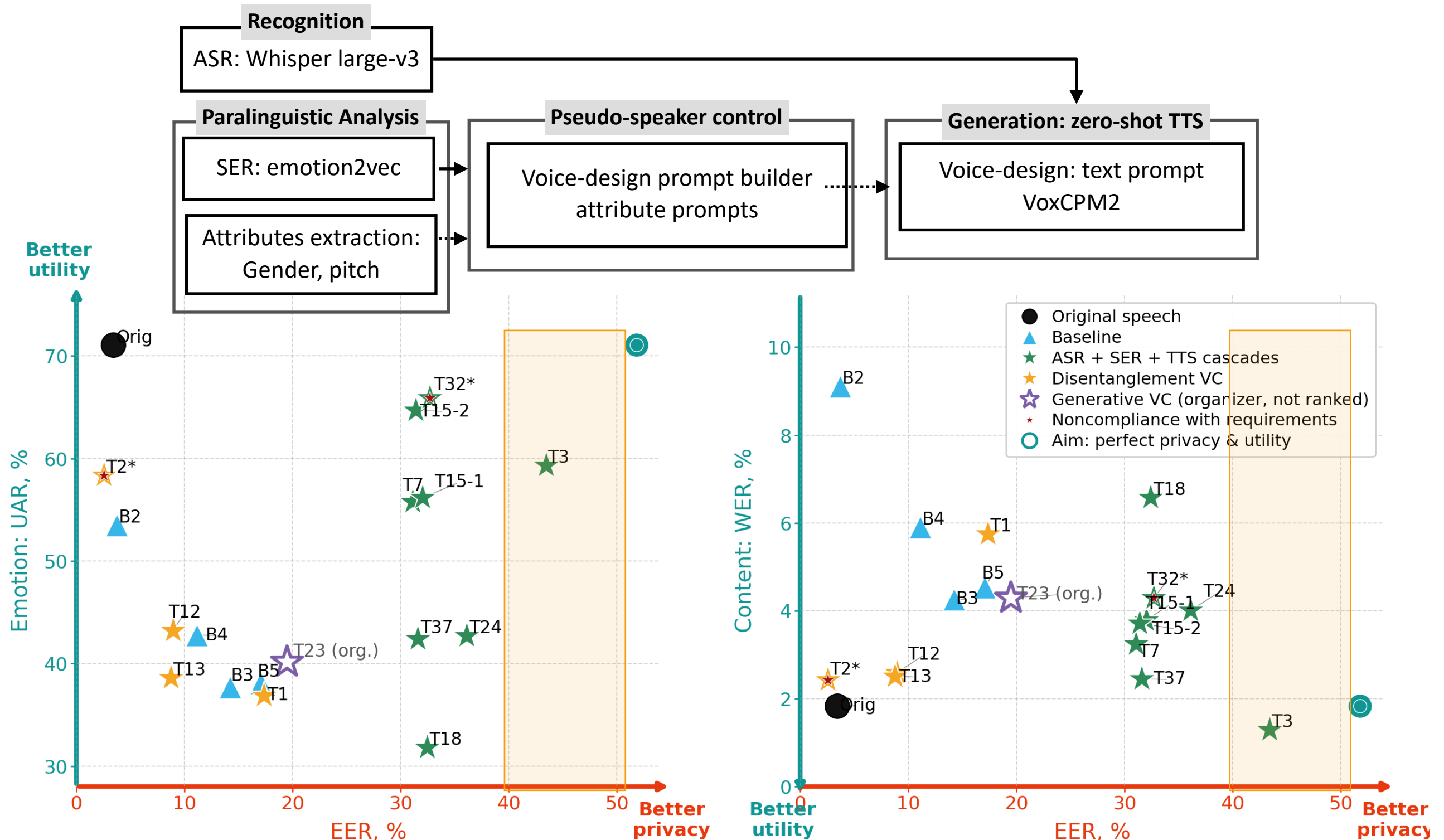
Track 1 results

- ASR + SER + TTS cascades dominate privacy, with all cascade systems exceeding 31%
- Emotion preservation varies substantially, so stronger privacy does not guarantee better utility



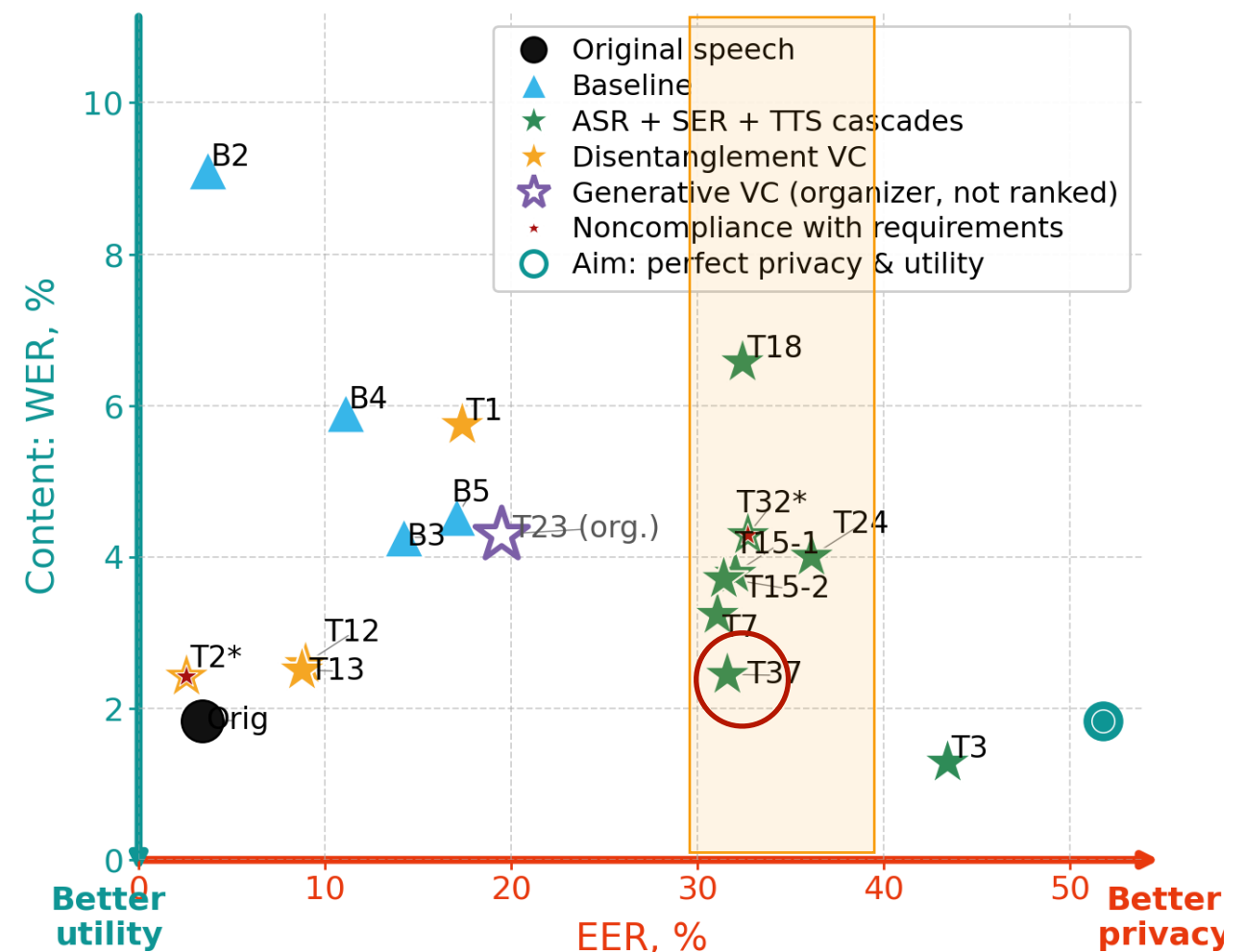
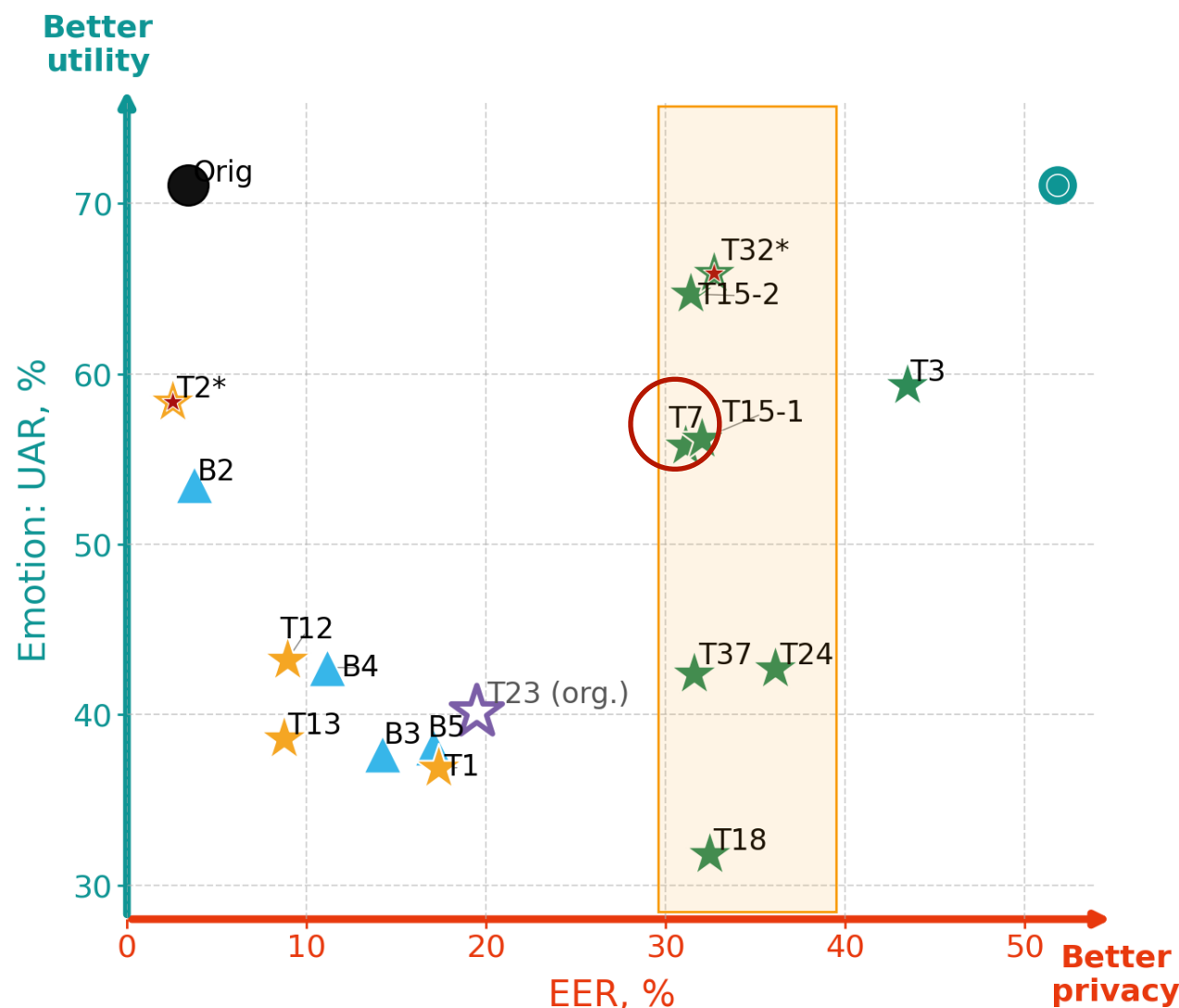
Track 1 results

- T3 is the only system in the $\geq 40\%$ EER category and records the lowest WER (1.29%)



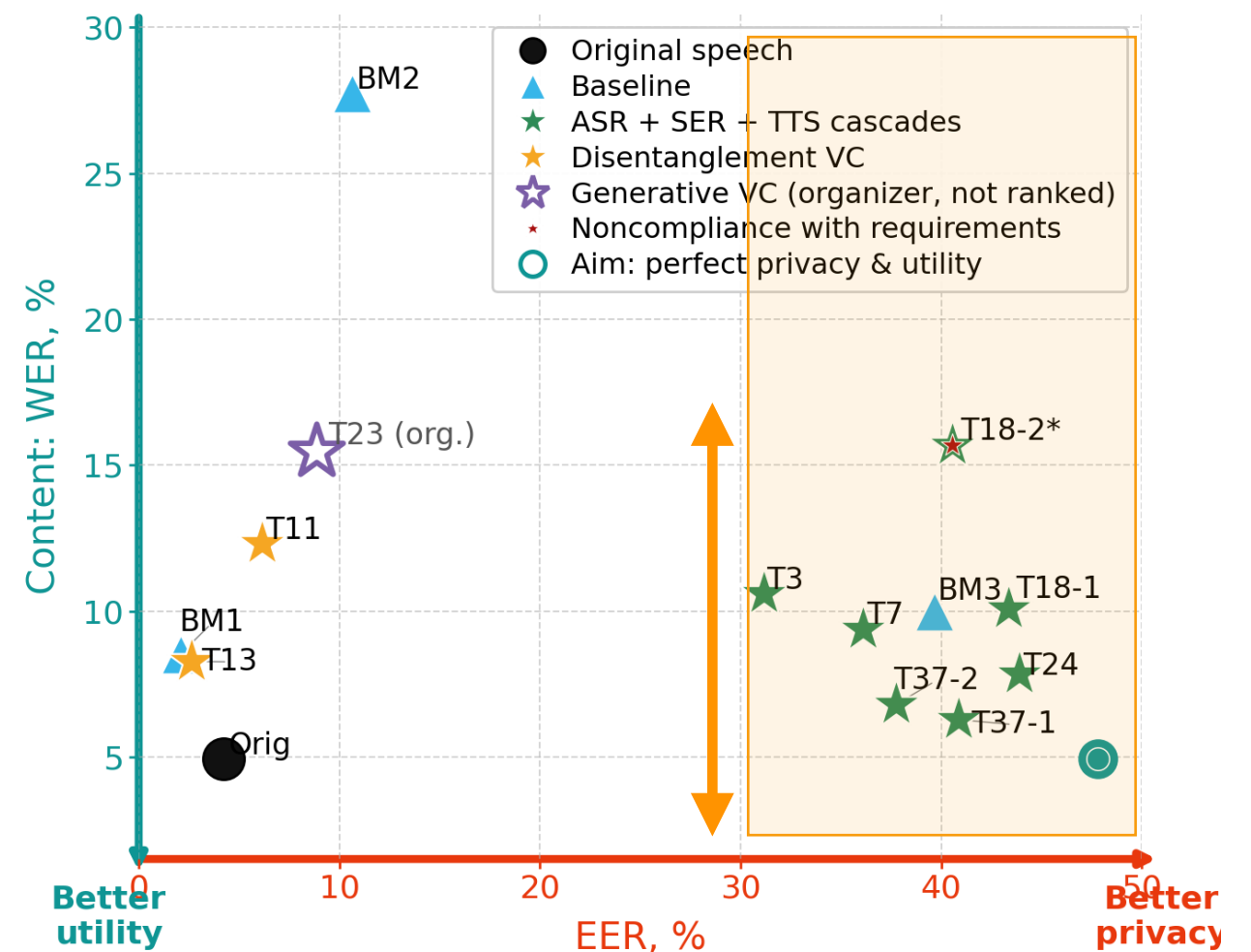
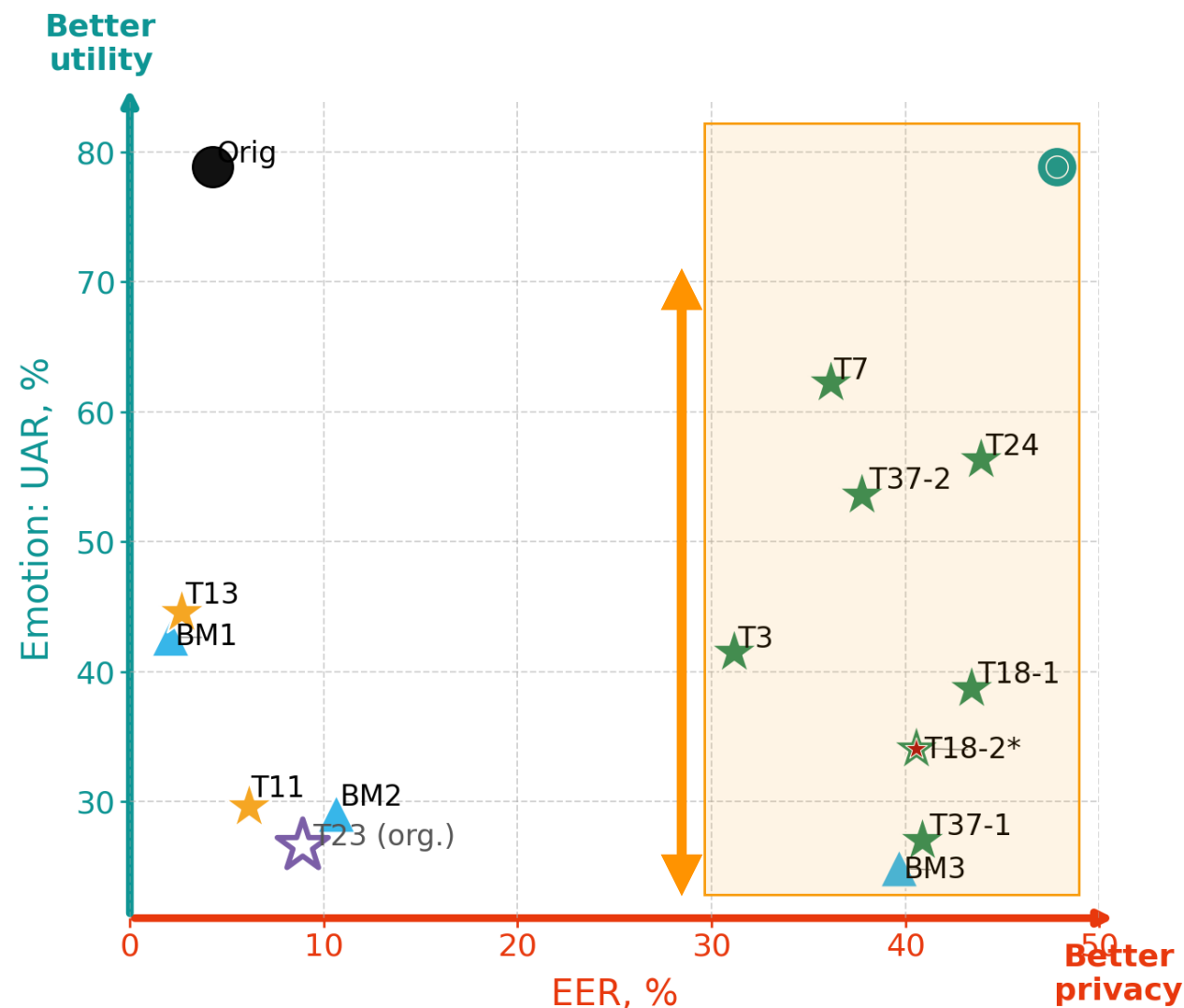
Track 1 results

- T37: Whisper + emotion2vec + Qwen3-TTS VoiceDesign with emotion instructions
 - the lowest WER (2.45%) among systems with $30\% < \text{EER} < 40\%$
- T7: Whisper + emotion2vec + VoxCPM2 with emotion-matched CREMA-D references and emotion instructions
 - the highest UAR (55.78%) among eligible systems with $30\% < \text{EER} < 40\%$



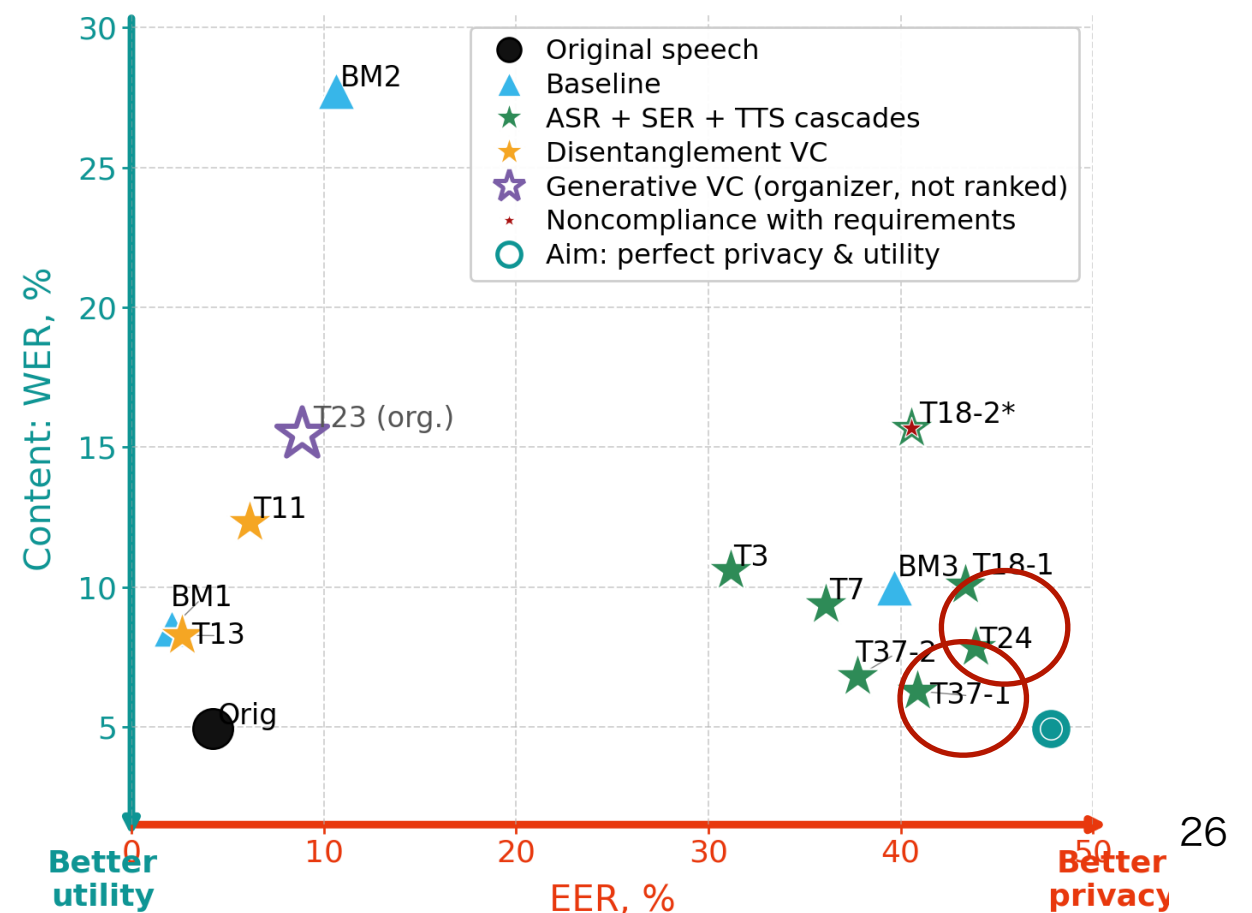
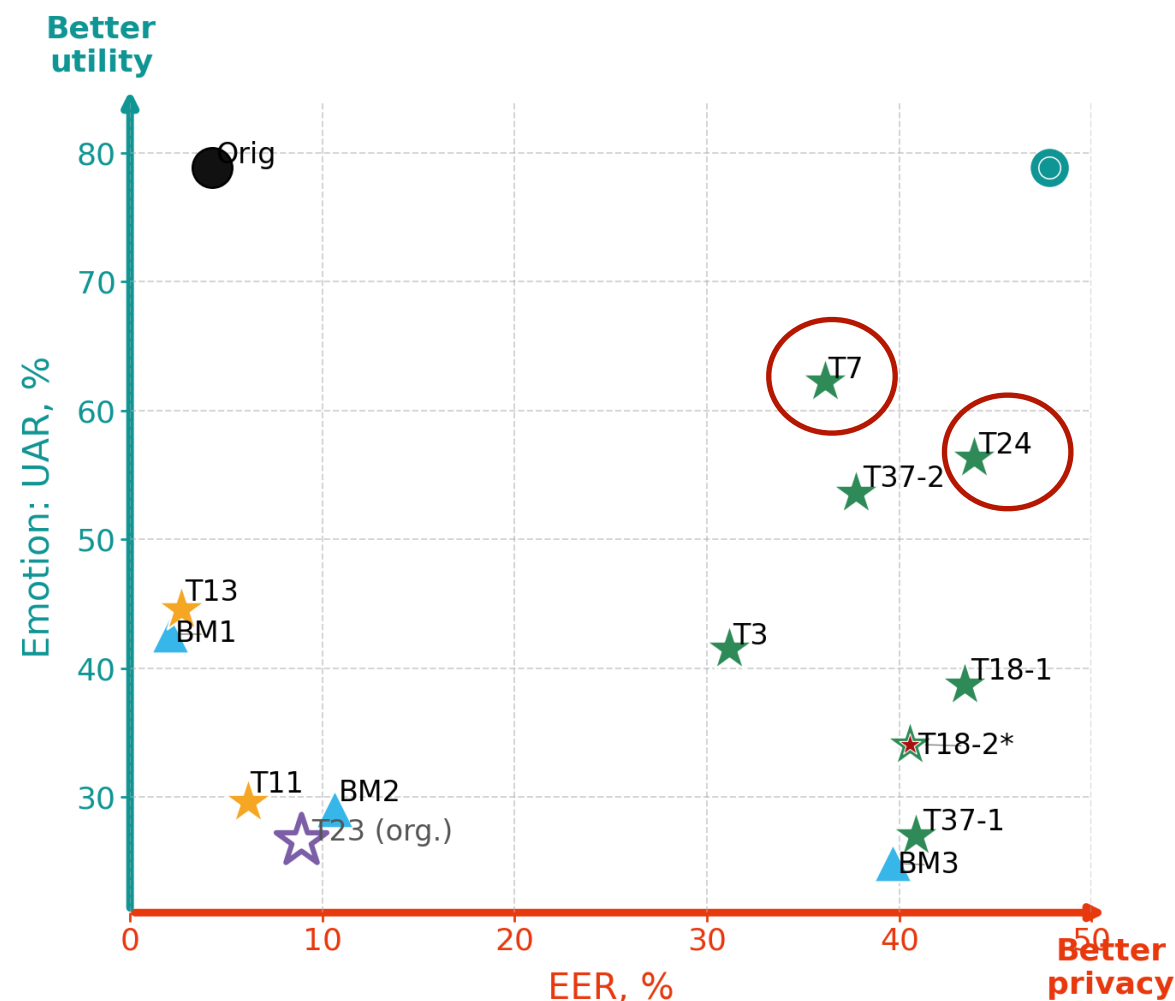
Track 2 results

- Cascaded systems achieve the strongest multilingual privacy, with several systems exceeding 40% EER
- Privacy does not ensure utility: WER and UAR vary substantially across systems and languages



Track 2 results

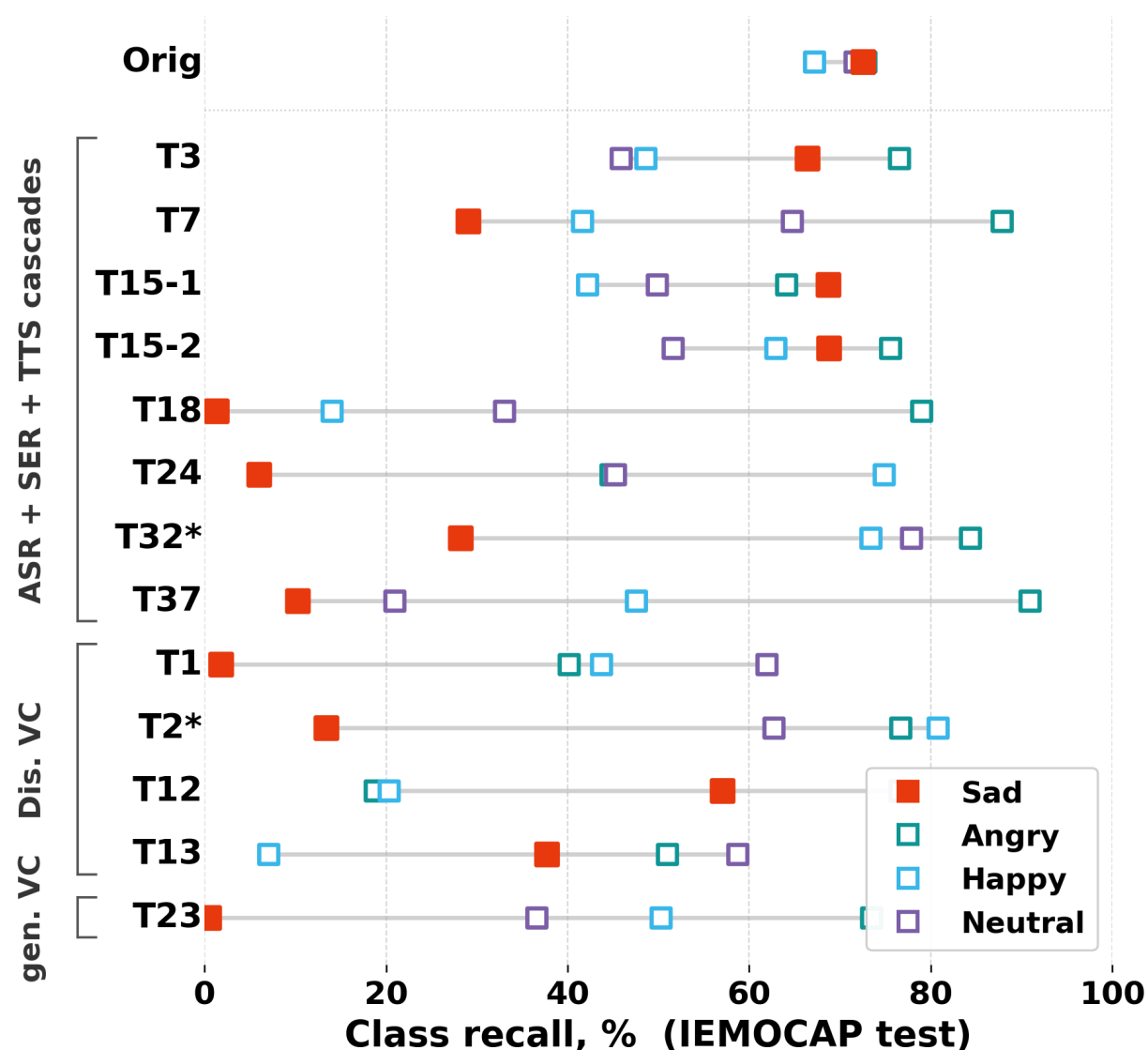
- T24: Whisper + emotion2vec + VoxCPM2 with language-matched emotional references
 - the highest UAR (56.34%) among systems with $EER \geq 40\%$
- T37-1: Whisper + Qwen3-TTS VoiceDesign with voice descriptions
 - the lowest WER (6.28%) among systems with $EER \geq 40\%$.
- T7: Whisper + emotion2vec + VoxCPM2 with language-matched CAMEO references and emotion instructions
 - the highest UAR (62.22%) among systems with $EER \geq 30\%$



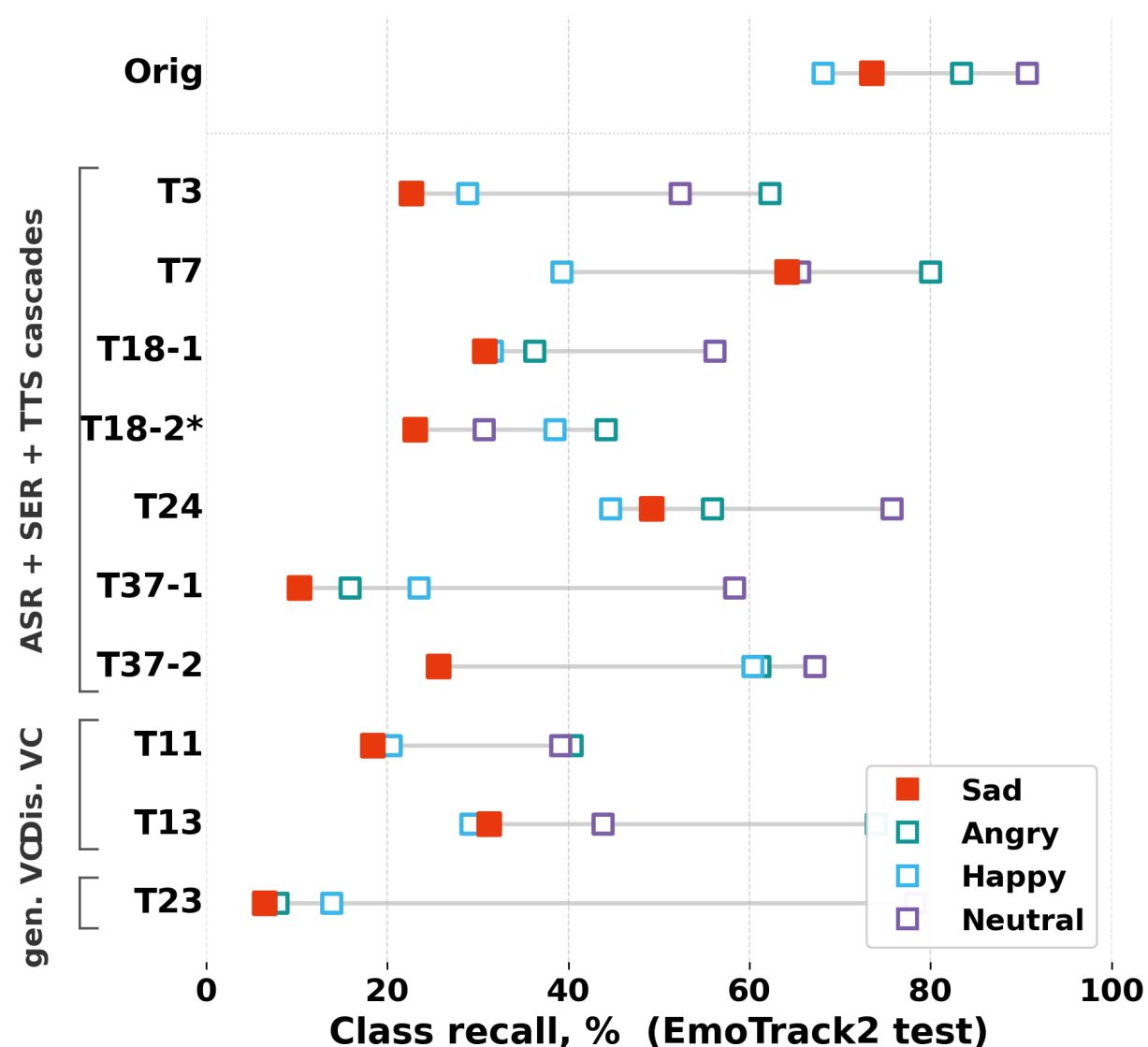
Emotion preservation by system

- Emotion preservation varies substantially across emotion classes and systems
- Several systems show particularly low recall for sadness

Track 1 emotion recall



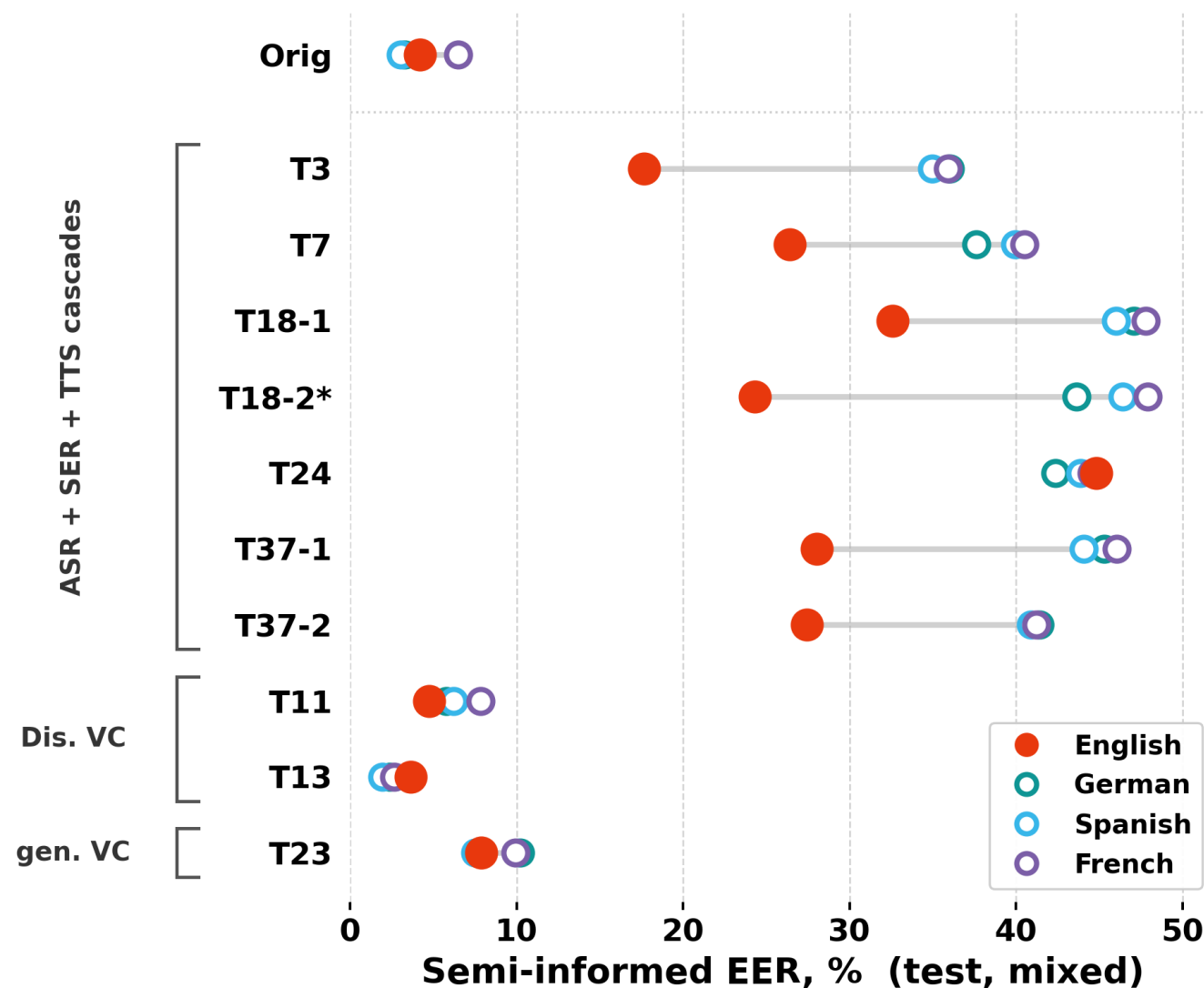
Track 2 emotion recall



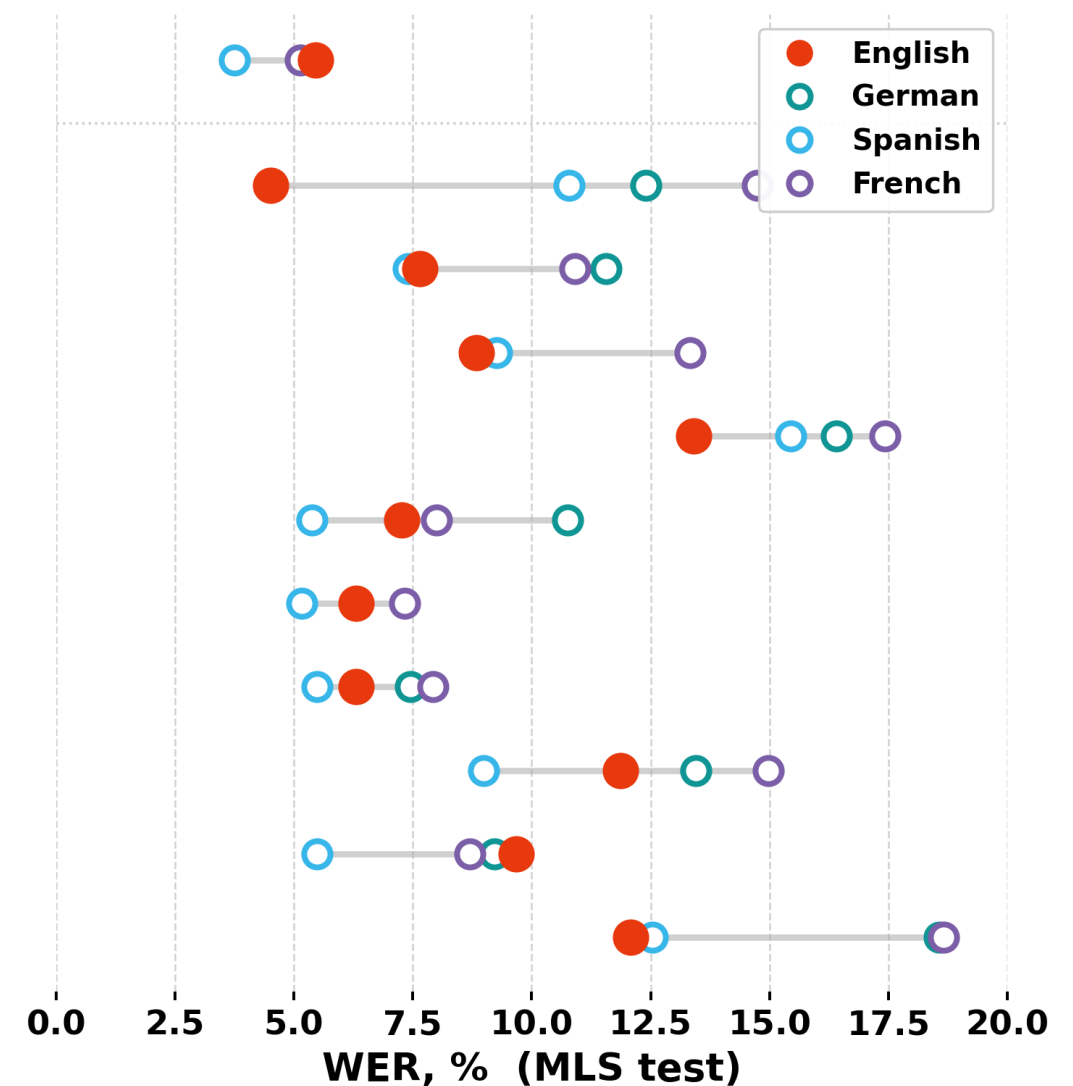
Content preservation by language

- Most cascaded systems show lower EER on English than on the other evaluated languages
- WER is generally higher in German and French than in English

(a) EER by language

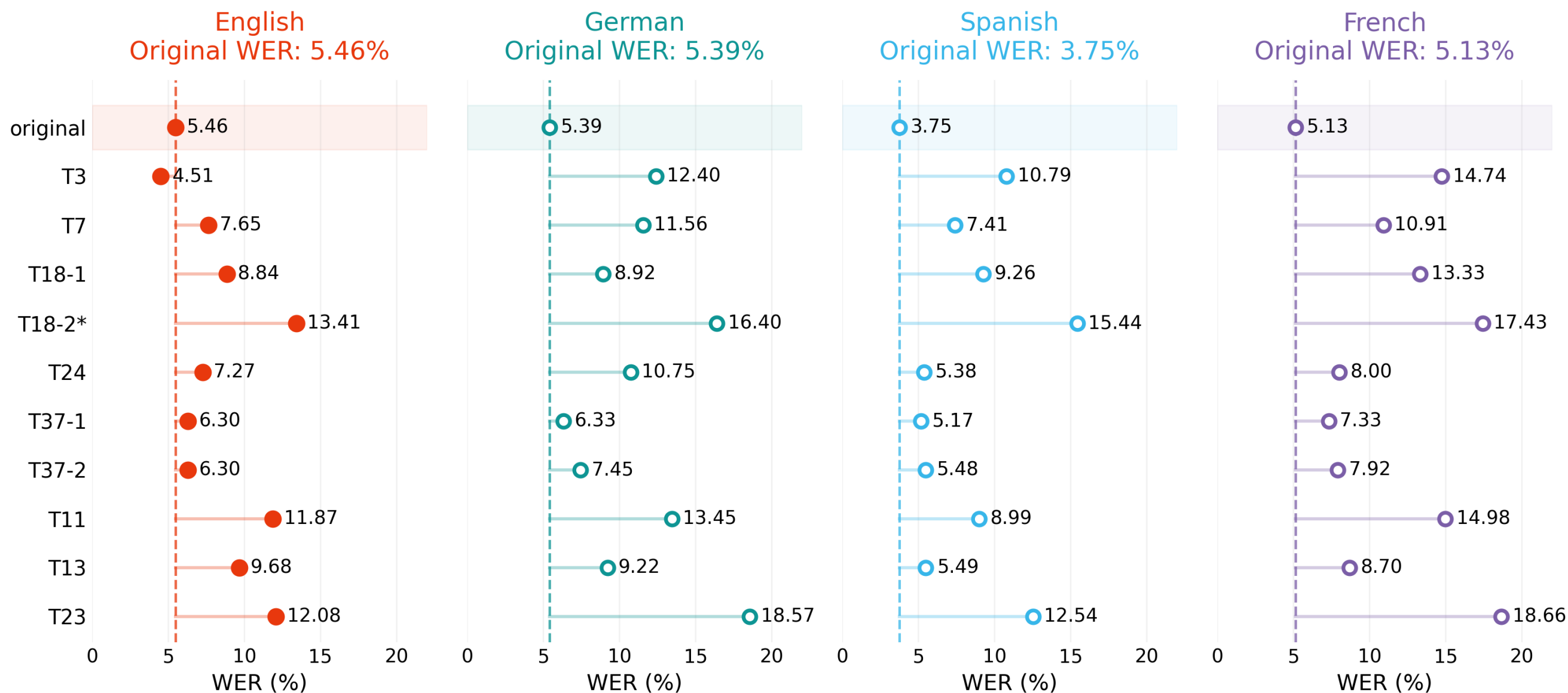


(b) WER by language



Content preservation by language

- For TTS-based cascades, content preservation is more challenging in German and French in these results
 - e.g. T3's WER is 4.51% in English, 12.40% in German, and 14.74% in French; original-data WERs were more similar across languages



Dashed line: original speech baseline in each language. Lower WER is better. | MLS test

Additional evaluation - attacker training data

- What if an attacker fine-tunes on the same team's English data from the other track, given that Tracks 1 and 2 contain different amounts of English data?

			ASV training data			ASV training data		
			Track 1 (LibriSpeech)			Track 2 (English)		
			ASV FT: train-clean-360, 23 GB			ASV FT: train-english, 4.3 GB		
			921 spk / 104,014 utt (~113/spk)			3,930 spk / 18,986 utt (~5/spk)		
ID	Track 1 system	Track 2 system	match (official)	cross	Δ	match (official)	cross	Δ
(1) ASR + SER + TTS cascades								
T3	voxcpm2 vd	voxcpm2 vd	49.74	49.07	-0.67	19.33	11.84	-7.48
T7	emotts	emotts	33.39	38.20	+4.81	27.50	15.25	-12.24
T18	feats8	T18-1 (final_6)	34.70	43.70	+9.00	35.23	18.57	-16.66
T18	feats8	T18-2 (final_3)	34.70	47.14	+12.44	25.42	22.81	-2.61
T24	ensemble2	voxcpm2_top50	36.82	47.54	+10.72	45.21	45.06	-0.15
T37	VD-E2V-T1	VD-E2V-T2	34.18	40.64	+6.46	27.99	15.53	-12.46
T37	VD-E2V-T1	VD-Pool-T2	34.18	41.04	+6.86	30.09	15.72	-14.37
(2) Disentanglement VC								
T13	adv_ep170	unit_ml	12.56	47.54	+34.98	4.30	14.84	+10.54
(3) Generative VC (hors concours)								
T23	smooth21	global	22.35	23.75	+1.40	9.26	9.66	+0.41

Δ = cross – match, Negative Δ indicates a stronger cross-track attacker (lower EER)

Additional evaluation - attacker training data

- Attacker training data has a substantial and asymmetric effect on privacy evaluation
 - For Track 1, the official attacker, fine-tuned on Track 1 anonymized data, is usually stronger than the cross-track attacker
 - For Track 2, the Track 1 attacker is often stronger than the official Track 2 attacker
 - Differences in training-data volume and per-speaker coverage may contribute to the observed attacker performance

Track 1 (LibriSpeech)	Track 2 (English)
ASV FT: train-clean-360, 23 GB	ASV FT: train-english, 4.3 GB
921 spk / 104,014 utt (~113/spk)	3,930 spk / 18,986 utt (~5/spk)

- T13 is a notable exception: the official attacker is stronger in both tracks, possibly because its Track 1 and Track 2 anonymization systems differ

ID	Track 1 system	Track 2 system	match (official)	cross	Δ	match (official)	cross	Δ
T13	adv_ep170	unit_ml	12.56	47.54	+34.98	4.30	14.84	+10.54

Additional evaluation - content leakage analysis

- Why do ASR + SER + TTS cascades fail to achieve an EER close to 50% ?
Can linguistic content support speaker linking in VPC 2026?
- Prior studies have demonstrated speaker linking from linguistic content and its use in multimodal attacks [2–4]
- We revisit text-only speaker verification on VPC 2026 using a BGE encoder with language-specific speaker-classification heads and cosine scoring on the official trials,
 - similar to [2], using BGE instead of BERT and evaluating both English and multilingual VPC 2026 data

1 Text representation

BGE encoder with a language-specific speaker-classification head.

2 Language-specific training

One class per training speaker
Track 1: train-clean-360
Track 2: train_{english, german, spanish, french}

3 VPC trial scoring

Use the official trial pairs and original enrollment/test texts; cosine similarity produces the EER.

[2] Ünal Ege Gaznepoglu, Anna Leschanowsky, Ahmad Aloradi, Prachi Singh, Daniel Tenbrinck, Emanuël A. P. Habets, and Nils Peters, “You Are What You Say: Exploiting Linguistic Content for VoicePrivacy Attacks,” in *Proc. Interspeech*, 2025, pp. 4238–4242.

[3] Ahmad Aloradi, Ünal Ege Gaznepoglu, Emanuël A. P. Habets, and Daniel Tenbrinck, “VoxATtack: A Multimodal Attack on Voice Anonymization Systems,” in *Proc. IEEE WASPAA*, 2025, pp. 1–5.

[4] Ashi Garg, Cristina Aggazzotti, Leibny Paola García-Perera, and Nicholas Andrews, “Multimodal Speaker Verification as a Threat to Speaker Anonymization,” *Proc. Interspeech*, 2026

Additional evaluation - content leakage analysis

- Why do ASR + SER + TTS cascades fail to achieve an EER close to 50% ?
Can linguistic content support speaker linking in VPC 2026?

Track 1 · English text-based EER

34.81%

Track 2 · Text-based EER by language

Language	English	German	Spanish	French
EER (%)	19.1	25.6	26.3	26.1

- Ideally, text-only speaker linking should give an EER near 50%, indicating no remaining link to speaker identity
- The VPC evaluation data contain linguistic cues that leak speaker identity. Even after acoustic voice resynthesis, speaker-specific text and trial patterns can remain linkable.

Findings

- From participants results

- ASR + SER + TTS cascades achieved the strongest privacy, but emotion preservation varied
- All submitted disentanglement VC systems achieved EERs below 20% under the official attacker
- Performance varied across emotion classes and languages

- From additional analysis

- Changing the attacker's training data substantially changed EER for several systems.
- Text-only verification revealed speaker-correlated linguistic content in the evaluation data

Congratulations!



T3 – Voicecode

Overall / Best trade-off — Track 1

Whisper + VoxCPM2 voice design with emotion-guided prompts and SER-aware candidate selection.

VOICECODE TECHNOLOGY PTE. LTD.

Sheng Zhao, Yibo Duan, Zihan Chai, Xuyang Xing, Maquan Yang, Shengchen Li

T24 – Caramelo

Overall / Best trade-off — Track 2

Whisper + VoxCPM2 with language-matched references, sampled from the top-50 emotion-similar candidates.

Federal University of Goiás (UFG), Unicamp, Federal University of Mato Grosso (UFMT), and São Paulo State University (Unesp), Brazil

Alef Iury Siqueira Ferreira, Lucas Hideki Ueda, Lucas Rafael Stefanel Gris, Frederico Santos de Oliveira, Augusto Seben da Rosa, Alexandre Costa Ferro Filho, and Victor Moreno

T37 – SHIVA

Best content preservation — Tracks 1 & 2

Whisper + Qwen3-TTS VoiceDesign: emotion instructions (Track 1);
random voice-description prompts (Track 2, T37-1).

Orange Research, France

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

T7 – AISRC

Best emotion preservation — Tracks 1 & 2

Whisper + emotion2vec + VoxCPM2 with emotion-matched references
(CREMA-D for English; CAMEO for the other languages).

Soongsil University, South Korea

Ngoc Hung Le, Kyujin Kim, Sangjun Park, and Souhwan Jung

Feedback on future challenges

- To help us organize future editions of the VoicePrivacy Challenge, we would greatly appreciate your feedback on participation in the VoicePrivacy Challenge 2026

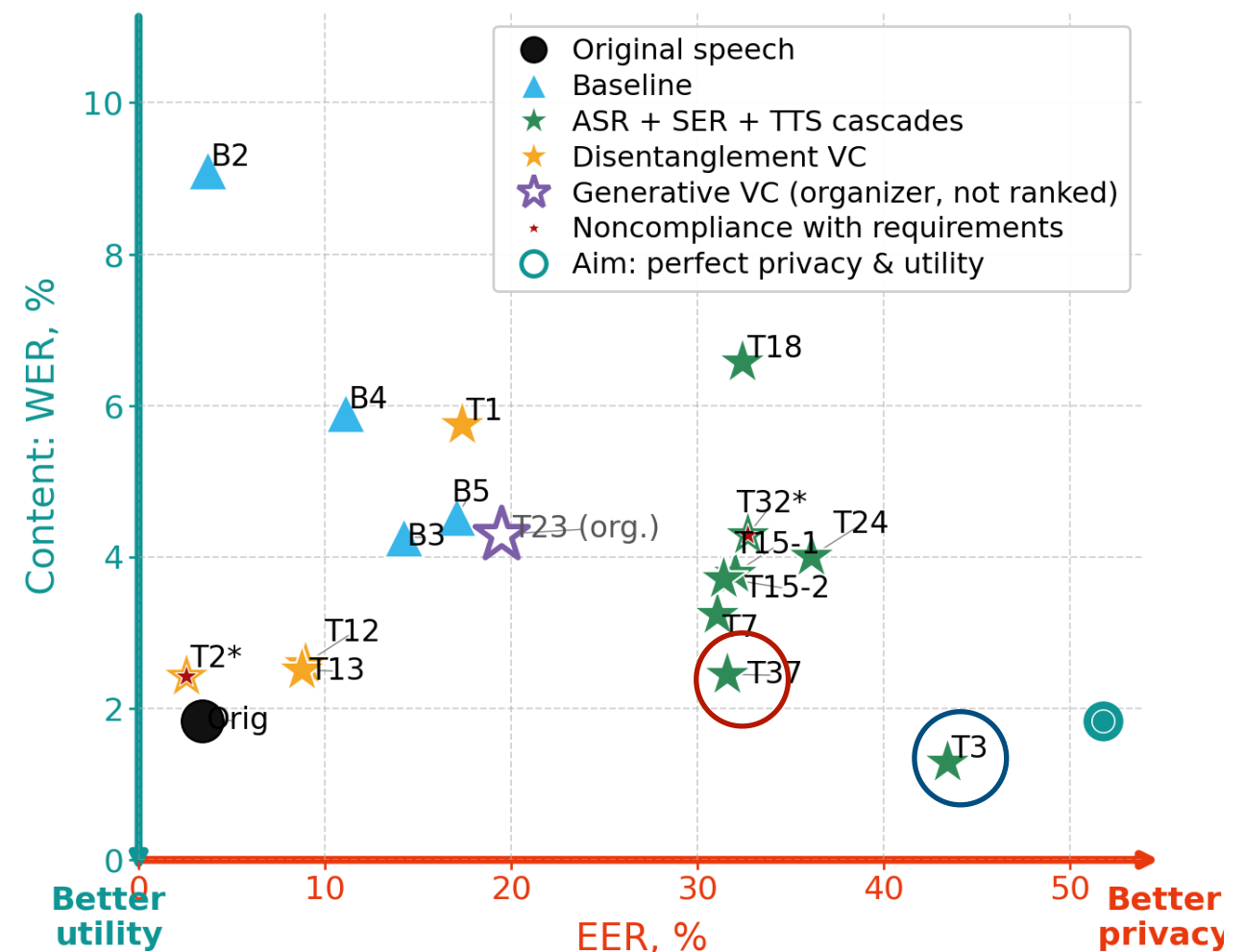
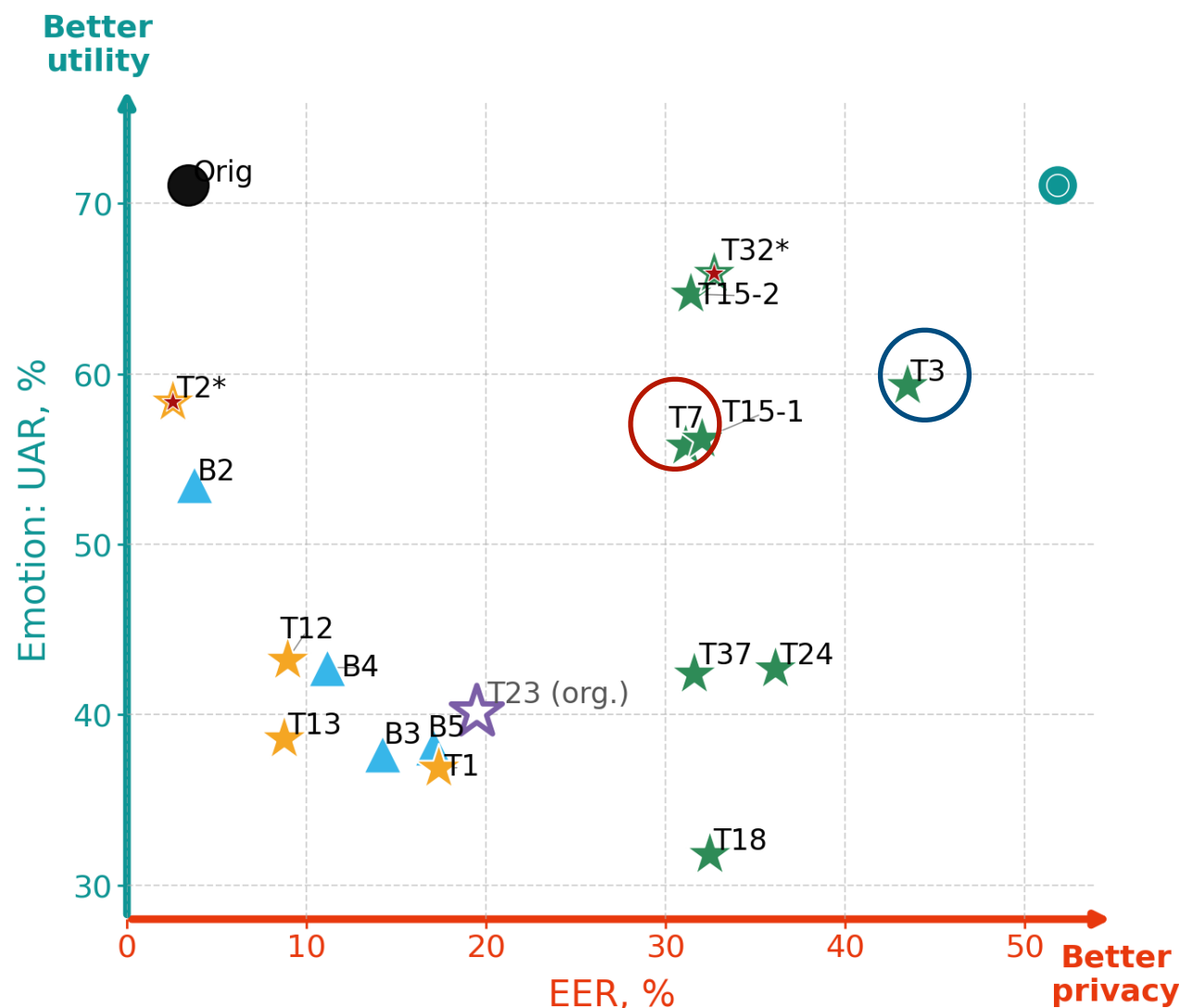


References: participants' papers

- [1] **T3 Voicecode** *VoxCPM2 Voice-Design Anonymization for VoicePrivacy 2026: English and Multilingual Systems*. Yibo Duan, Sheng Zhao, Xuyang Xing
- [2] **T7 AISRC** *AISRC System Description for Voice Privacy 2026 Challenge*. Ngoc Hung Le, Kyujin Kim, Sangjun Park, Souhwan Jung
- [3] **T15 Radboud-Aalto (ZZZ)** *Radboud-Aalto Submission to the Voice Privacy Challenge 2026*. Jiusi Zheng, Martha Larson, Tom Bäckström
- [4] **T18 RocknRiver** *RocknRiver: Multilingual Speaker Anonymization System for Voice Privacy 2026 Challenge*. Gia-Huy Bui, Minh-Xuan Phan, Quoc-Huy Nguyen, Candy Olivia Mawalim, Natthawut Kertkeidkachorn, Masashi Unoki
- [5] **T24 Caramelo** *Emotion-Aware Speaker Anonymization with Zero-Shot TTS Foundation Models for VoicePrivacy 2026*. Alef Iury Siqueira Ferreira, Lucas Hideki Ueda, Lucas Rafael Stefanel Gris, Frederico Santos de Oliveira, Augusto Seben da Rosa, Alexandre Costa Ferro Filho, Victor Moreno
- [6] **T32 Chameleos** *Two-Stage Speaker Anonymization with Emotion-Conditioned Resynthesis and Mild McAdams Post-Processing: A CosyVoice2 System for the VoicePrivacy 2026 Challenge*. Jiarui Xu, Zhiyu Wan, Sihan Xie
- [7] **T37 Orange SHIVA** *Natural Language Instructions for Voice Anonymization: Orange Research Submission for the Voice Privacy 2026 Challenge*. Rachid Benhsina, Olivier Le Blouch, Rayane Bakari, Nicolas Gengembre
- [8] **T1 Xihua-Voice-Processing** *FAR-CAP-DSF System for the VoicePrivacy 2026 Challenge*. Hanbing Tian, Yi Guo, Fuqiang Hu
- [9] **T2 PSST1** *PSST1 System Description for VoicePrivacy Challenge 2026*. Théo Nguyen, Ben Luks, Tom Bäckström, Rainer Martin
- [10] **T11 AI4AI** *Enhancing Selection-Based Voice Anonymization via Soft-HuBERT and Statistical F0 Transformation*. Yilin Han, Ruixi Jiang, Tianyu Song, Lijie Liu, Qihang Lu, Wenxin Fu, Yingming Gao, Ya Li
- [11] **T12 M-PSIvacy** *Listen to the Features: Voice Anonymization Driven by Content Embedding Matching over Signal Reconstruction*. Adrien Schneider, Kacper Zabkowski, Anderson Augusma, Frédérique Letué, Maria Camila Pinzon, Dominique Vaufreydaz
- [12] **T13 FBK-SpeechTek** *Discrete Units, Adversarial Vocoding, and the Identity That Remains: The FBK-SpeechTek Systems for VoicePrivacy 2026*. Siavosh Sepanta, Marco Matassoni, Alessio Brutti
- [13] **T23 DKU Voice-Privacy-Team** *A Training-Free Projection-Guided Conditional Flow-Matching System for VoicePrivacy 2026*. Xiang Shi, Han Zhu, Ming Li, Xiaoxiao Miao

Track 1 results

- T3 is the only system in the $\geq 40\%$ EER category and records the lowest WER (1.29%)
- T37 (SHIVA): Whisper + emotion2vec + Qwen3-TTS VoiceDesign with emotion instructions
 - the lowest WER (2.45%) among systems with $30\% < \text{EER} < 40\%$
- T7 (AISRC): Whisper + emotion2vec + VoxCPM2 with emotion-matched CREMA-D references and emotion instructions
 - the highest UAR (55.78%) among eligible systems with $30\% < \text{EER} < 40\%$



Track 2 results

- T24 (Caramelo): Whisper + emotion2vec + VoxCPM2 with language-matched emotional references
 - the highest UAR (56.34%) among systems with EER $\geq 40\%$
- T37-1 (SHIVA): Whisper + Qwen3-TTS VoiceDesign with voice descriptions
 - the lowest WER (6.28%) among systems with EER $\geq 40\%$.
- T7(AISRC): Whisper + emotion2vec + VoxCPM2 with language-matched CAMEO references and emotion instructions
 - the highest UAR (62.22%) among systems with EER $\geq 30\%$

