The VoicePrivacy 2024 Challenge

4th Symposium on Security and Privacy in Speech Communication

6th September 2024
Kos Island, Greece

Natalia Tomashenko, Xiaoxiao Miao, Pierre Champion, Sarina Meyer, Xin Wang, Emmanuel Vincent, Michele Panariello, Nicholas Evans, Junichi Yamagishi, Massimiliano Todisco

Japan Science and Technology Agency

Organizers

Natalia Tomashenko
Inria, France

Xiaoxiao Miao
Singapore Institute of Technology, Singapore

Pierre Champion
Inria, France

Sarina Meyer
University of Stuttgart, Germany

Xin Wang
NII, Japan

Emmanuel Vincent
Inria, France

Michele Panariello
EURECOM, France

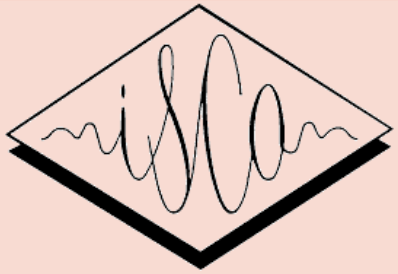
Nicholas Evans
EURECOM, France

Junichi Yamagishi
NII, Japan

Massimiliano Todisco
EURECOM, France

Acknowledgment

ISCA



Sponsor

nijta

Challenge Participants

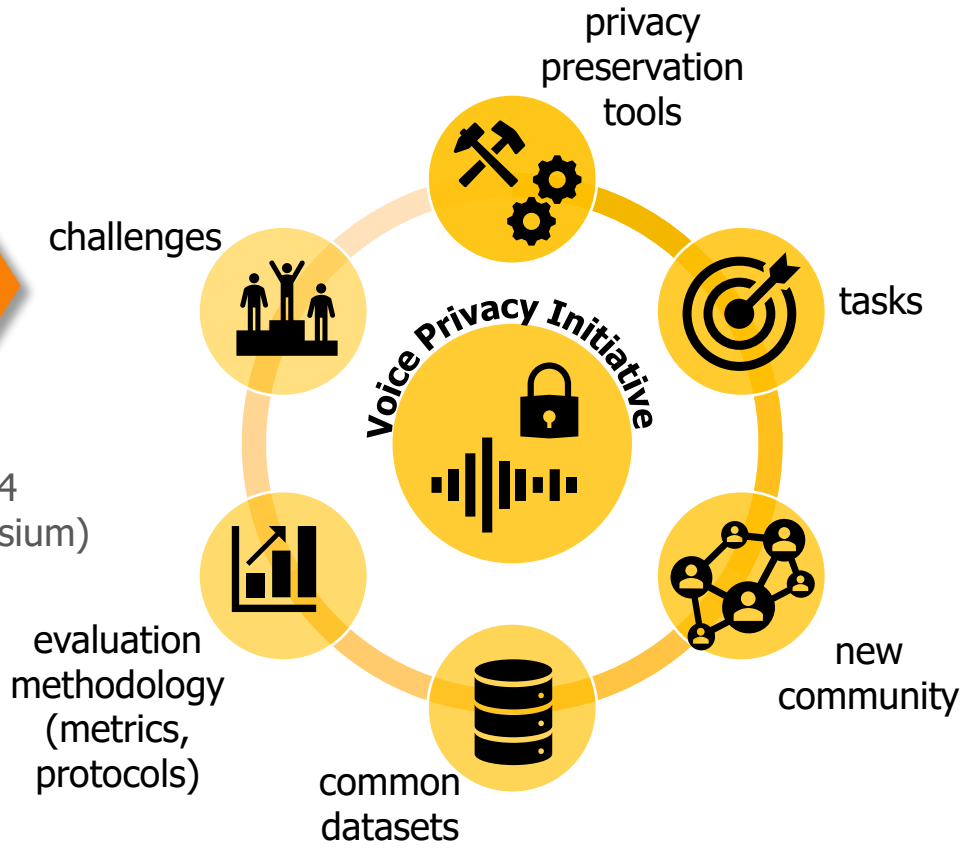
VoicePrivacy Challenges

Promote the development of privacy preservation tools for speech technology

2020 — **1st** VoicePrivacy Challenge
- SS at Interspeech 2020
- Satellite workshop at Speaker Odyssey 2020

2022 — **2nd** edition
- Workshop at Interspeech 2022 (2nd SPSC Symposium)

2024 — **3rd** edition
- Workshop at Interspeech 2024 (4th SPSC Symposium)



Privacy-enhancing technologies and challenge focus

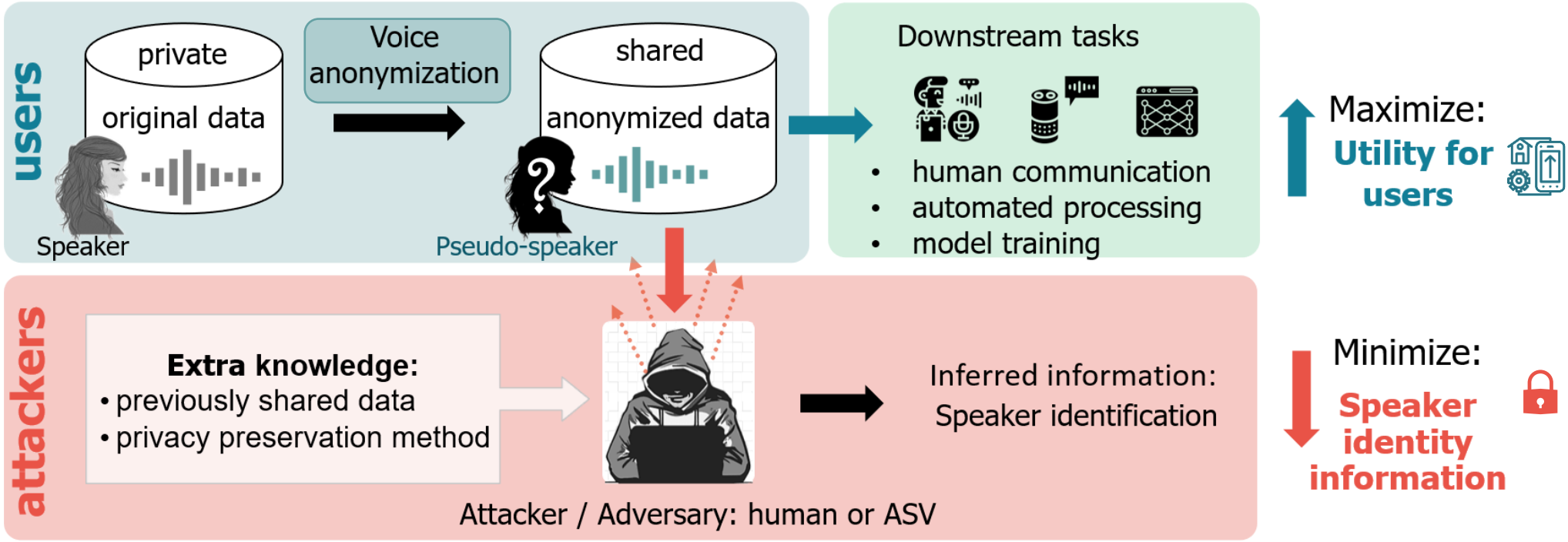
Anonymization

✕ remove personally identifiable information in the speech signal

✓ keep other characteristics unchanged
- linguistic content

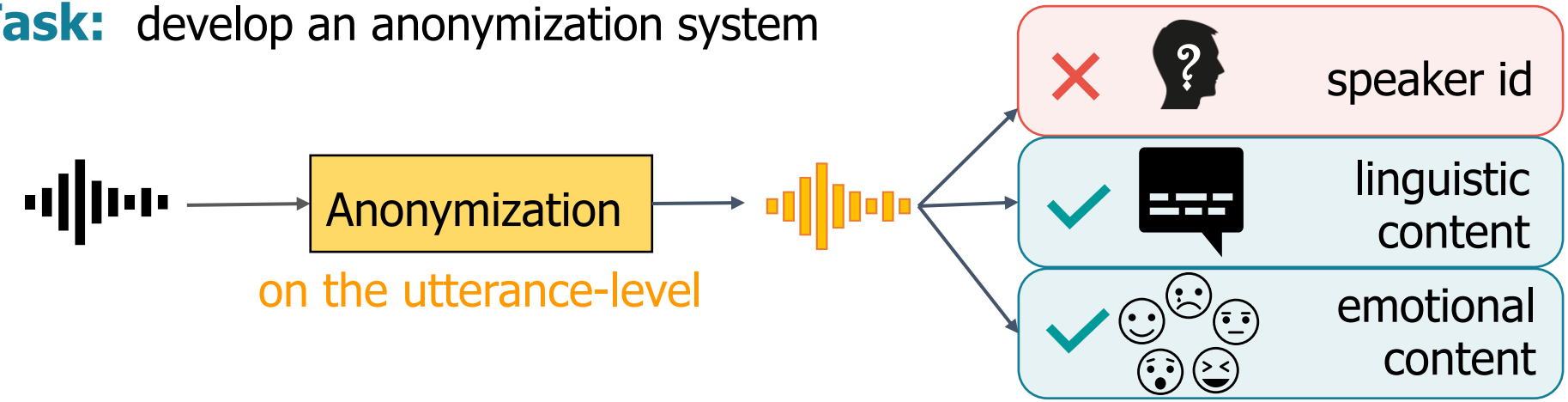
★new for VPC 2024: emotions

Anonymization task



Challenge task and requirements

Task: develop an anonymization system



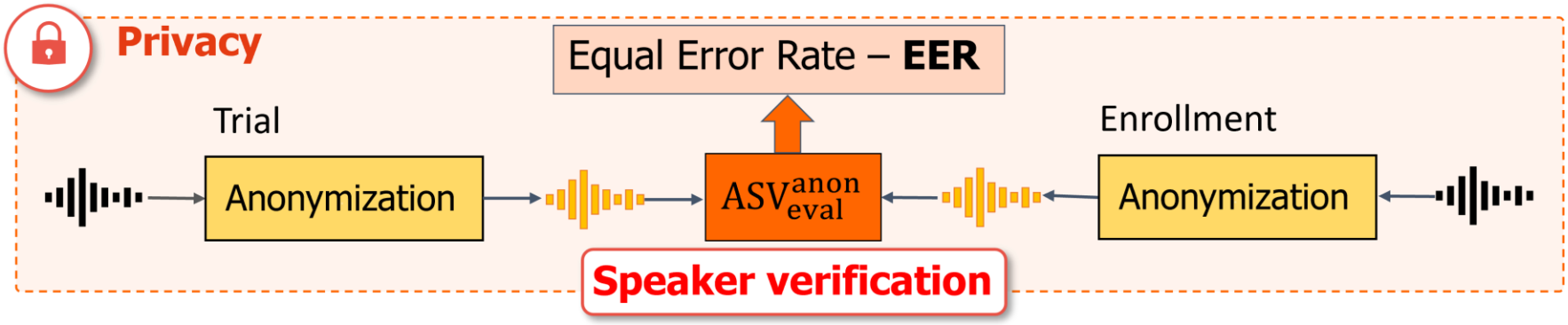
We provide:

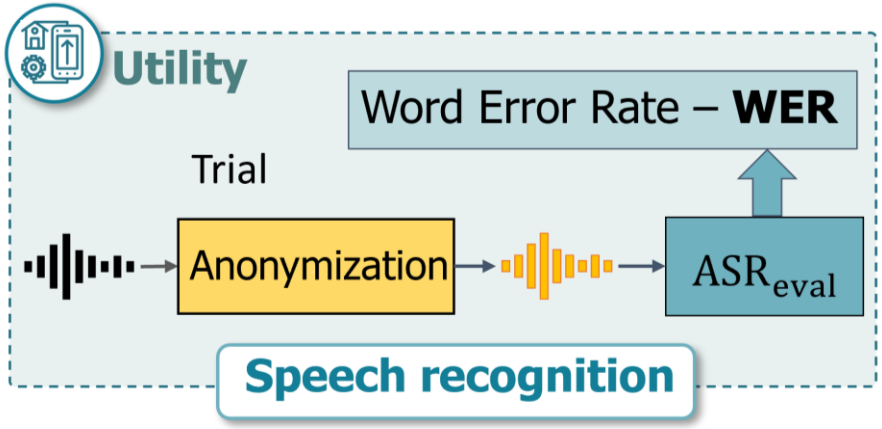
- training, development & evaluation data
- 6 baseline anonymization systems
- evaluation scripts and metrics

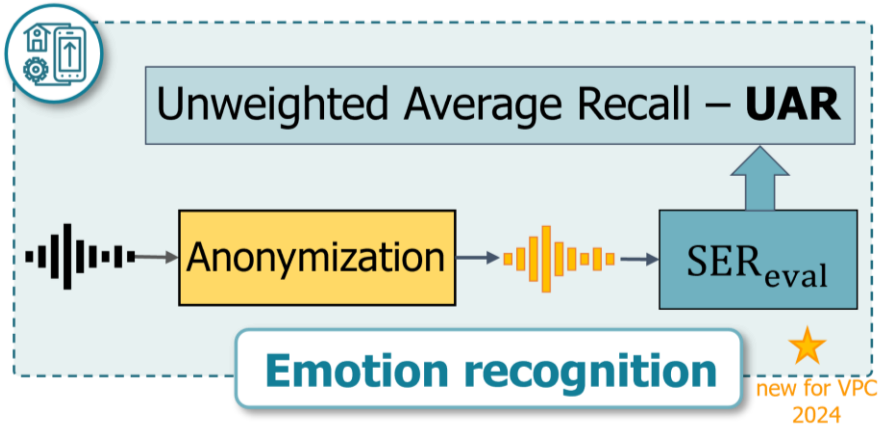
Participants:

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

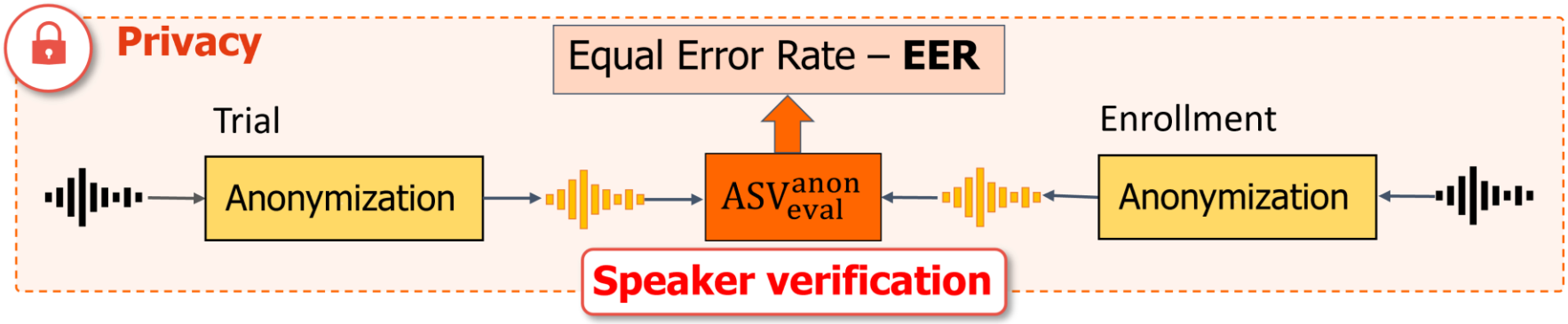
Evaluation


Privacy
Equal Error Rate – EER
Trial
Enrollment
Anonymization
ASV_{eval}^{anon}
Anonymization
Speaker verification


Utility
Word Error Rate – WER
Trial
Anonymization
ASR_{eval}
Speech recognition


Unweighted Average Recall – UAR
Anonymization
SER_{eval}
Emotion recognition
new for VPC 2024

Attacker model for privacy evaluation



Semi-informed attacker:

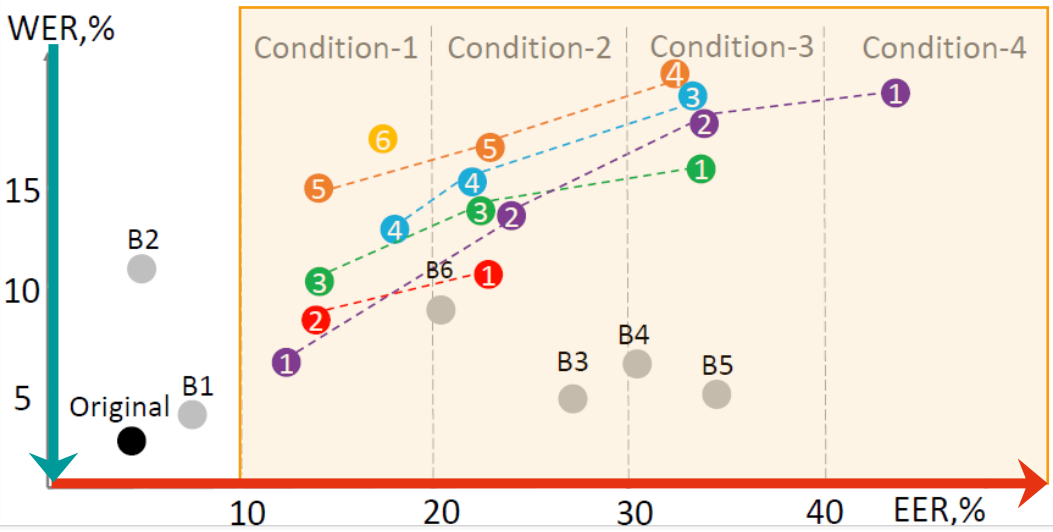
- has access to the anonymization system
- anonymizes enrollment & training data on the *utterance level*

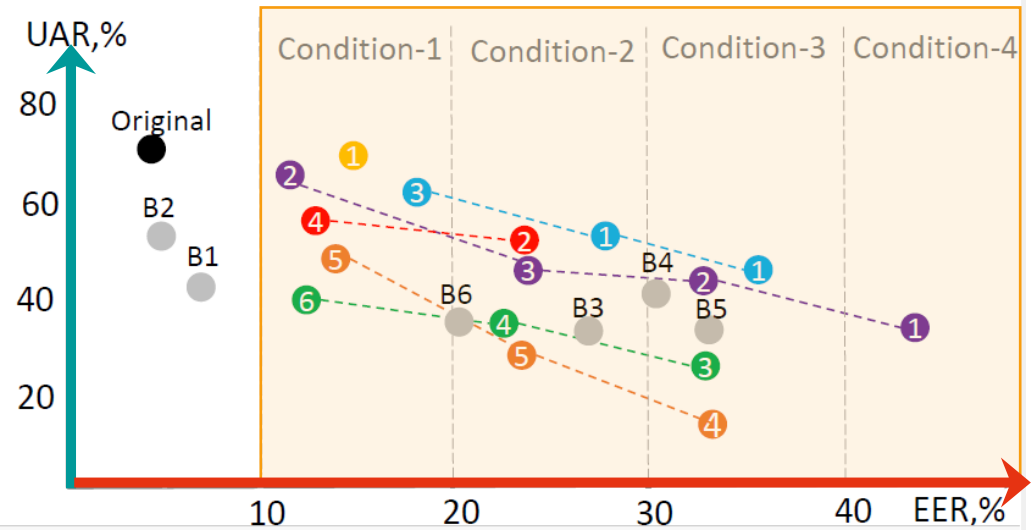
★ new for VPC 2024: *utterance level anonymization for enrollment (and trials) data*

Ranking policy

- **4** evaluation **conditions** (minimum target **privacy** requirements)
- To measure the **privacy-utility trade-off** of any solution at multiple operating points

1. EER ≥ 10%
2. EER ≥ 20%
3. EER ≥ 30%
4. EER ≥ 40%.

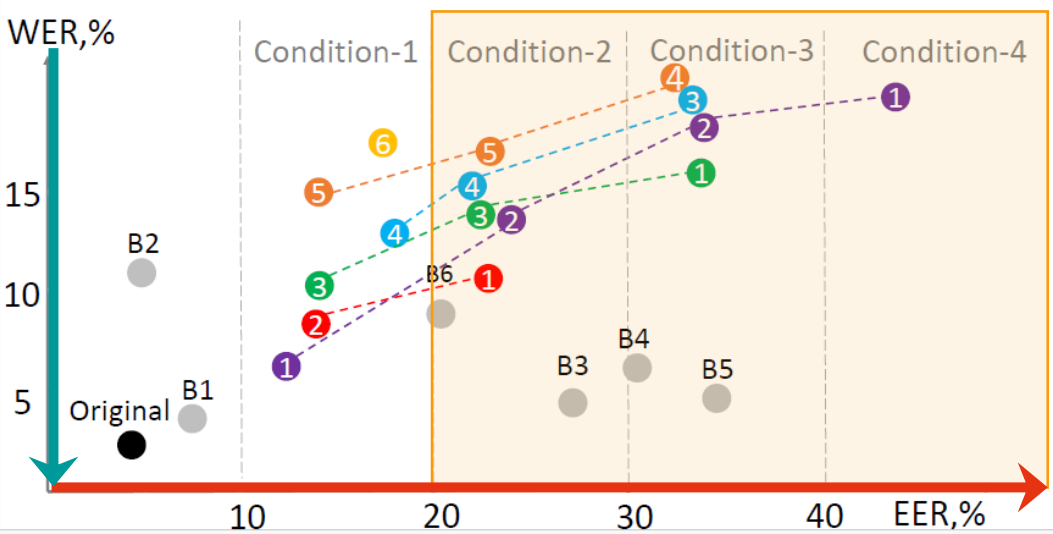


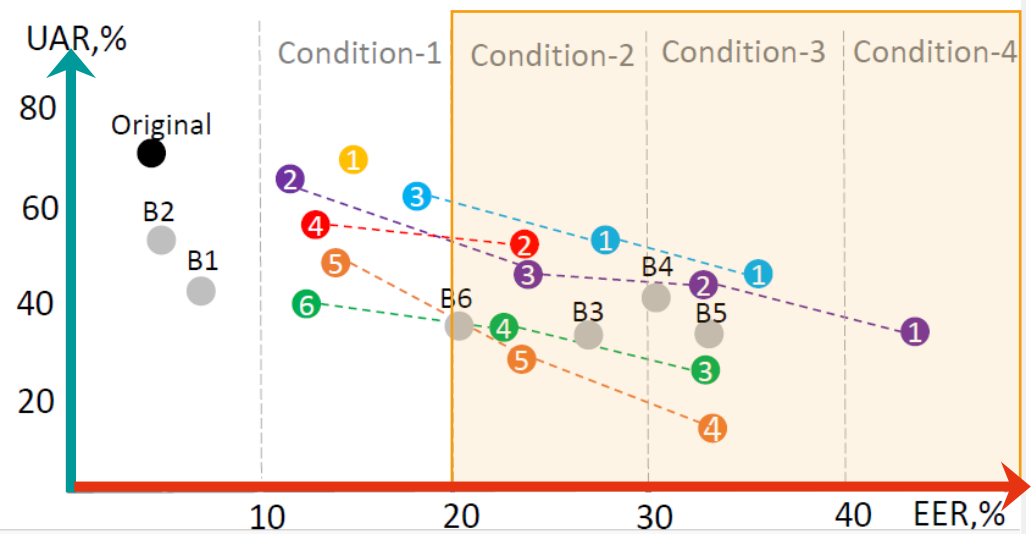


Ranking policy

- **4** evaluation **conditions** (minimum target **privacy** requirements)
- To measure the **privacy-utility trade-off** of any solution at multiple operating points

1. EER ≥ 10%
2. EER ≥ 20%
3. EER ≥ 30%
4. EER ≥ 40%.

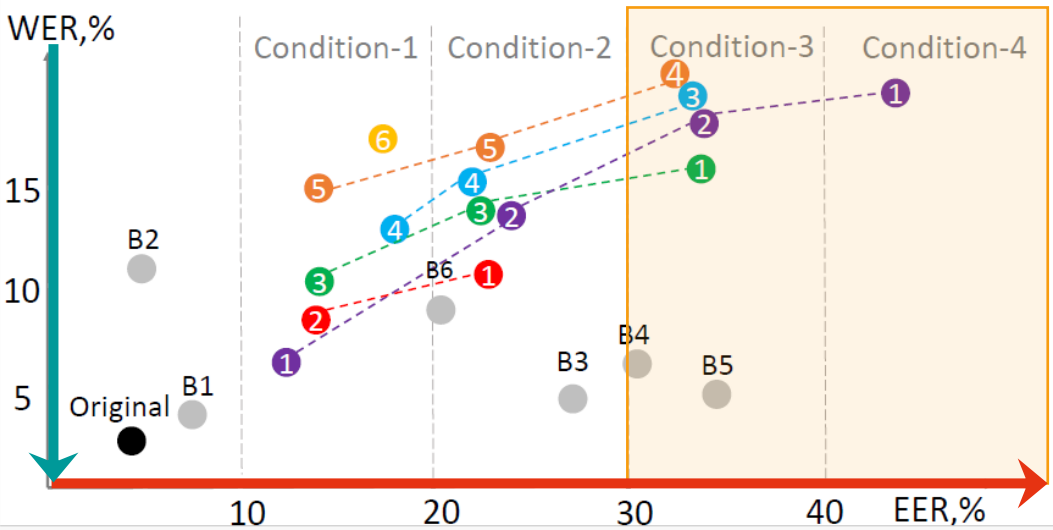


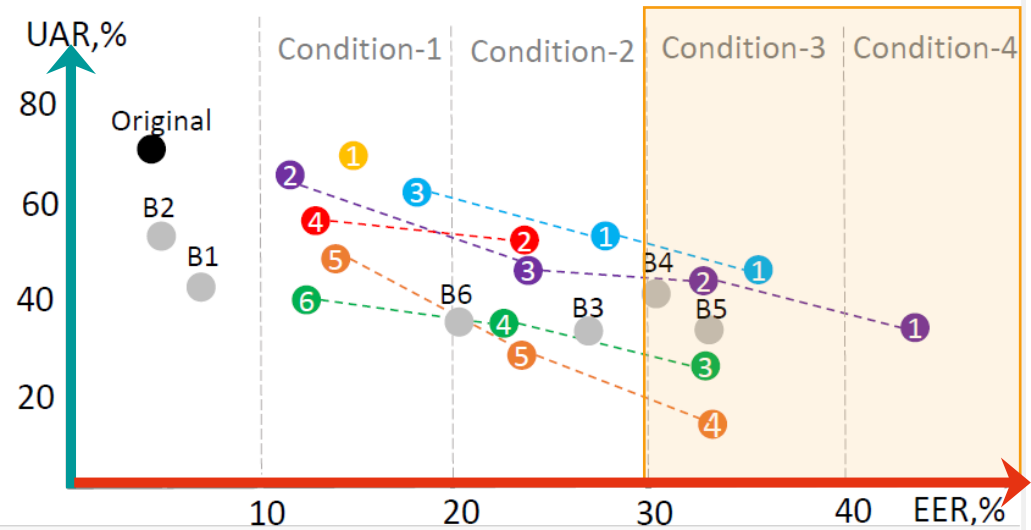


Ranking policy

- **4** evaluation **conditions** (minimum target **privacy** requirements)
- To measure the **privacy-utility trade-off** of any solution at multiple operating points

1. EER ≥ 10%
2. EER ≥ 20%
3. EER ≥ 30%
4. EER ≥ 40%.

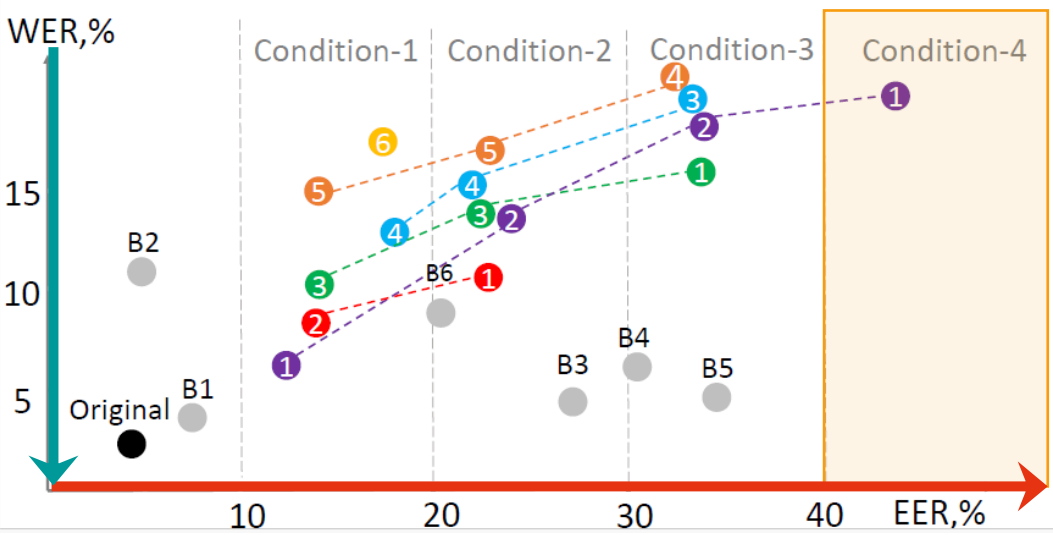


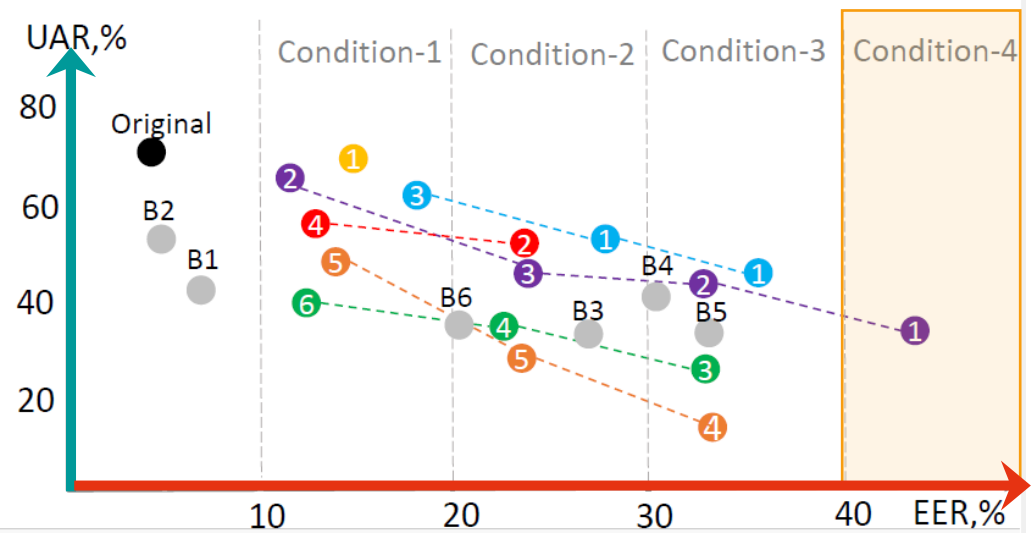


Ranking policy

- **4** evaluation **conditions** (minimum target **privacy** requirements)
- To measure the **privacy-utility trade-off** of any solution at multiple operating points

1. EER ≥ 10%
2. EER ≥ 20%
3. EER ≥ 30%
4. EER ≥ 40%.





Data and pretrained models

Training resources

Data

ESD ★★★★
CREMA-D ★★
RAVDESS ★★★★
VGAF from EmotiW challenge
MSP-Podcast ★★
CMU-MOSEI
EMO-DB
SAVEE

VCTK ★★★★★★
LJSpeech
LibriSpeech-train ★★★★★
Libri-light-train ★★
LibriTTS-train ★★★★

VoxCeleb-1,2 ★★★

MUSAN noise
RIR dataset

Models

WavLM ★★★★
Whisper ★★
HuBERT ★★★
XLS-R ★★
ContentVec
w2v-BERT
ECAPA2
ECAPA-TDNN
Encodec
Bark
NaturalSpeech3
NVIDIA Hifi-GAN Vocoder (en-US)
CRDNN on CommonVoice 14.0 English (No LM)
wav2vec 2.0 ★★★
wav2vec2-large-robust-12-ft-emotion-msp-dim

Software with pretrained models

Resemblyzer ★
VITS
PIPER pretrained on VITS
RVC-Project
DISSC

★– multiple requests

Development & Evaluation

LibriSpeech

✓ Speaker verification

✓ Speech recognition

IEMOCAP

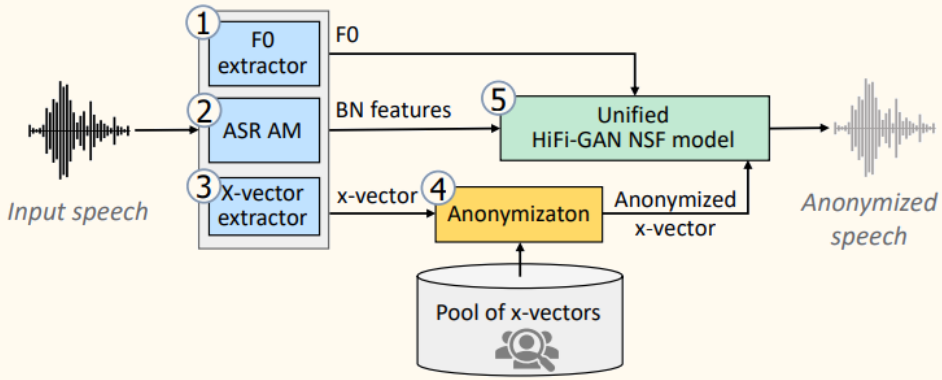
with 4 classes:

neutral, sadness, anger & happiness

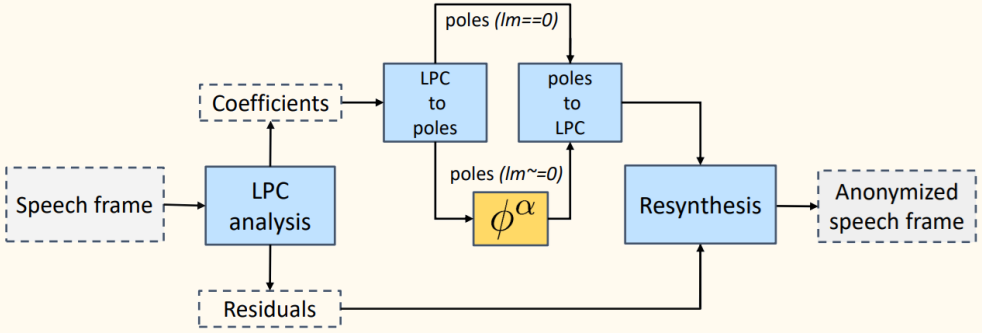
✓ Emotion recognition

Baselines from VPC-2020/2022

B1 x-vectors and a neural source-filter model

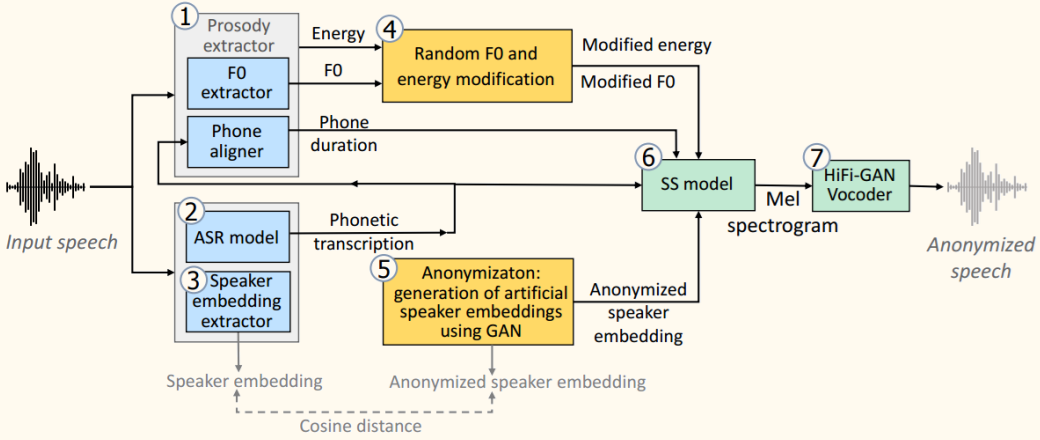


B2 McAdams coefficient

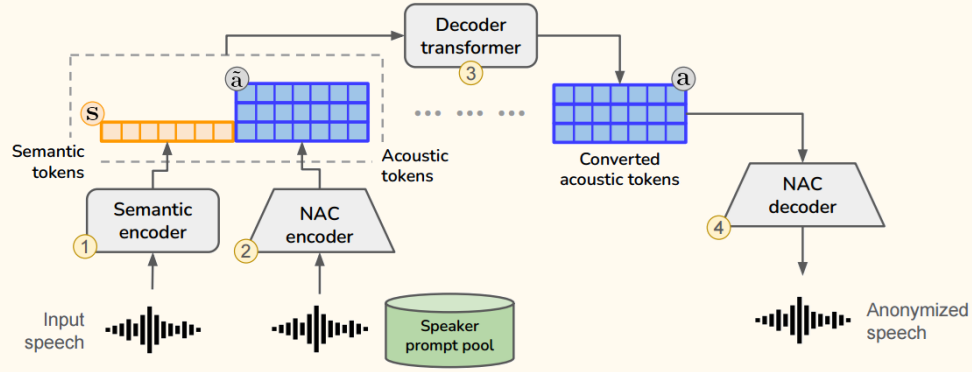


New baselines

B3 phonetic transcriptions and a generative adversarial network

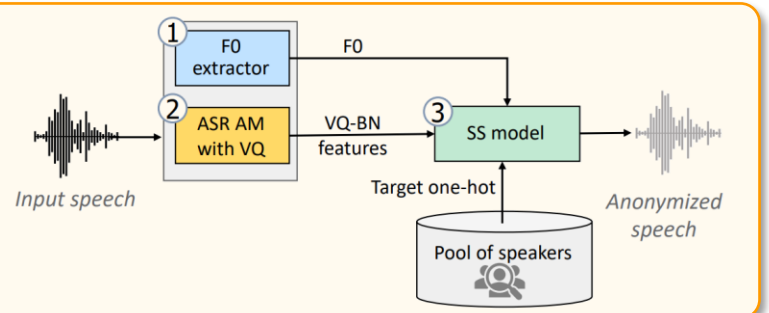


B4 neural audio codec language modeling



B5 ASR-BN with VQ and a pretrained *wav2vec2* model

B6 Similar to B5, but without using a pretrained *wav2vec2* model.



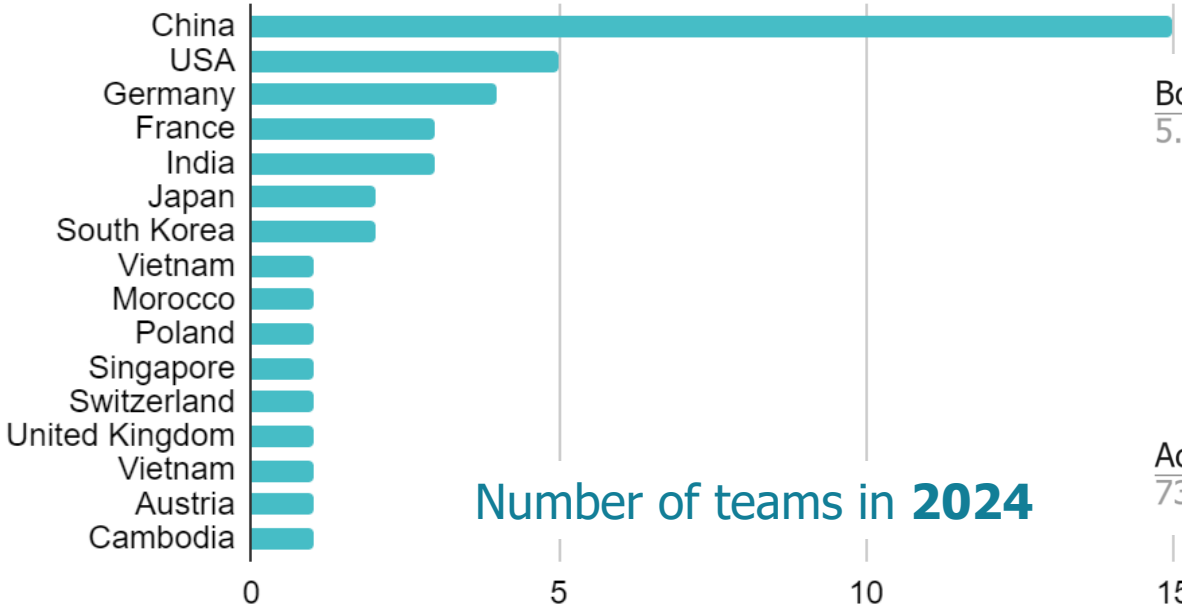
What is new in VPC-2024?

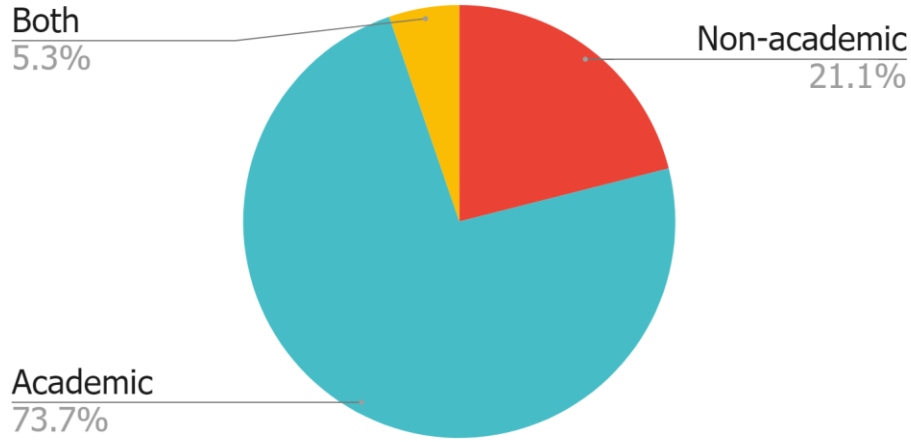
- Utterance-level anonymization (remove voice distinctiveness and intonation)
- New utility metric: UAR for emotion recognition
- Extended list of data & pretrained models, based on the participants' requests (for training)
- Only objective evaluation
- Models for utility evaluation (ASR and SER) are trained on original data
- 4 new baselines
- Reduced evaluation complexity and speed

Participants

- Registered teams: **40** (**107** participants) from **16** countries
- Teams submitted results: **13**
- Submitted anonymization systems: **36**

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





Teams and systems

Team name	Team ID	Affiliation	# Systems
Anemone	T7	Institute of Acoustics, Chinese Academy of Sciences, China ● University of the Chinese Academy of Sciences, China	2
JHU CLSP	T8	Johns Hopkins University, United States	5
LongYuan	T9	Auditory Intelligence Computing Group (AIC), Nanjing University of Posts and Telecommunications, China ● Nanjing Longyuan InformationTechnology Co.Ltd, China	1
NPU-NTU	T10	Audio, Speech and Language Processing Group (ASLP@NPU) ● School of Computer Science, Northwestern Polytechnical University ● Nanyang Technological University ● The Hong Kong Polytechnic University	2
NTU-NPU	T12	Nanyang Technological University, China ● Institute for Infocomm Research, A*STAR, Singapore ● Audio, Speech and Language Processing Group (ASLP@NPU), China ● The Hong Kong Polytechnic University	6
ADRES	T14	Univ. Grenoble Alpes, CNRS, Grenoble LIG & LJK, France ● Institute of Digital Research and Innovation, CADT, Phnom Penh, Cambodia	2
NKU HLT Lab	T17	HLT Lab, College of Computer Science, Nankai University, China	1
Q	T18	Qifu Technology, China ● Fudan University, Shanghai, China	2
DFKI_SLT	T19	Speech and Language Technology, German Research Center for Artificial Intelligence (DFKI), Germany ● Quality and Usability Lab, Technische Universitat Berlin, Germany	3
USTC-PolyU	T25	NERC-SLIP, University of Science and Technology of China, China ● Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong	2
V-Beam	T30	Sogang University, Seoul, Korea ● Ewha Womans University, Seoul, Korea	2
KIT-ISL	T33	Karlsruhe Institute of Technology (KIT), Germany ● Carnegie Mellon University (CMU), USA	2
Orange_shiva	T38	Orange, France	6

See teams ids also at: https://www.spsc2024.mobileds.de/#program

Submitted systems

1 Neural voice-conversion methods

T8-2, T19-1,2 kNN-VC

with attribute disentanglement:

content, speaker, F0, emotion, …

T7-1,2 fusion of VC models on the model parameter level
T9 GMM-blender
T14-1,2 content (VQ-VAE), prosody, x-vect. anonymization
T18-1,2 based on FreeVC + emotion encoder
T19-3 VQ MI based disentanglement
T25-1,2 content (VQ-BN) & non-content (GST) disentanglement + transfer non-content information from utterances of other speakers with the same predicted emotions
T38-1 DISSC, **2** +prosody preservation, **3** +MSP-podcast, **4** +randomness to prosody, **5** +prosody prediction using MSP-podcast, **6** +emotion embeddings

2 Using neural codecs

T10 disentangled neural codec
T12-1 NaturalSpeech 3 FACodec + white Gaussian noise to speaker embedding & cross-gender conversion
T17 NaturalSpeech 3 FACodec + PANO

3 Cascade ASR+TTS-based systems

T30-2 Whisper ASR + TTS + integration class emotion prototype & F0
T8-1 Whisper + VITS-TTS
T30-1 phonetic transcriptions + modified F0 & energy,…

4 Hybrid / fusion methods

T8-3,4,5 Admixture (among 2 methods, randomly select one for each utterance)

Approaches to emotion preservation

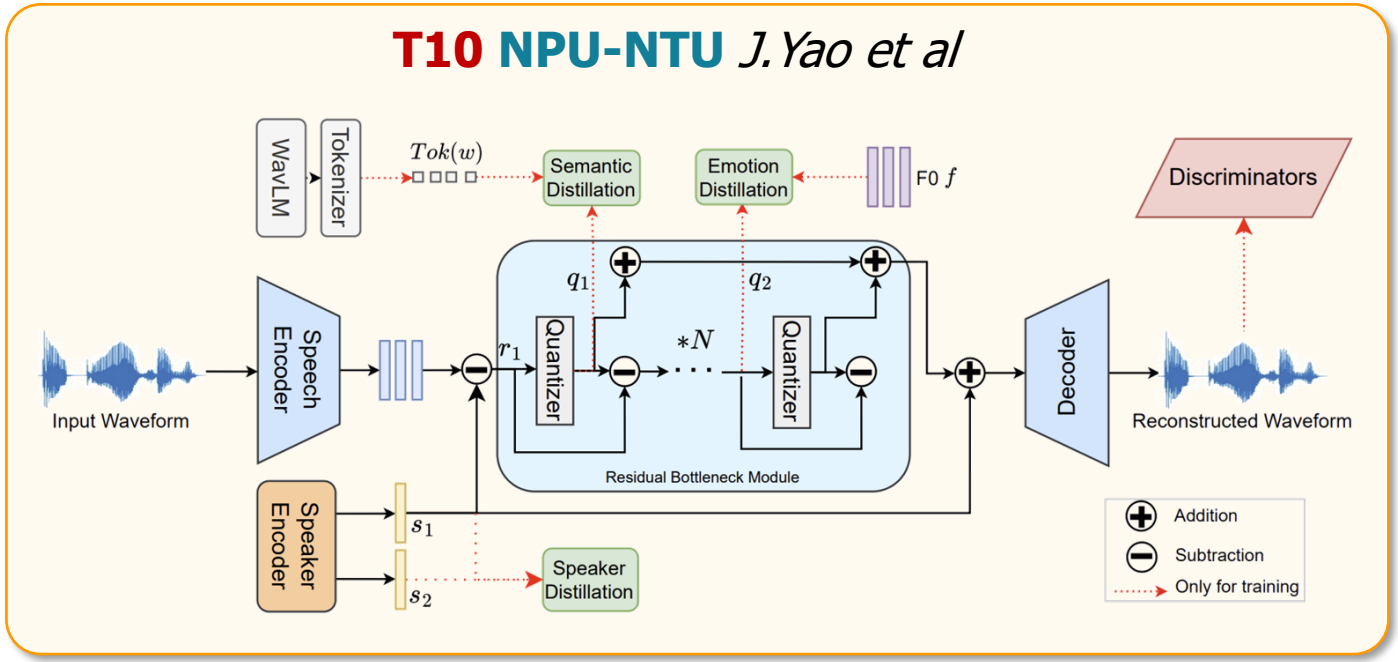
Emotion encoders

- based on pretrained models: wav2vec2 **T7-1**, **T18-1**
- global style token (GST) **T7-2**, **T25-1**

Emotion distillation

- **T10**

disentangled neural codec & sequential disentanglement for linguistic content, speaker id & emotion



Anonymization strategies (selected)

T8-1 ASR+TTS – nearly perfect anonymization (EER=48%) (then adjust utility-privacy tradeoff for emotions using admixture with a kNN-VC system)

T30-2 ASR+TTS

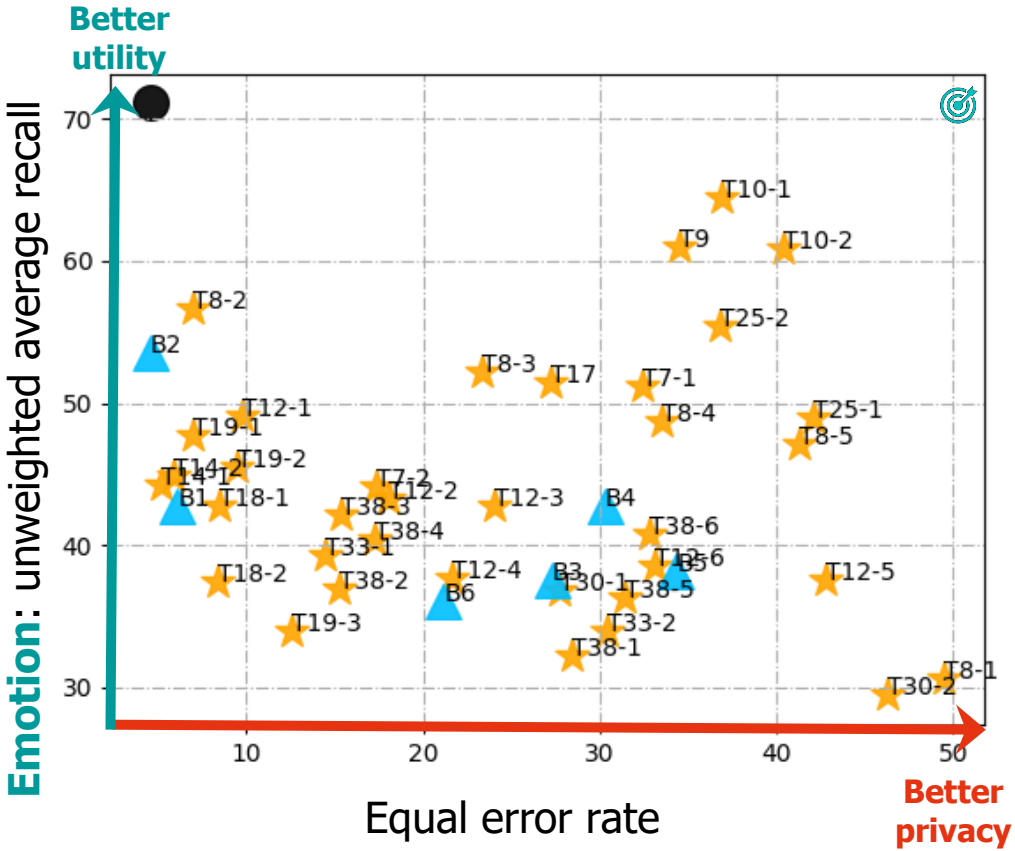
T10 combine averaged speaker id (from a speaker pool, as in VPC **B1**) and a randomly generated speaker id from a Gaussian distribution: $s_{\text{anon}} = \alpha\bar{s} + (1-\alpha)\hat{s}$

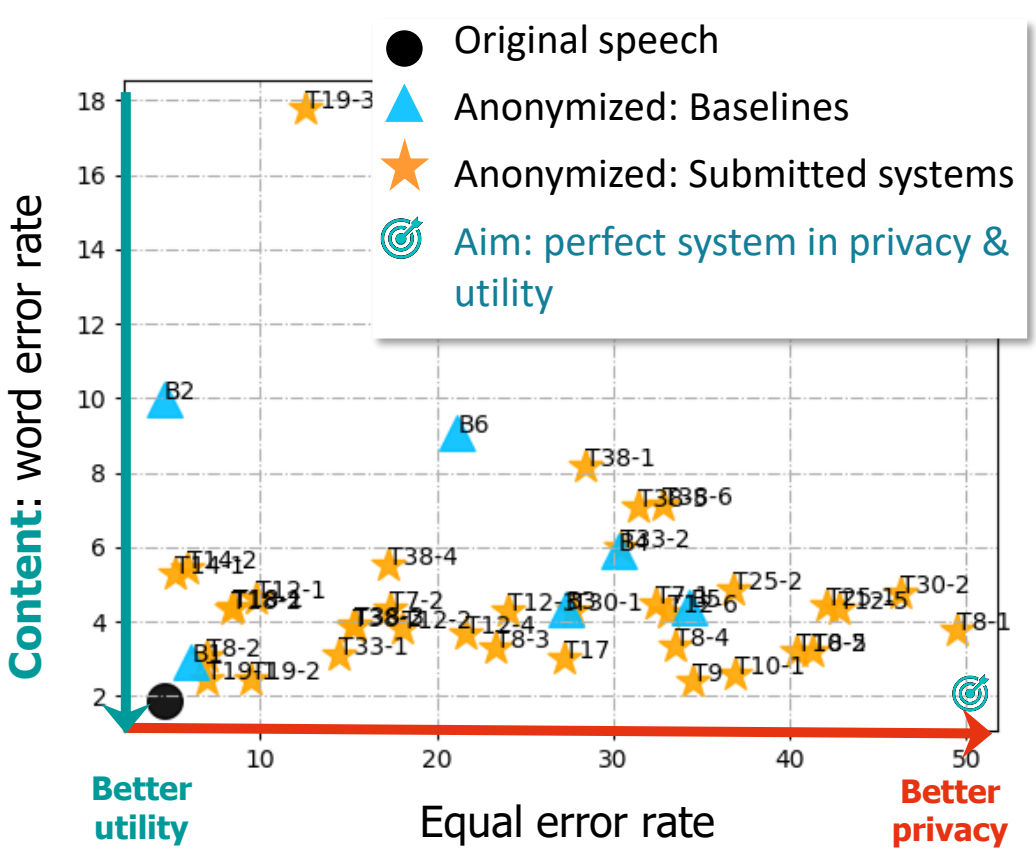
T12-5, 6 based on B5 /BN-VQ for content, 1-hot target speaker vectors from the pool/ T12-5: mean reversion F0, T12-6 + additive white Gaussian noise (AWGN)

T25-1,2 content & anonymization parts are derived from B5, speaker pool with emotion labels (ESD / ESD+LibriSpeech), random selection of utterance of a target speaker (1-hot) (consistent in emotion with the source speaker utterance)

T9 GMM Blender

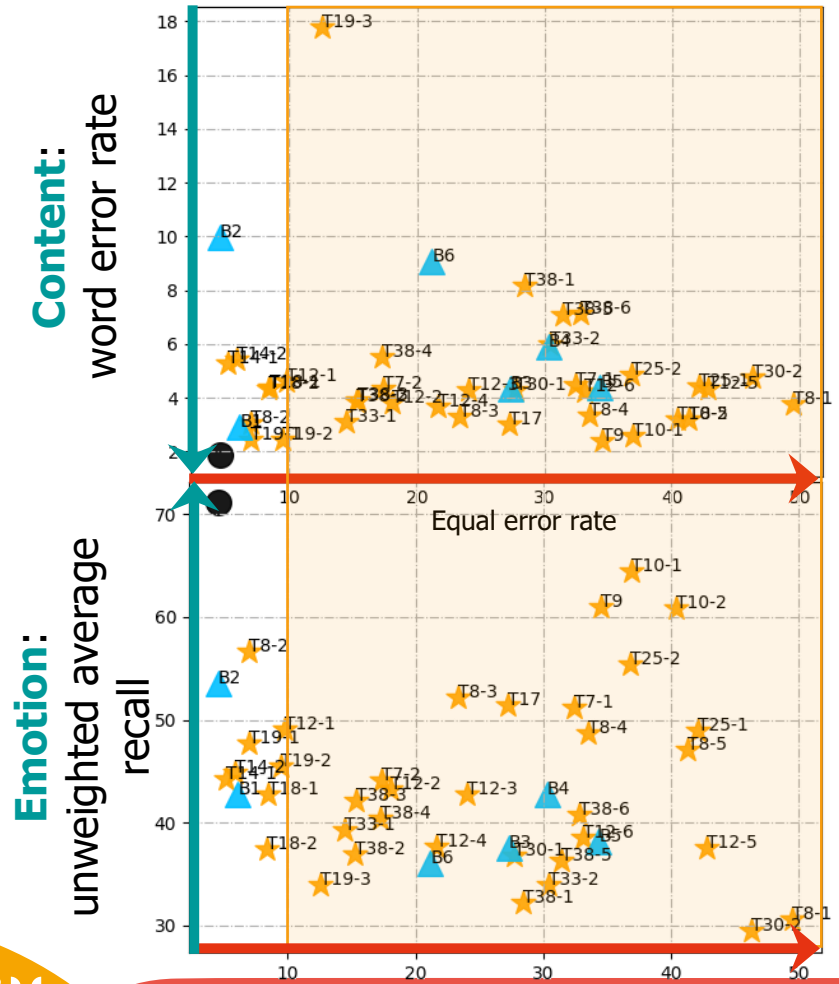
Results: privacy vs utility



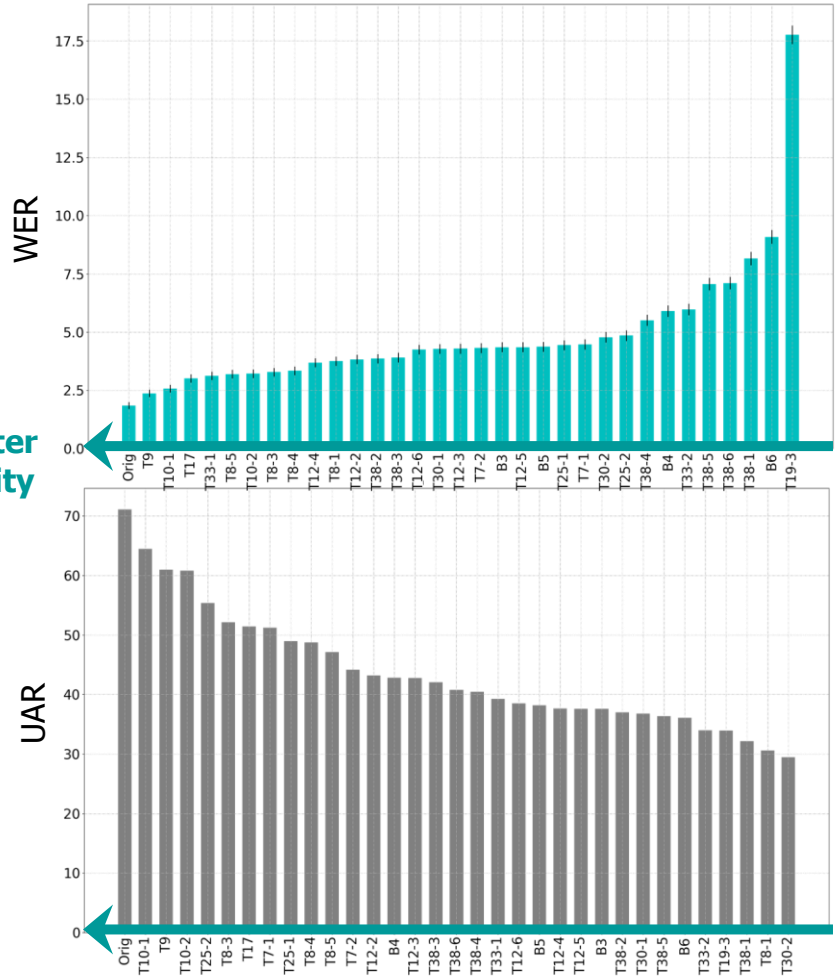


Results on test data, ASV models were retrained by organizers on the anonymized participants' data

Results: privacy vs utility



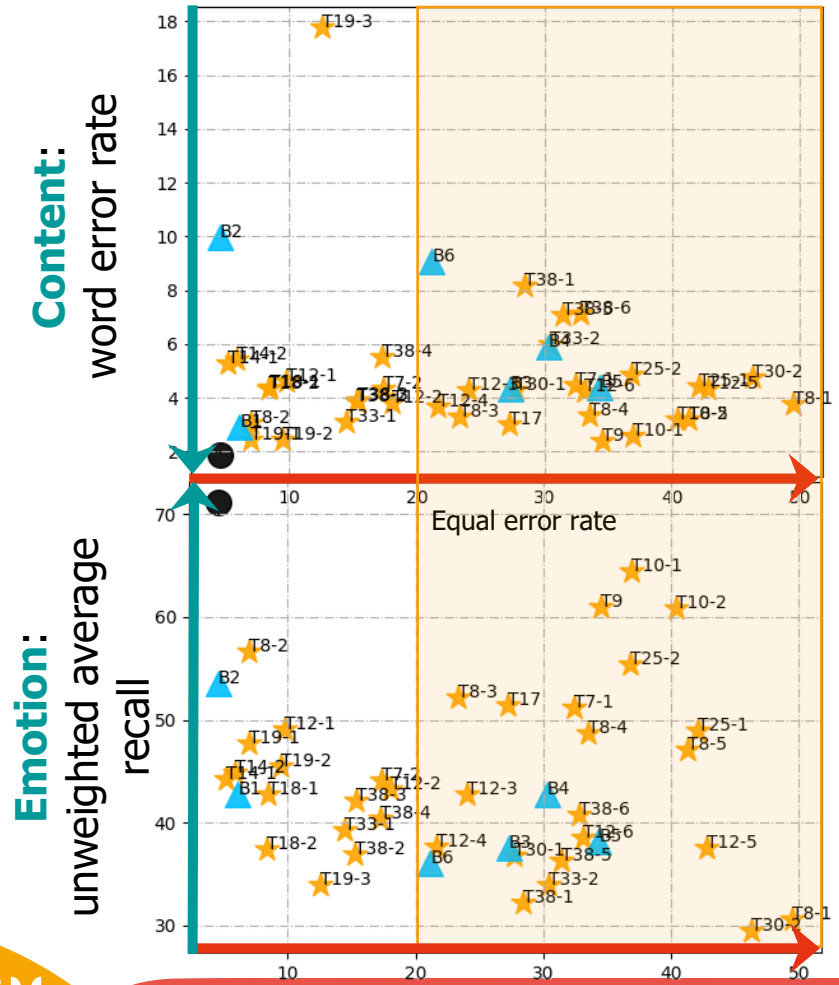
1. EER ≥ 10%
2. EER ≥ 20%
3. EER ≥ 30%
4. EER ≥ 40%



*Possible incompliance of systems with the following VPC2024 requirements/setup:

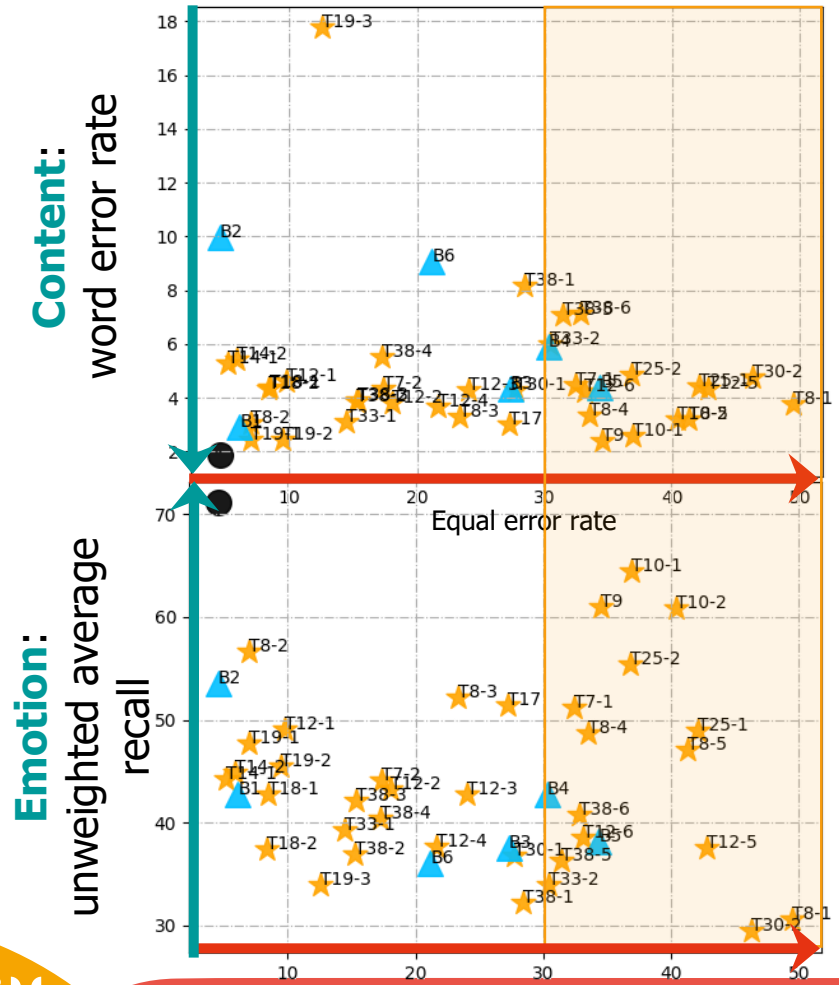
- T17: "the trial utterances shall be processed independently of each other"
- T25: "IEMOCAP is not included in the VPC training data"

Results: privacy vs utility



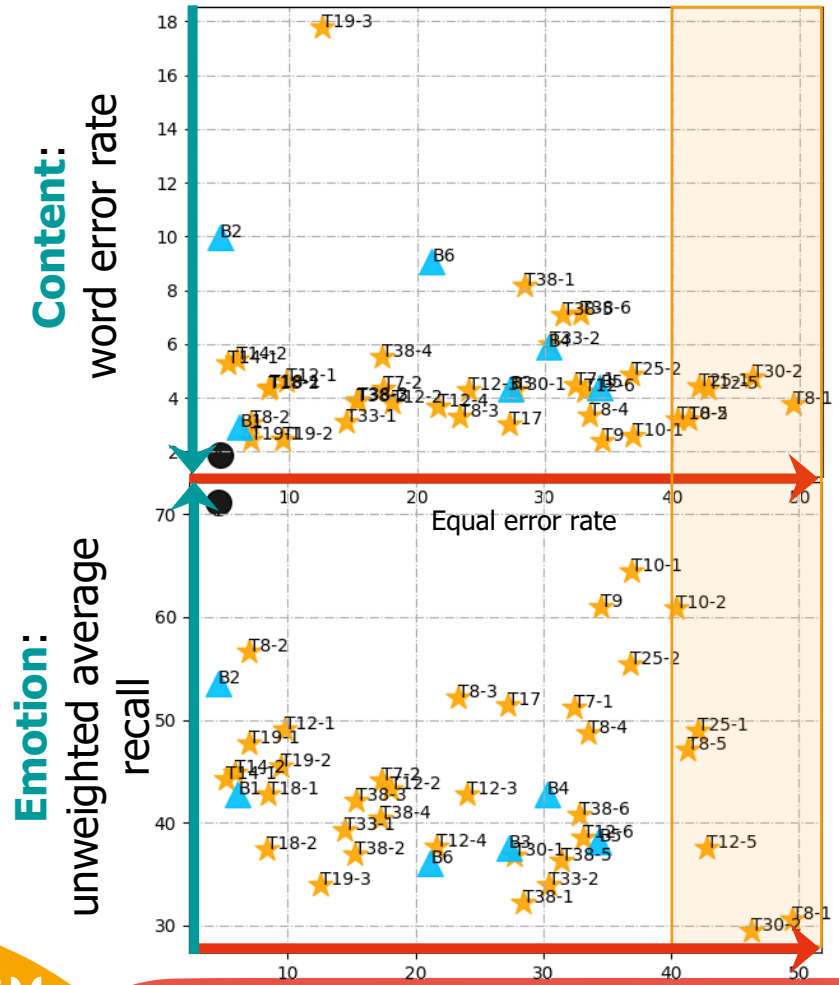
1. EER ≥ 10%
2. EER ≥ 20%
3. EER ≥ 30%
4. EER ≥ 40%

Results: privacy vs utility

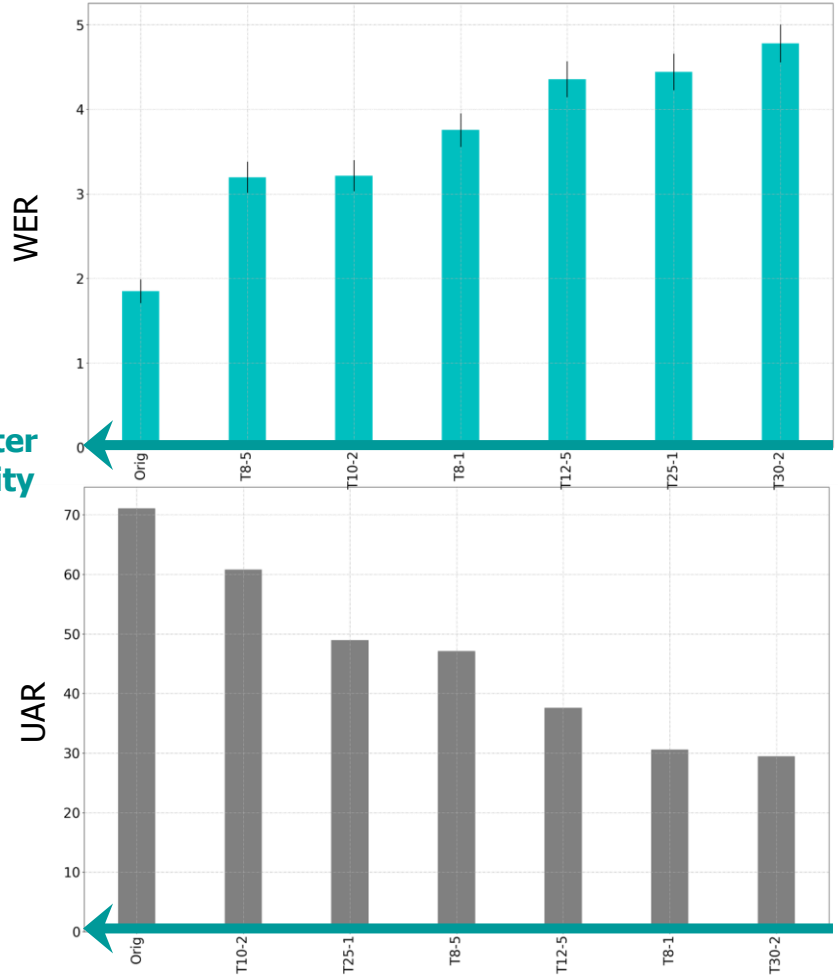


1. EER ≥ 10%
2. EER ≥ 20%
3. EER ≥ 30%
4. EER ≥ 40%

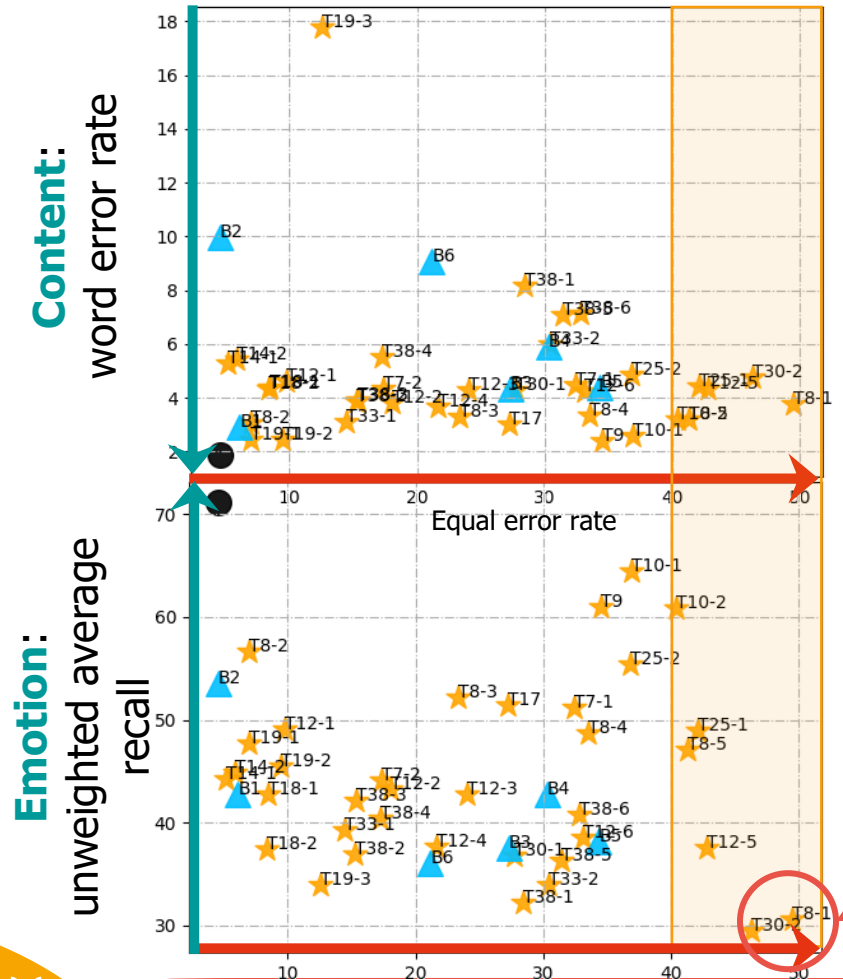
Results: privacy vs utility



1. EER ≥ 10%
2. EER ≥ 20%
3. EER ≥ 30%
4. EER ≥ 40%

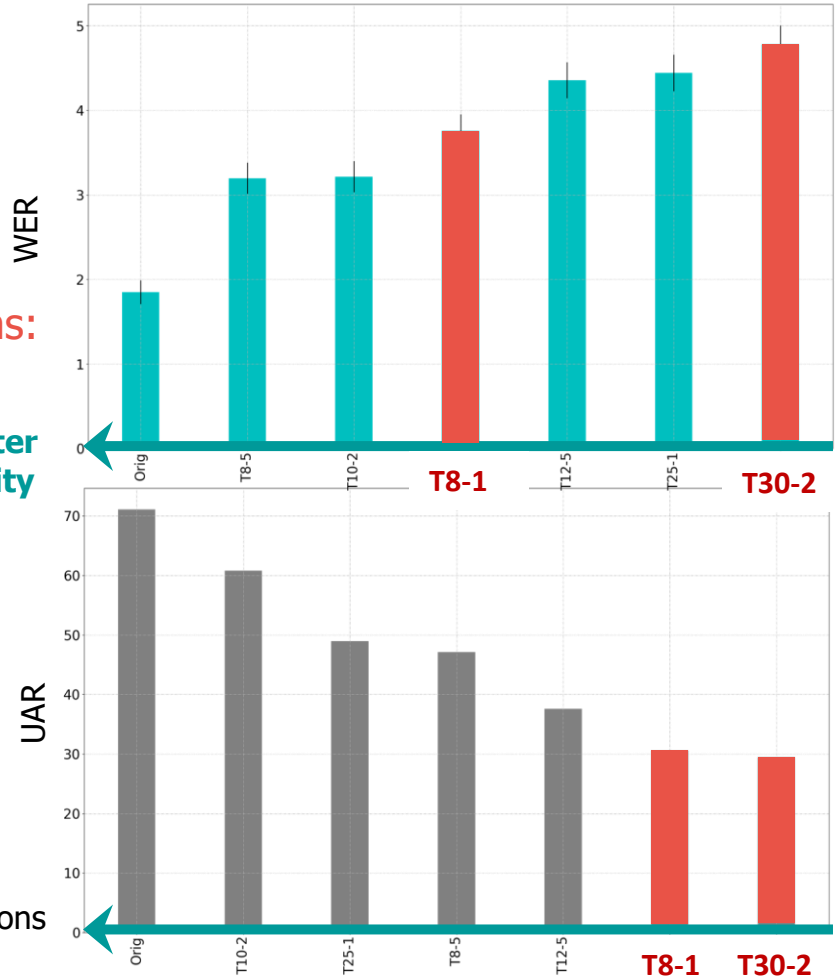


Results: privacy vs utility



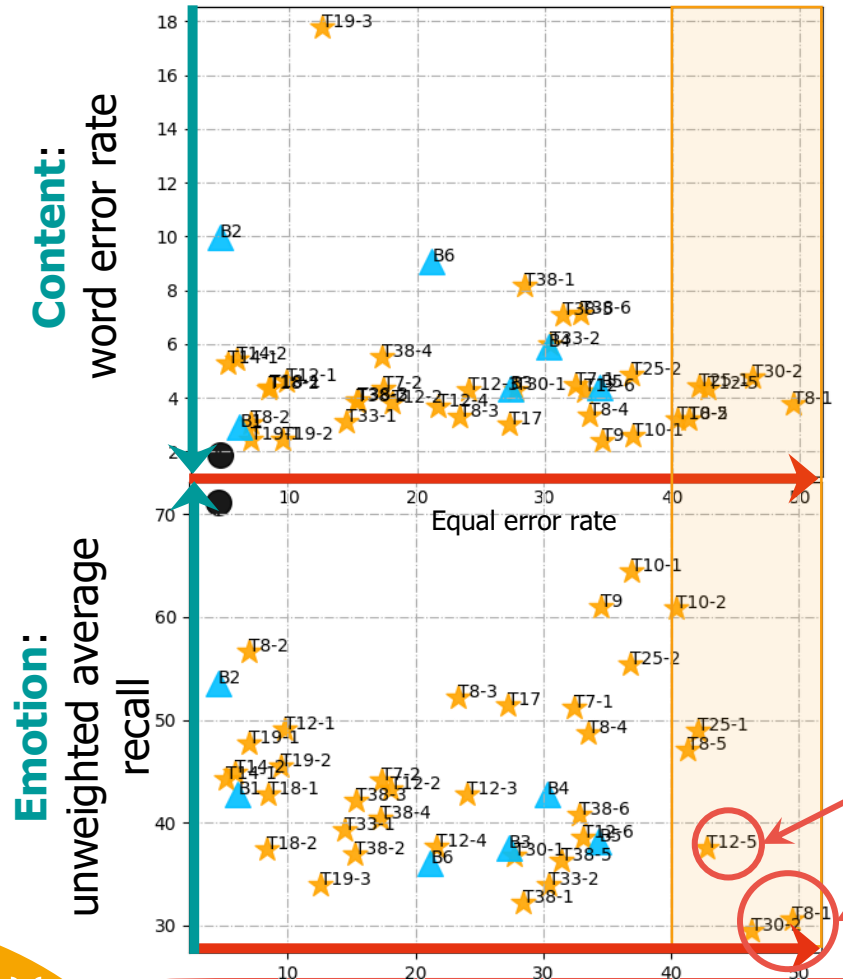
1. EER ≥ 10%
2. EER ≥ 20%
3. EER ≥ 30%
4. EER ≥ 40%

Best privacy systems:



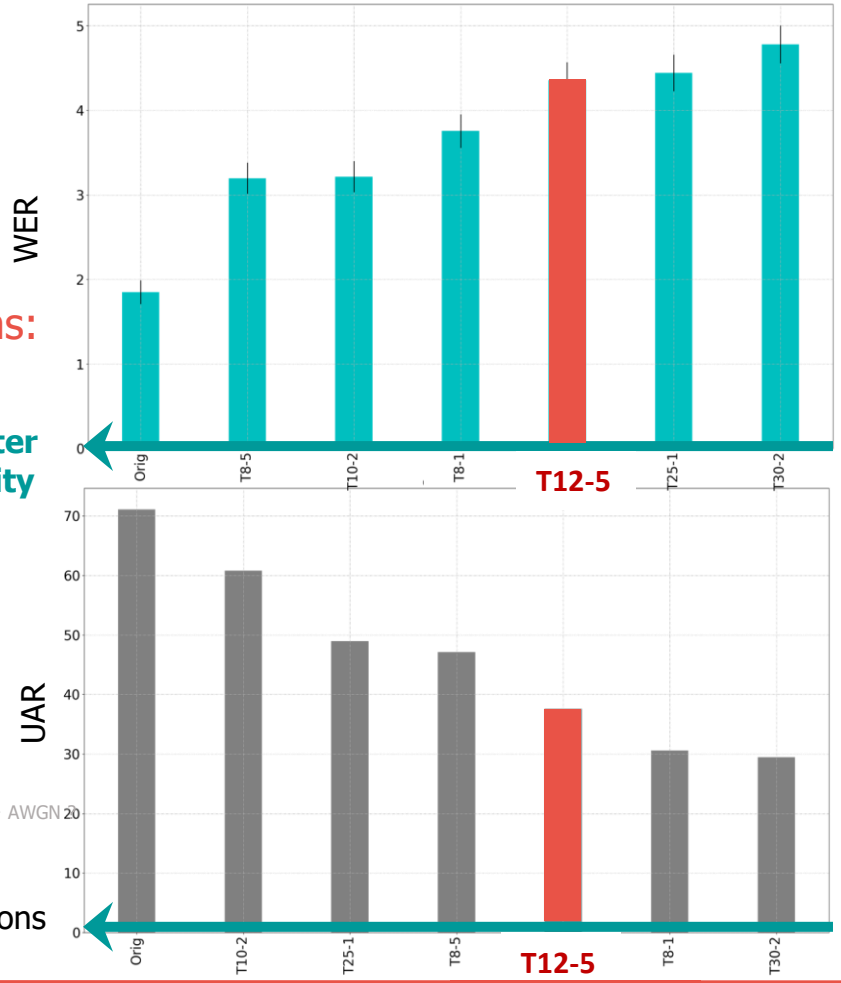
ASR + TTS
do not preserve emotions

Results: privacy vs utility



1. EER ≥ 10%
2. EER ≥ 20%
3. EER ≥ 30%
4. EER ≥ 40%

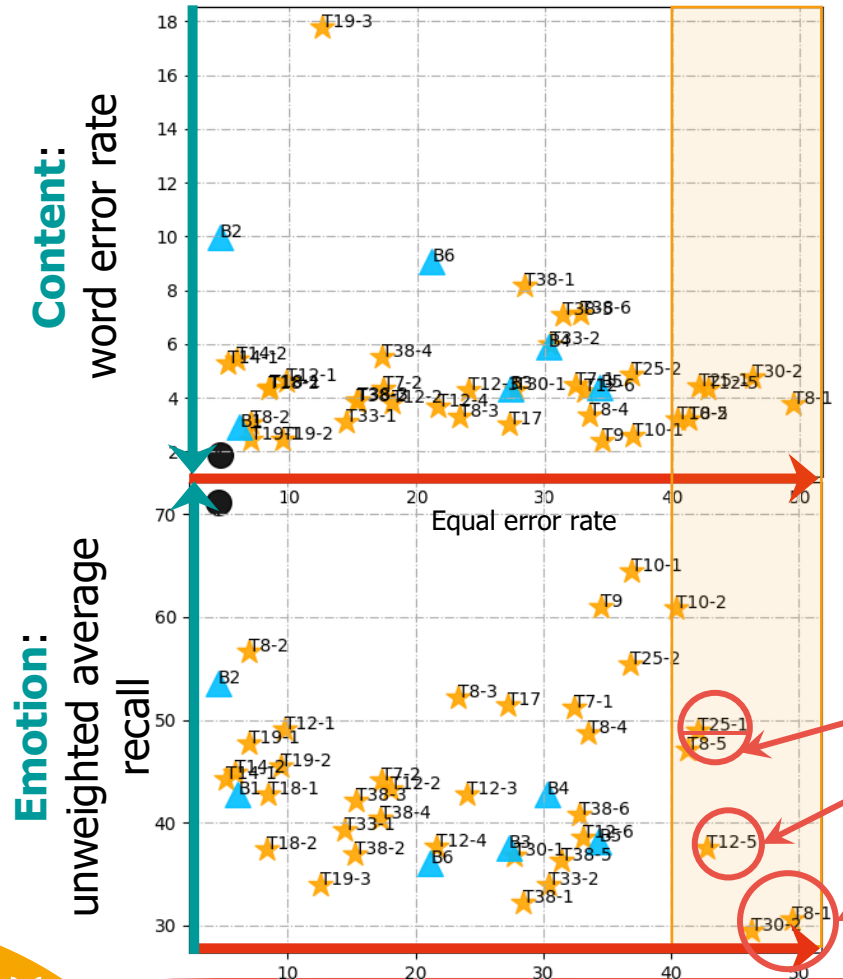
Best privacy systems:



Based on B5 + mean reversion F0 + AWGN

ASR + TTS do not preserve emotions

Results: privacy vs utility



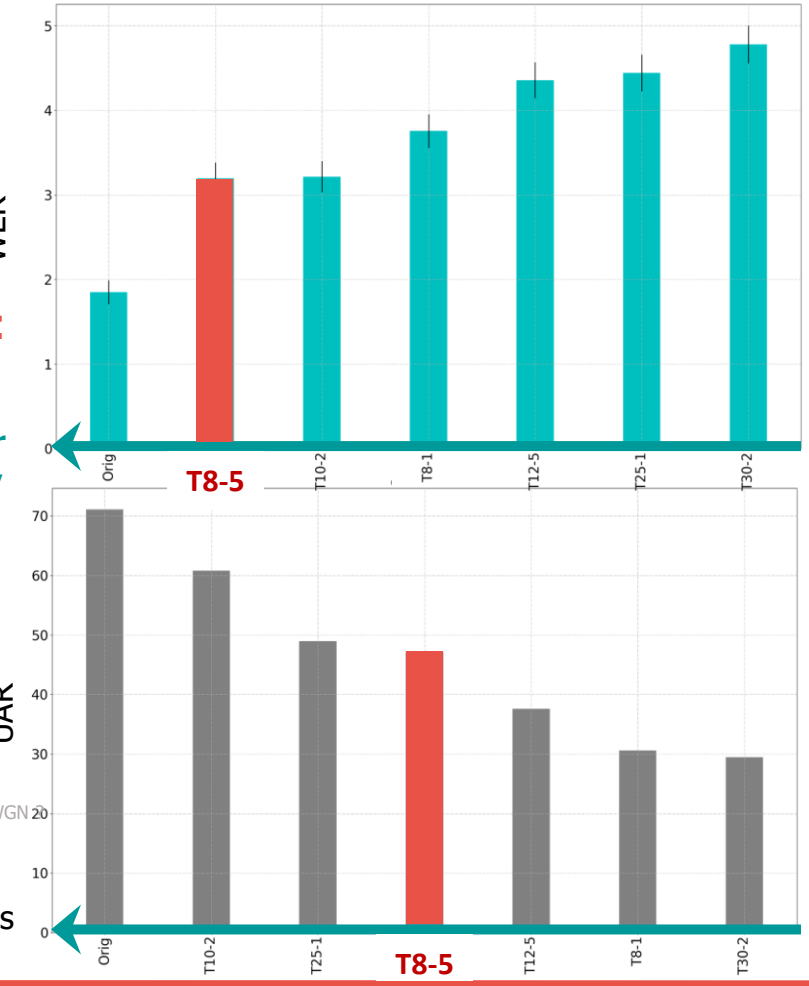
1. EER ≥ 10%
2. EER ≥ 20%
3. EER ≥ 30%
4. EER ≥ 40%

Best privacy systems:

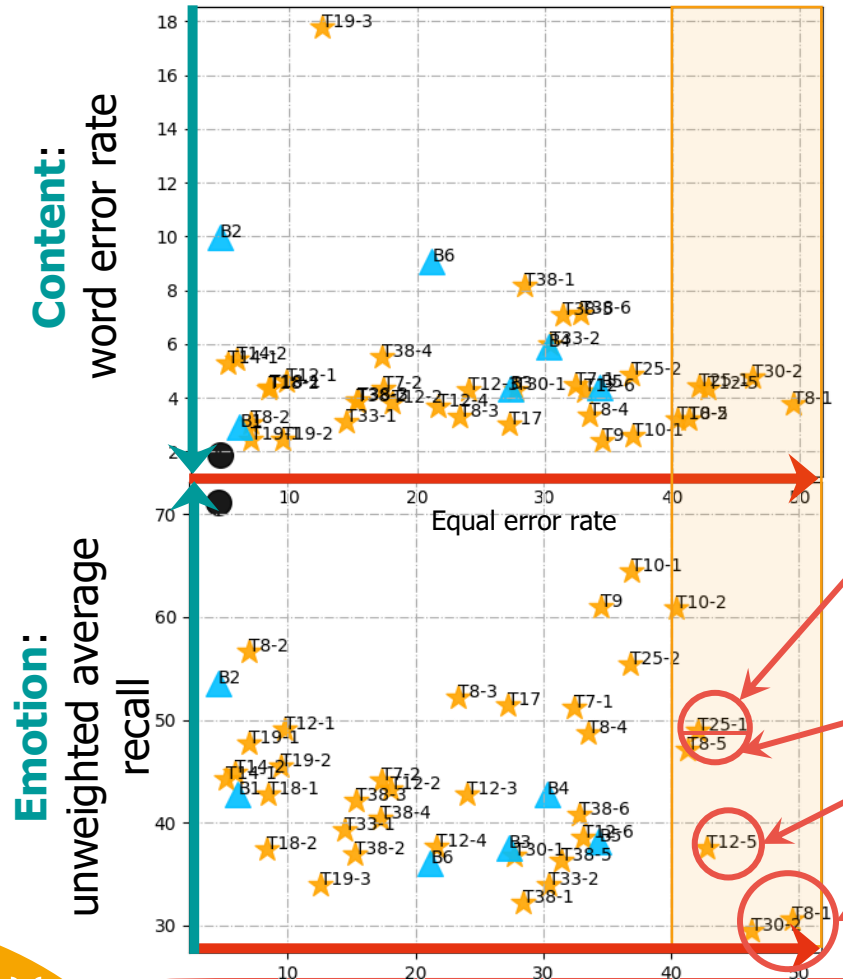
admixture

Based on B5 + mean reversion F0 + AWGN

ASR + TTS do not preserve emotions

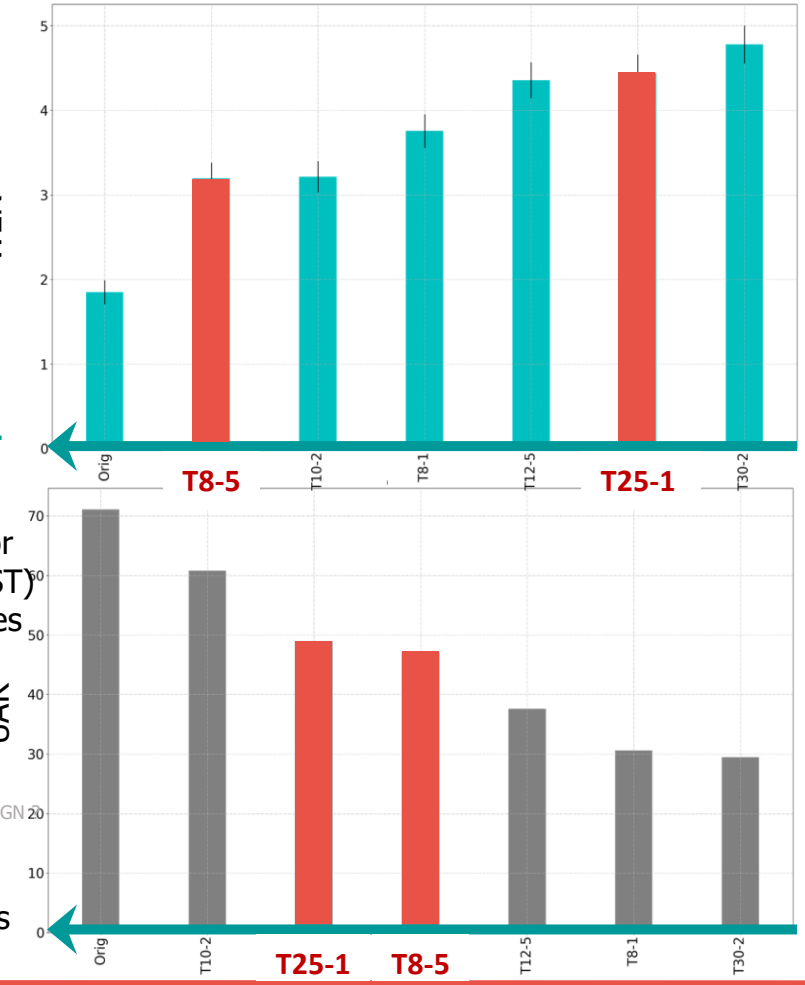


Results: privacy vs utility



1. EER ≥ 10%
2. EER ≥ 20%
3. EER ≥ 30%
4. EER ≥ 40%

Best privacy systems:



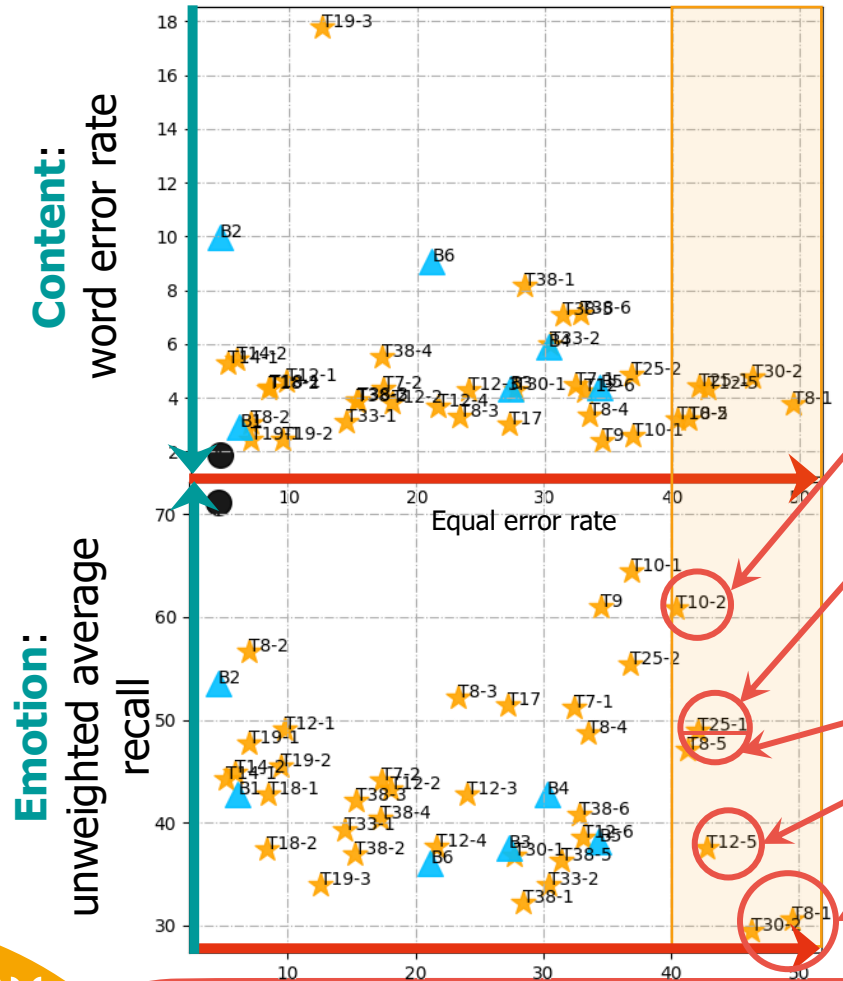
VC: disent. content-& non-content, VQ-BN for content & emotion (GST) transfer from utterances of target speakers

admixture

Based on B5 + mean reversion F0 + AWGN

ASR + TTS do not preserve emotions

Results: privacy vs utility

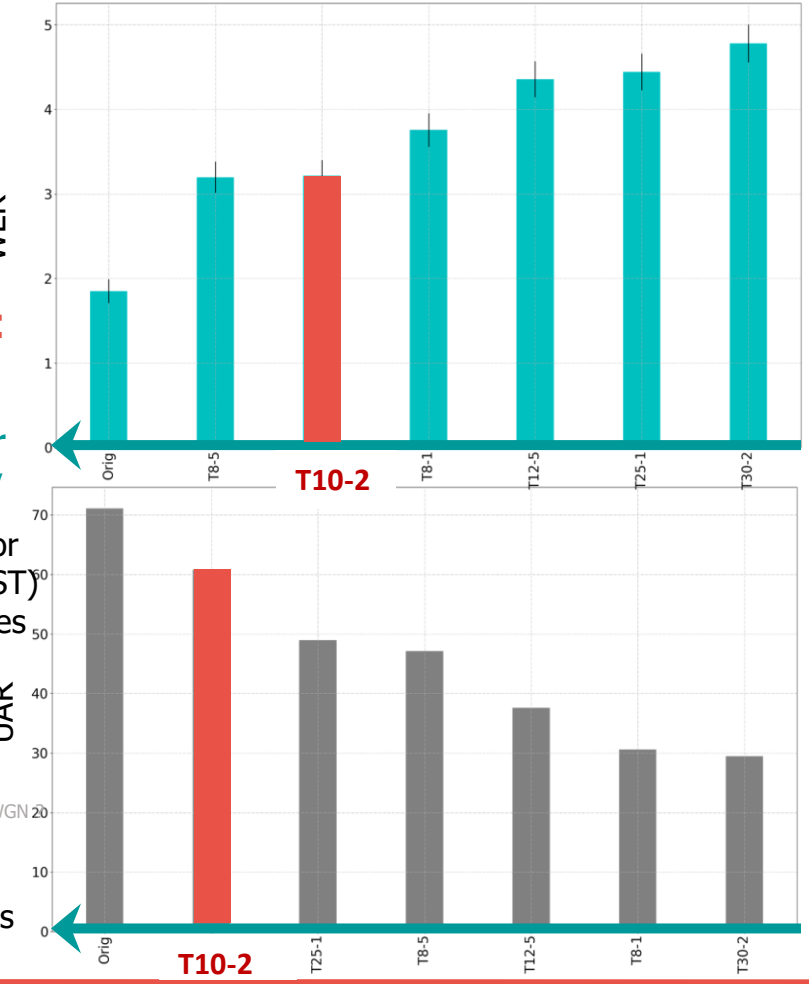


1. EER ≥ 10%
2. EER ≥ 20%
3. EER ≥ 30%
4. EER ≥ 40%

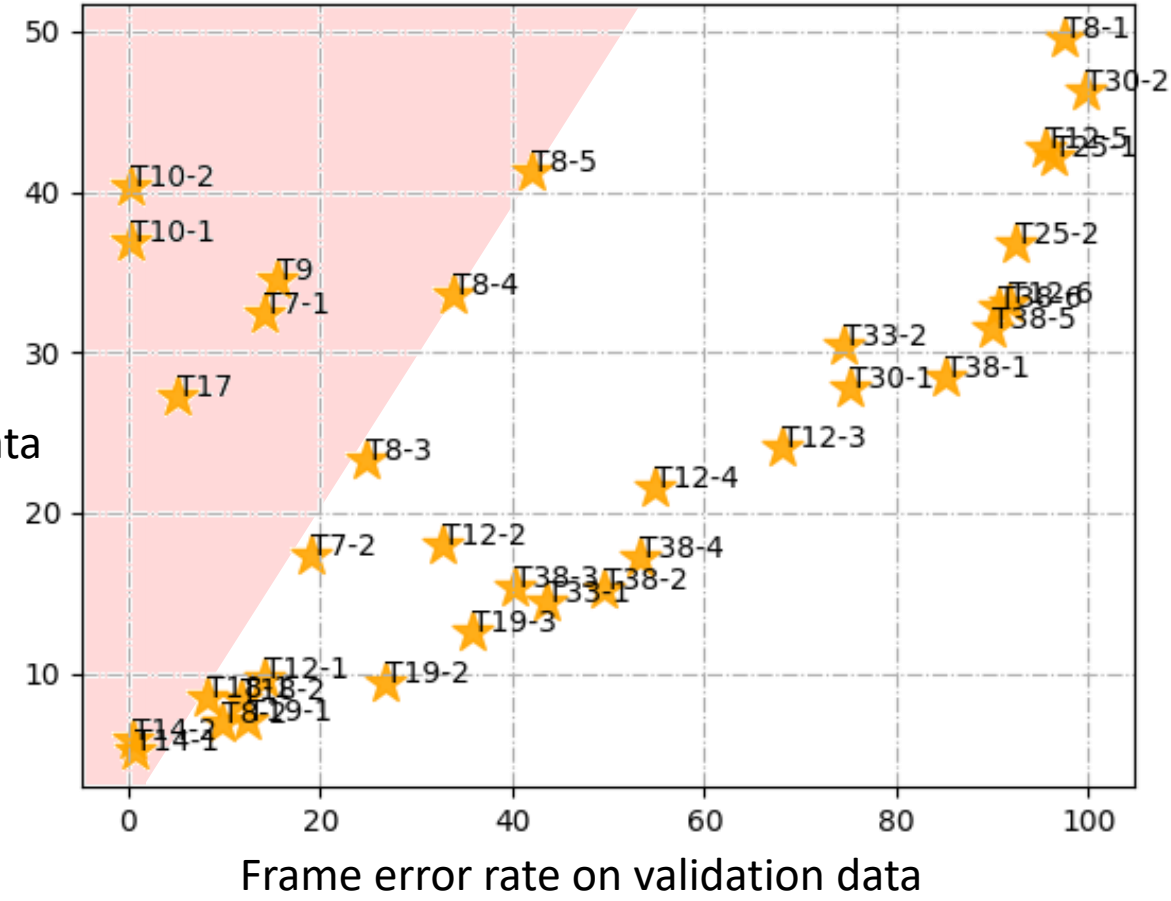
Best privacy systems:

- sequential disentanglement of attributes
- VC: disent. content & non-content; VQ-BN for content & emotion (GST) transfer from utterances of target speakers
- admixture
- Based on B5 + mean reversion F0 + AWGN
- ASR + TTS do not preserve emotions

Better utility



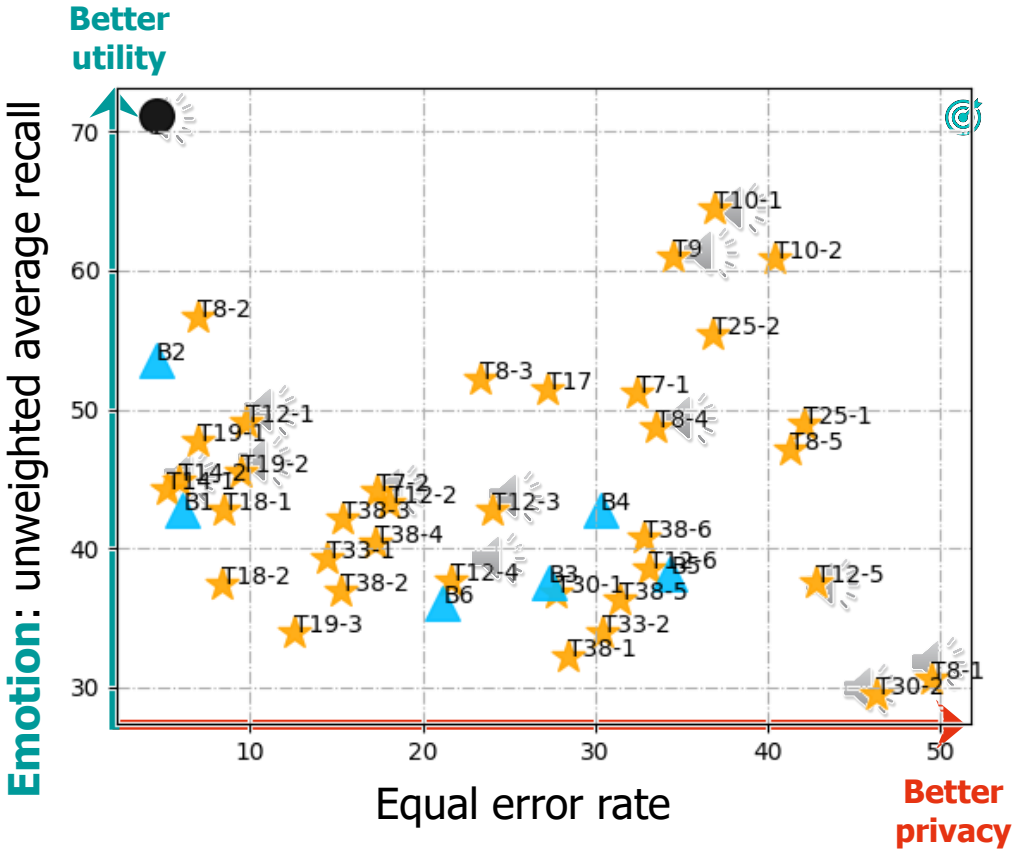
Privacy evaluation reliability

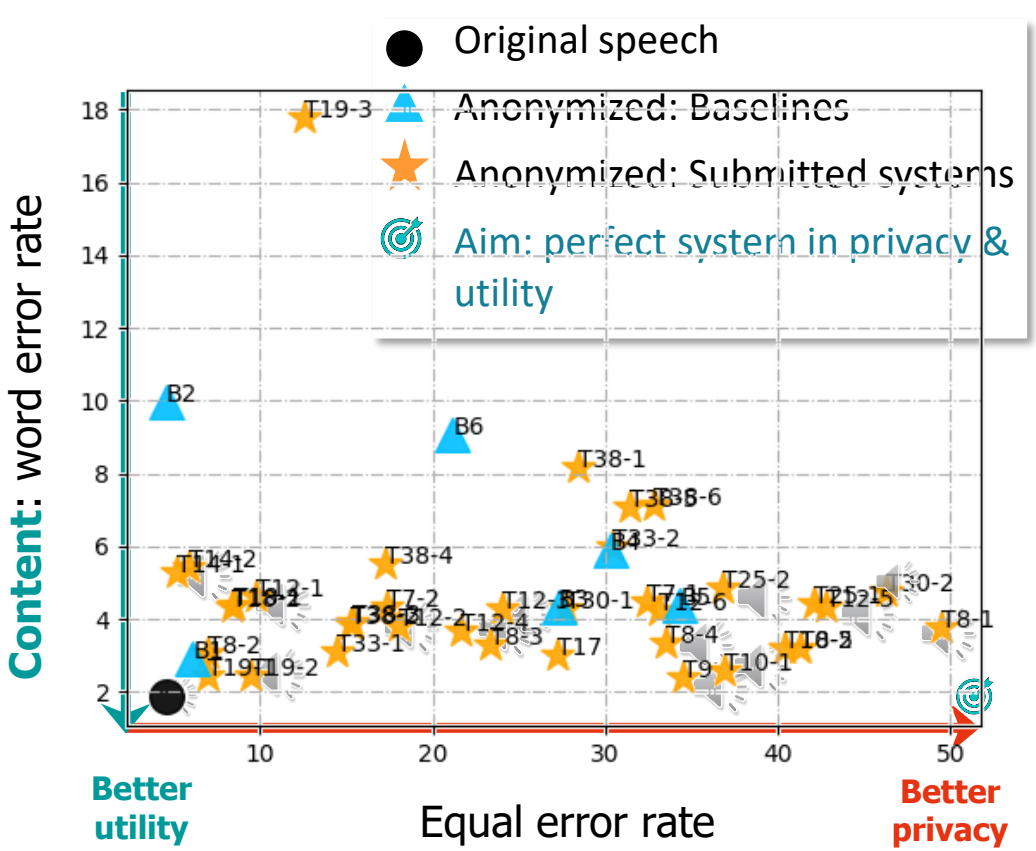


Systems with low FER on validation and high EER on test: suboptimal privacy evaluation?

?

Results 2024: privacy vs utility





Participant findings (selected)

- ✓ **T10** A serial disentanglement (distillation) strategy for content, speaker and emotion state – best results in emotion preservation (for all privacy categories) & best trade-off of privacy and both utility metrics in the last privacy category (EER ≥ 40%)
- ✓ **T8** Admixture using anonymization methods with different properties – a solution to flexibly adjust the privacy-utility threshold (the attacker model used in the VPC-2024 may be suboptimal for such method)
- ✓ **T38** Importance of using expressive datasets to train emotion encoders (but the best systems in UAR (**T10**) do not use them?)
- ✓ **T12** Methods to improve anonymization baselines B3 & B5

Conclusions

- ✓ Noticeable progress in anonymization methods: VC-based methods (with feature disentanglement and speech synthesis – more effective than signal-processing based)
- ✓ Possibility of using speech and emotion recognition with speaker anonymisation for speaker privacy protection
- ✓ Using large-scale (self-)supervised pre-trained models methods for speech processing models/encoders (WavLM, wav2vec2, Whisper,…) in VC-based methods to represent specific attributes (content, speaker, emotion,…)
- ✓ Diversity of approaches (and results)
- ✓ Evaluation challenges:
  - consistency & reliability of results (significant difference in EER for some systems, …)
  - setting thresholds for minimum target privacy requirements (task-dependent?)
  - strong and realistic attack models

Perspectives

✓ Anonymization methods

- improved information disentanglement and attribute modeling
- transferability across languages
- hybrid approaches with other privacy-preservation methods
- multi-modality

✓ Attributes

- (speaker , emotion, content, intonation, gender, accent, age, health status,…): controllable attribute modification

✓ Evaluation

- stronger and more realistic attack models, using metadata
- consider real-word scenarios vs theoretically motivated
- privacy vs utility trade-off – development & better ranking policy

References: participants' papers

T7 **Anemone** Emotional Speech Anonymization: Preserving Emotion Characteristics in Pseudo-speaker Speech Generation. *Hua Hua, Zengqiang Shang, Xuyuan Li, Peiyang Shi, Chen Yang, Li Wang, Pengyuan zhang*

T8 **JHU CLSP** HLTCOE JHU Submission to the Voice Privacy Challenge 2024. *Henry Li Xinyuan, Zexin Cai, Ashi Garg, Kevin Duh, Leibny Paola García-Perera, Sanjeev Khudanpur, Nicholas Andrews, Matthew Wiesner*

T9 **LongYuan** Speaker anonymization system with sentiment transfer and feature interpolation. *Tao Tan, Shutao Liu, Yibo Duan, Sheng Zhao, Xi Shao*

T10 **NPU-NTU** NPU-NTU System for Voice Privacy 2024 Challenge *Jixun YAO, Nikita Kuzmin, Qing Wang, Pengcheng Guo, Ziqian Ning, Dake Guo, Kong Aik Lee, Eng-Siong Chng, Lei Xie*

T12 **NTU-NPU** NTU-NPU System for Voice Privacy 2024 Challenge. *Nikita Kuzmin, Hieu-Thi Luong, Jixun YAO, Lei Xie, Kong Aik Lee, Eng-Siong Chng*

T14 **ADRES** Exploring Vector-quantized Variational Auto-Encoder with Prosody Parameters for Speaker Anonymization. *Sotheara Leang, Anderson Augusma, Dominique Vaufreydaz, Eric Castelli, Sethserey Sam, Frédérique Letué*

T17 **NKU HLT Lab** PANO: Facodec Anonymization System Enhanced with Prosody Anonymization. *Jiabei He, Jiaming Zhou, Haoqin Sun, Hui Wang, Yong Qin*

T18 **Q** Emotion-Enhanced Speaker Anonymisation Using the FreeVC Framework. *Yuqi Li, Yuanzhong Zheng, Jingyi Fang, Jinming Chen*

T19 **DFKI_SLT** Comparing Speech Anonymization Efficacy by Voice Conversion Using KNN and Disentangled Speaker Feature Representations. *Arnab Das, Carlos Franzreb, Tim Herzig, Philipp Pirlet, Tim Polzehl*

T25 **USTC-PolyU** A Voice Anonymization Method Based on Content and Non-content Disentanglement for Emotion Preservation. *Wenju Gu, Zeyan Liu, Liping Chen, Rui Wang, Chenyang Guo, Wu Guo, Kong Aik Lee, Zhen-Hua Ling*

T30 **V-Beam** Voice Anonymization Using Emotion-Enriched Feature Integration with STT and TTS Models. *Jeongae Lee, Taeje Park, Yeawon You*

T33 **KIT-ISL** Voice Privacy - Investigating Voice Conversion Architecture with Different Bottleneck Features. *Seymanur Akti, Tuan Nam Nguyen, Yining Liu, Alex Waibel*

T38 **Orange_Shiva** Tuning DISSC for Voice Privacy Challenge 2024. *Olivier Le Blouch, Rayane BAKARI, Nicolas Gengembre*

Congratulations!

T10 - **NPU-NTU**

Anonymization systems with the best trade-off of privacy protection and emotion preservation in all privacy categories & the best trade-off of privacy protection and preservation of both linguistic content and emotion in the 4th privacy category (EER>=40%)

Northwestern Polytechnical University, Nanyang Technological University, and Hong Kong Polytechnic University

Jixun Yao, Nikita Kuzmin, Qing Wang, Pengcheng Guo, Ziqian Ning, Dake Guo, Kong Aik Lee, Eng-Siong Chng, and Lei Xie

T9 - **LongYuan**

Anonymization system with the best trade-off of privacy protection and linguistic content preservation in the 1st, 2nd, and 3rd privacy categories (10<=EER<40%)

Nanjing University of Posts and Telecommunications, Nanjing Longyuan Information Technology Co. Ltd

Tao Tan, Shutao Liu, Yibo Duan, Sheng Zhao, and Xi Shao

T8 - **JHU CLSP**

Anonymization system with the best trade-off of privacy protection and linguistic content preservation in the 4th privacy category (EER>=40%)

Johns Hopkins University

Henry Li Xinyuan, Zexin Cai, Ashi Garg, Kevin Duh, Leibny Paola Garcia, Sanjeev Khudanpur, Nicholas Andrews, and Matthew Wiesner

- **5th December 2024** - Submission of results and system descriptions
- **9th December 2024** - Submission of 2-page papers to ICASSP 2025 (by invitation)
- **30th December 2024** Paper acceptance notification
- **13th January 2025** Camera-ready 2-page papers
- **6th-11th April 2025** ICASSP-2025
- **11th June 2025** - OJ-SP papers (by invitation)

Thank you!

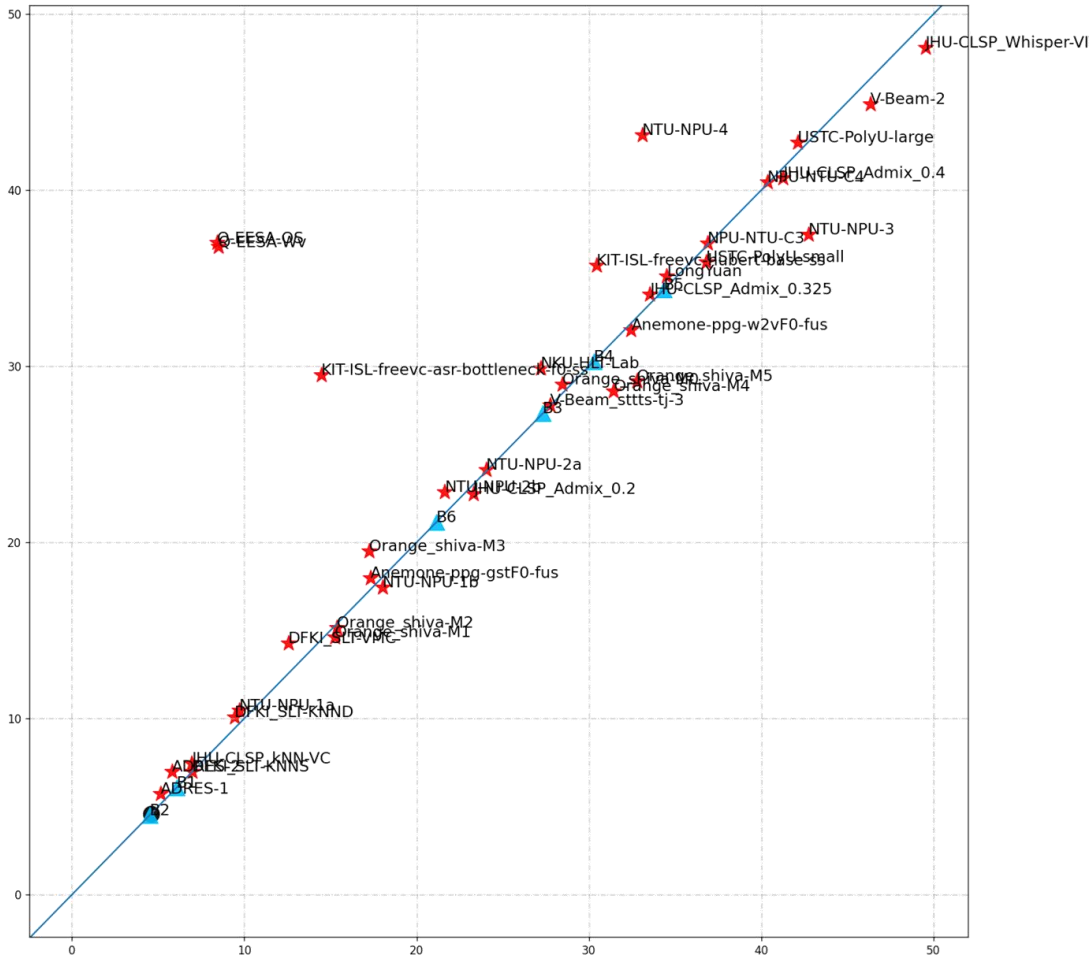
https://www.voiceprivacychallenge.org/attacker/
attacker.challenge@inria.fr

https://www.voiceprivacychallenge.org
organisers@lists.voiceprivacychallenge.org

Teams and systems

Team name	Team ID	Affiliation	System ID	System ID (paper)
Anemone	T7	Institute of Acoustics, Chinese Academy of Sciences, China University of the Chinese Academy of Sciences, China	T7-1	ppg-w2vF0-fusion
			T7-2	ppg-gstF0-fusion
JHU CLSP	T8	Johns Hopkins University, United States	T8-1	Whisper-VITS TTS
			T8-2	kNN-VC
			T8-3	Admixture (p=0.2)
			T8-4	Admixture (p=0.325)
			T8-5	Admixture (p=0.4)
LongYuan	T9	Auditory Intelligence Computing Group (AIC), Nanjing University of Posts and Telecommunications, China Nanjing Longyuan InformationTechnology Co.Ltd, China	T9	GMM-Blender
NPU-NTU	T10	Audio, Speech and Language Processing Group (ASLP@NPU) School of Computer Science, Northwestern Polytechnical University Nanyang Technological University The Hong Kong Polytechnic University	T10-1	C3
			T10-2	C4
NTU-NPU	T12	Nanyang Technological University, China Institute for Infocomm Research, A*STAR, Singapore Audio, Speech and Language Processing Group (ASLP@NPU), China The Hong Kong Polytechnic University	T12-1	1a
			T12-2	1b
			T12-3	2a
			T12-4	2b
			T12-5	3
			T12-6	4
ADRES	T14	Univ. Grenoble Alpes, CNRS, Grenoble LIG, France Institute of Digital Research and Innovation, CADT, Phnom Penh, Cambodia Univ. Grenoble Alpes, CNRS, LJK, Grenoble, France	T14-1	v1
			T14-2	v2
NKU HLT Lab	T17	HLT Lab, College of Computer Science, Nankai University, China	T17	Facodec
Q	T18	Qifu Technology, China Fudan University, Shanghai, China	T18-1	EESA(Wv)
			T18-2	EESA(OS)
DFKI_SLT	T19	Speech and Language Technology, German Research Center for Artificial Intelligence (DFKI), Germany Quality and Usability Lab, Technische Universitat Berlin, Germany	T19-1	KNNS
			T19-2	KNND
			T19-3	VMC
USTC-PolyU	T25	NERC-SLIP, University of Science and Technology of China, China Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong	T25-1	large: ESD+LibriTTS
			T25-2	small: ESD only
V-Beam	T30	Sogang University, Seoul, Korea Ewha Womans University, Seoul, Korea	T30-2	2 - V-Beam_method2
			T30-1	3 V-Beam_sttts-tj-method3
KIT-ISL	T33	Karlsruhe Institute of Technology (KIT), Germany Carnegie Mellon University (CMU), USA	T33-2	freevc-hubert-base-ss
			T33-1	freevc-asr-bottleneck-f0-ss
Orange_shiva	T38	Orange, France	T38-1	M0
			T38-2	M1
			T38-3	M2
			T38-4	M3
			T38-5	M4
			T38-6	M5

EER results verification (retrained vs submitted)



Results: privacy vs utility