The VoicePrivacy 2022 Challenge

Second Symposium on Security and Privacy in Speech Communication

23-24th September 2022
Incheon, Korea

Natalia Tomashenko
Xin Wang
Xiaoxiao Miao
Hubert Nourtel
Pierre Champion
Massimiliano Todisco
Emmanuel Vincent
Nicholas Evans
Junichi Yamagishi
Jean-François Bonastre
Michele Panariello

Organizers

Natalia Tomashenko
LIA, University of Avignon, France

Xin Wang
NII, Japan

Xiaoxiao Miao
NII, Japan

Hubert Nourtel
Inria, France

Pierre Champion
Inria, LIUM, France

Massimiliano Todisco
EURECOM, France

Emmanuel Vincent
Inria, France

Nicholas Evans
EURECOM, France

Junichi Yamagishi
NII, Japan
University of Edinburgh, UK

Jean-François Bonastre
LIA, University of Avignon, France

Michele Panariello
EURECOM, France

Acknowledgment

- ISCA

INTERSPEECH 2022
September 18 - 22 • Incheon Korea

- INTERSPEECH satellite event organizers Jong Won Shin, Yunjung Kim, Wenwu Wang
- SPSC Symposium organizers
- HYUNDAI ASAN Hyeseon Lucy Chung, Gahyung Han, …
- VoicePrivacy Challenge participants

Introduction: VoicePrivacy Initiative

Promote the development of privacy preservation tools for speech technology

- ✓2020 **First** VoicePrivacy Challenge
  - Special session at Interspeech 2020
  - Satellite workshop at Speaker Odyssey 2020
- ✓2022 **Second** edition
  - Satellite ISCA workshop at Interspeech 2022 (SPSC Symposium)

Privacy preservation for speech and challenge focus

Anonymization

✗ remove personally identifiable information in the speech signal

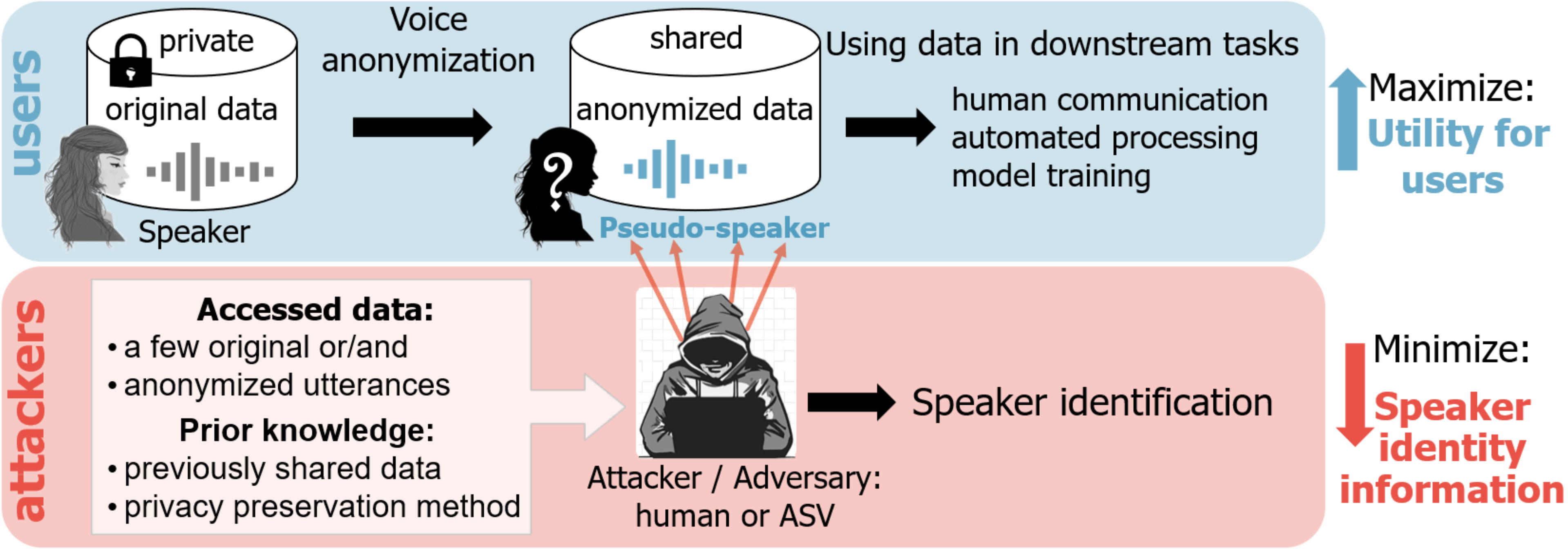
✓ keep all other characteristics unchanged

- linguistic content
- paralinguistic attributes
- speech intelligibility/naturalness

…

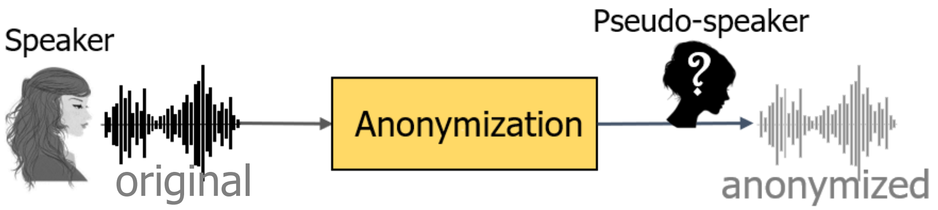
Anonymization task

- Privacy preservation is formulated as a **game** between **users** (share some data) & **attackers** (access this data or data derived from it and wish to infer information about the users)



Challenge task and requirements

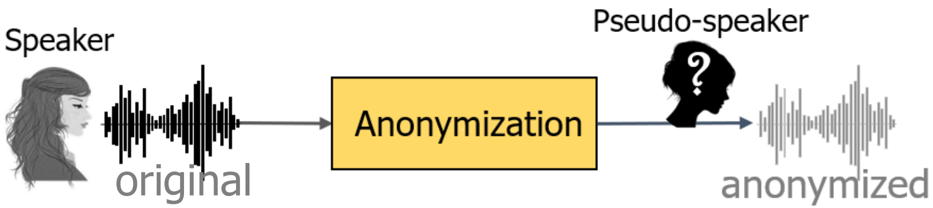
Task: develop an anonymization system



- ✓ conceal the speaker identity;
- ✓ leave the linguistic content and paralinguistic attributes unchanged;
- ✓ ensure that all trial utterances from a given speaker are uttered by the same pseudo-speaker while trial utterances from different speakers are uttered by different pseudo-speakers (speaker-level anonymization; voice distinctiveness preservation)

Challenge task and requirements

Task: develop an anonymization system



- ✓ conceal the speaker identity;
- ✓ leave the linguistic content and paralinguistic attributes unchanged;
- ✓ ensure that all trial utterances from a given speaker are uttered by the same pseudo-speaker while trial utterances from different speakers are uttered by different pseudo-speakers (speaker-level anonymization; voice distinctiveness preservation)

We provide:

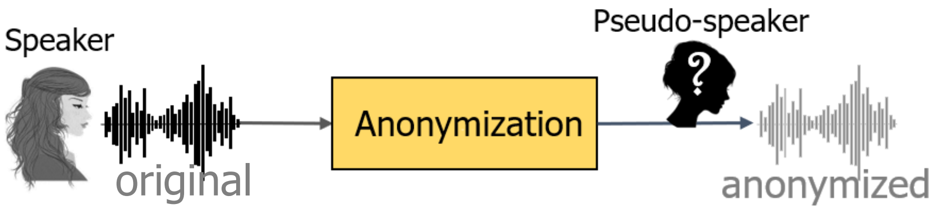
- ✓ training, development and evaluation datasets
- ✓ 3 different baseline anonymization systems
- ✓ evaluation scripts and metrics

Participants:

- ✓ apply their developed anonymization systems, run evaluation scripts
- ✓ submit objective evaluation results and anonymized speech data to the organizers

Challenge task and requirements

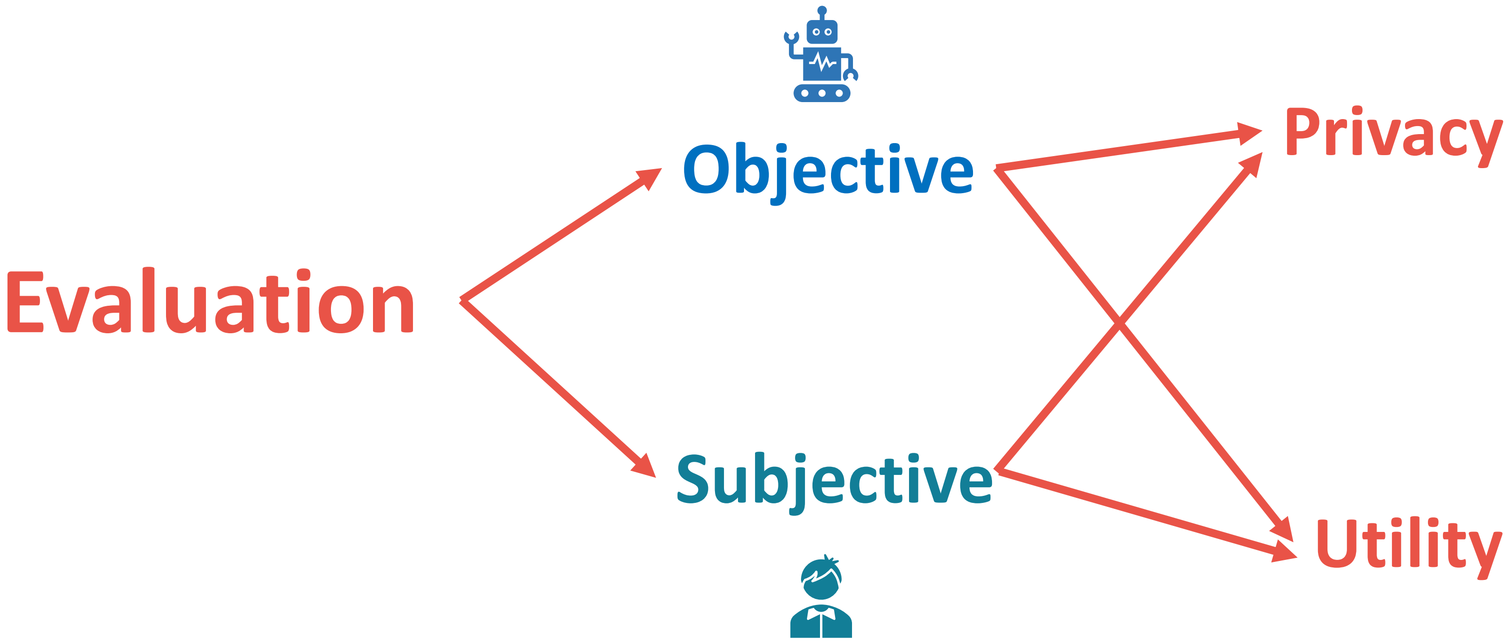
Task: develop an anonymization system



- ✓ conceal the speaker identity;
- ✓ leave the linguistic content and paralinguistic attributes unchanged;
- ✓ ensure that all trial utterances from a given speaker are uttered by the same pseudo-speaker while trial utterances from different speakers are uttered by different pseudo-speakers (speaker-level anonymization; voice distinctiveness preservation)

We provide:
- ✓ training, development and evaluation datasets
- ✓ 3 different baseline anonymization systems
- ✓ evaluation scripts and metrics

Participants:
- ✓ apply their developed anonymization systems, run evaluation scripts
- ✓ submit objective evaluation results and anonymized speech data to the organizers


Evaluation
Objective
Subjective
Privacy
Utility

Objective evaluation: primary privacy and utility metrics

Privacy

$\mathbf{ASV_{eval}}$ Automatic speaker verification system = **attacker**

Equal error rate $\mathrm{EER} = P_{\mathrm{fa}}(\theta_{\mathrm{EER}}) = P_{\mathrm{miss}}(\theta_{\mathrm{EER}})$

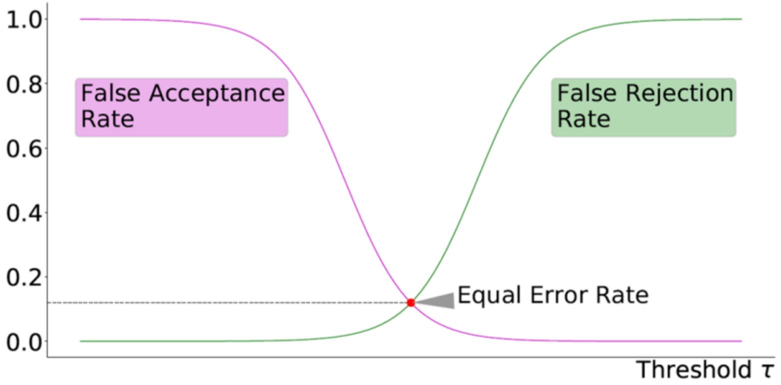


Figure from [E .Vincent 2022]

larger EER => better privacy

Utility

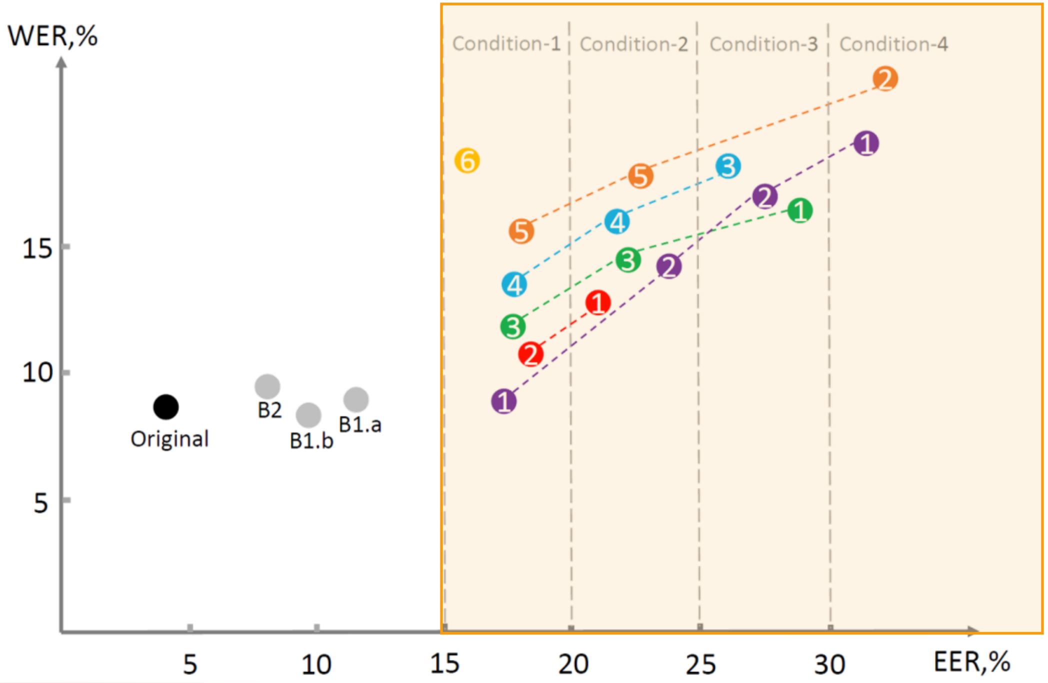
$\mathbf{ASR_{eval}}$ Automatic speech recognition system

Word error rate $$\mathrm{WER} = \frac{N_{\mathrm{sub}} + N_{\mathrm{del}} + N_{\mathrm{ins}}}{N_{\mathrm{ref}}}$$

smaller WER => better utility

Ranking policy:

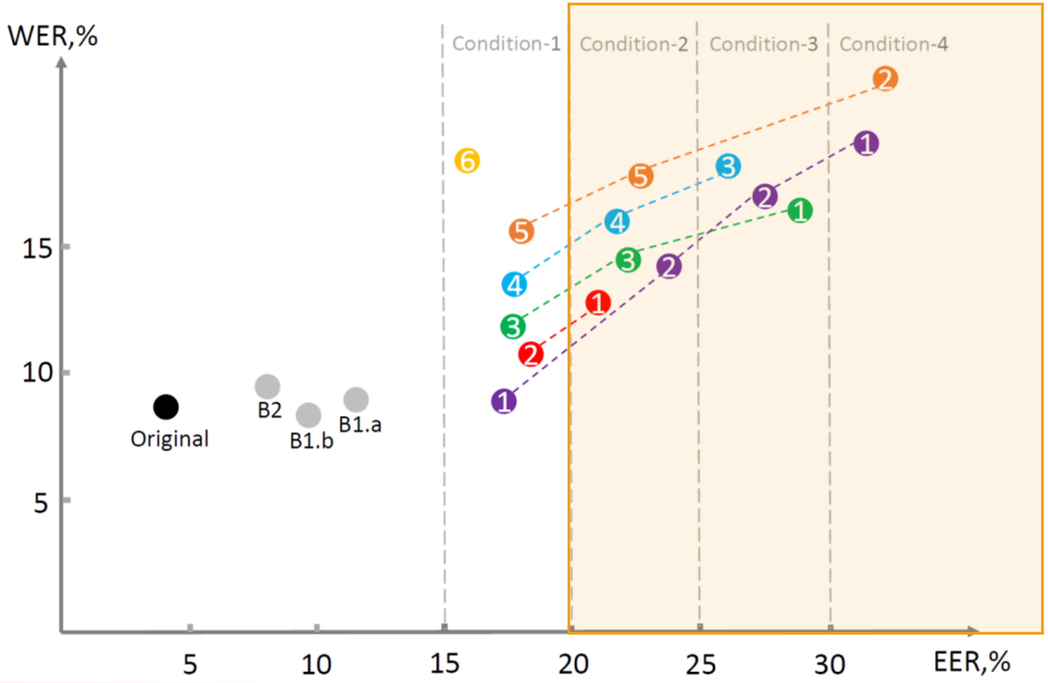
- ⭐ New to the 2022 edition:
- Use of **multiple** evaluation **conditions** specified with a set of **minimum target privacy** requirements:
- To measure the **privacy-utility trade-off** of any solution at multiple operating points



1. EER ≥15%
2. EER ≥ 20%
3. EER ≥ 25%
4. EER ≥ 30%.

Ranking policy:

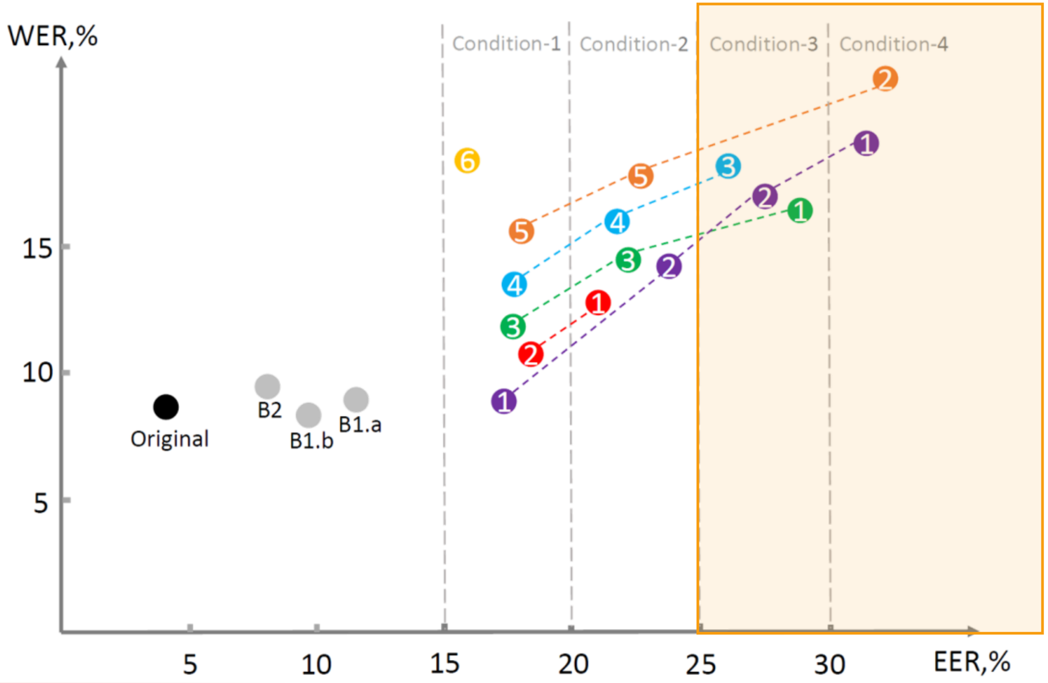
- ⭐ New to the 2022 edition
- Use of **multiple** evaluation **conditions** specified with a set of **minimum target privacy** requirements:
- To measure the **privacy-utility trade-off** of any solution at multiple operating points



1. EER ≥15%
2. EER ≥ 20%
3. EER ≥ 25%
4. EER ≥ 30%.

Ranking policy:

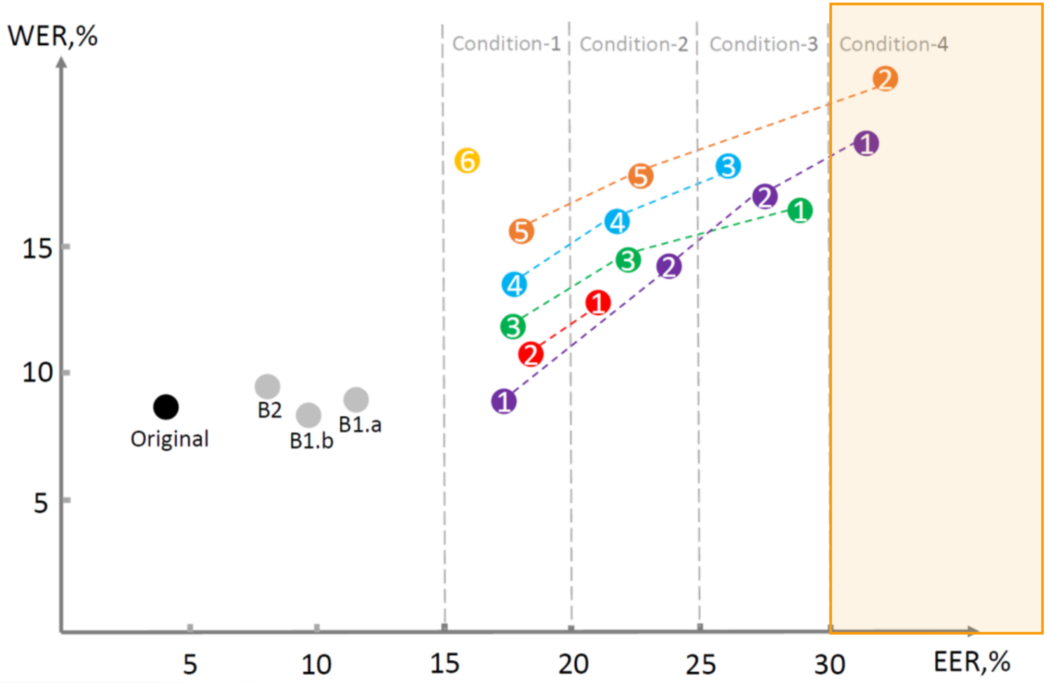
- ⭐ New to the 2022 edition
- Use of **multiple** evaluation **conditions** specified with a set of **minimum target privacy** requirements:
- To measure the **privacy-utility trade-off** of any solution at multiple operating points



1. EER ≥15%
2. EER ≥ 20%
3. EER ≥ 25%
4. EER ≥ 30%.

Ranking policy:

- ⭐New to the 2022 edition
- Use of **multiple** evaluation **conditions** specified with a set of **minimum target privacy** requirements:
- To measure the **privacy-utility trade-off** of any solution at multiple operating points

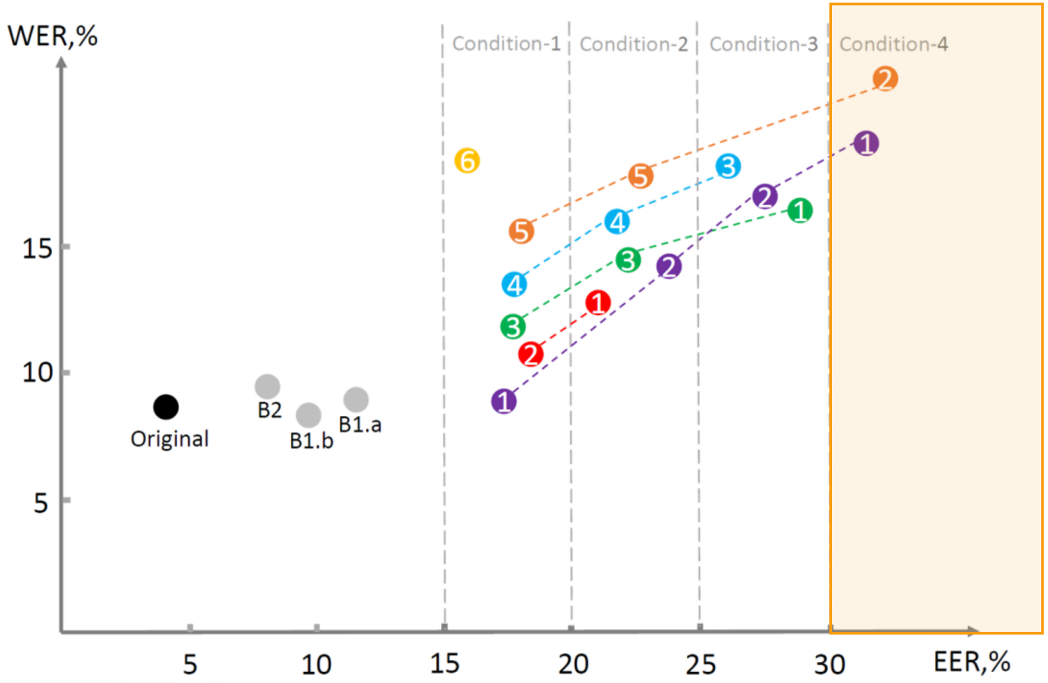


1. EER ≥15%
2. EER ≥ 20%
3. EER ≥ 25%
4. EER ≥ 30%.

Ranking policy:

$$\mathrm{EER}_{aver} = 0.5 \cdot \mathrm{EER}_{LibriSpeech} + 0.4 \cdot \mathrm{EER}_{VCTK\_diff} + 0.1 \cdot \mathrm{EER}_{VCTK\_common}$$

$$\mathrm{WER}_{aver} = 0.5 \cdot \mathrm{WER}_{LibriSpeech} + 0.5 \cdot \mathrm{WER}_{VCTK}$$



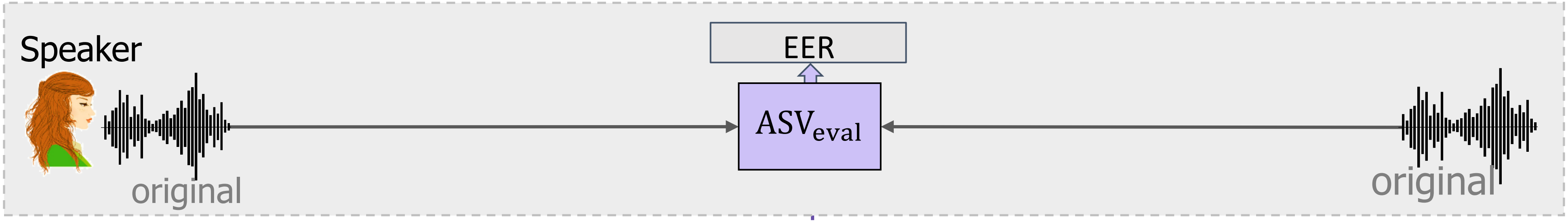
1. EER ≥15%
2. EER ≥ 20%
3. EER ≥ 25%
4. EER ≥ 30%.

Objective privacy evaluation: automatic speaker verification

- **Original** (no anonymization)

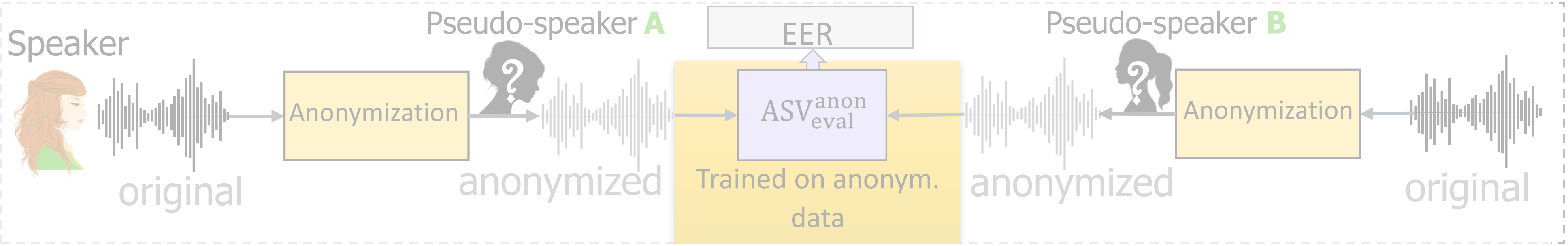
Trials | Enrollment

1



- **Semi-informed attacker:** (☆new stronger attacker in 2022 edition w.r.t 2020)
  - retrains the ASV system anonymized data on **utterance-level** ← more efficient than speaker-level
  - anonymizes enrollment data on **speaker-level**

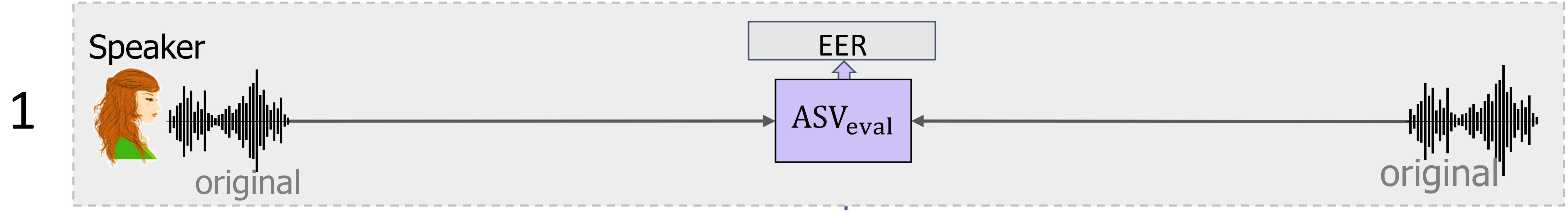
2



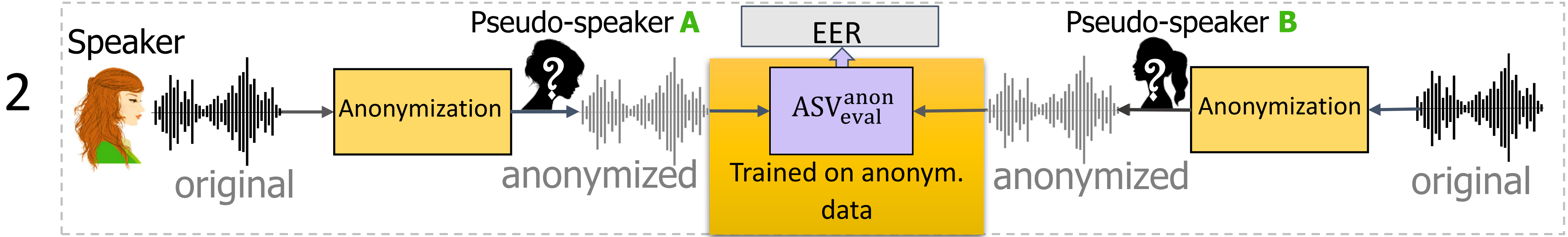
Objective privacy evaluation: automatic speaker verification

- **Original** (no anonymization)

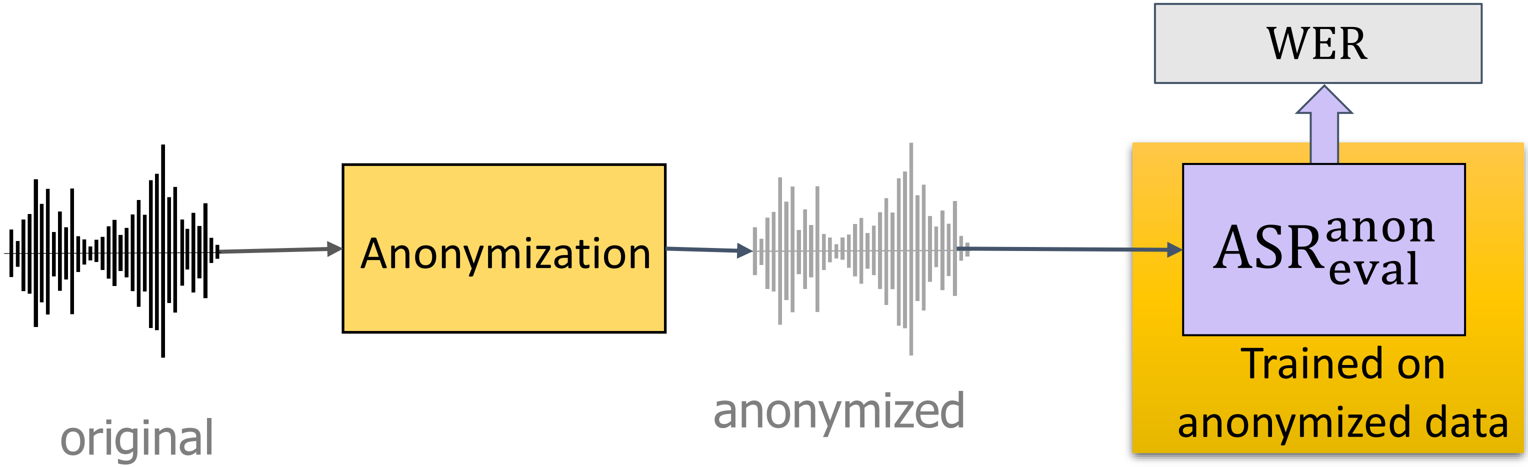
Trials | Enrollment



- **Semi-informed attacker:** (★new stronger attacker in 2022 edition w.r.t 2020)
  - retrains the ASV system anonymized data on **utterance-level** ← more efficient than speaker-level
  - anonymizes enrollment data on **speaker-level**



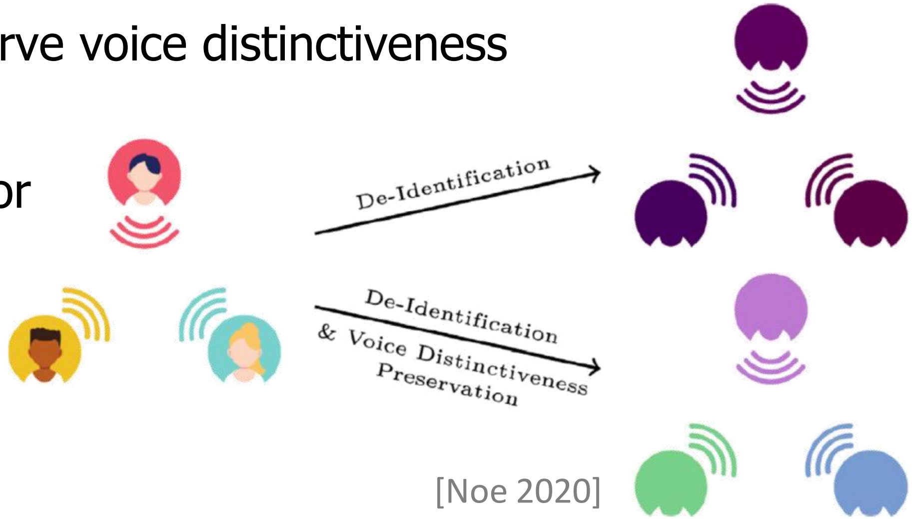
Objective utility evaluation



Automatic speech recognition (ASR) system trained on anonymized data
Metric: Word error rate (WER), lower WER => better utility

Secondary utility metrics

1. **Pitch correlation** between original and anonymized utterances ρ^{F_0}
   - intonation should be preserved in anonymized speech
   - $\rho^{F_0} \leq 1$, higher is better
   - requirement for all datasets: $\rho^{F_0} > 0.3$
2. **Gain of voice distinctiveness** G_{VD}
   - aims to evaluate the requirement to preserve voice distinctiveness
   - relies on voice similarity matrices
   - important to keep distinguishable voices for multi-party human conversation



[Noe 2020]

Secondary utility metrics

1. **Pitch correlation** between original and anonymized utterances ρ^{F_0}
   - intonation should be preserved in anonymized speech
   - $\rho^{F_0} \leq 1$, higher is better
   - requirement for all datasets: $\rho^{F_0} > 0.3$
2. **Gain of voice distinctiveness** G_{VD}
   - aims to evaluate the requirement to preserve voice distinctiveness
   - relies on voice similarity matrices

$$G_{\mathrm{VD}} = 10 \log_{10} \left(D_{\mathrm{diag}}(M_{aa}) / D_{\mathrm{diag}}(M_{oo}) \right)$$

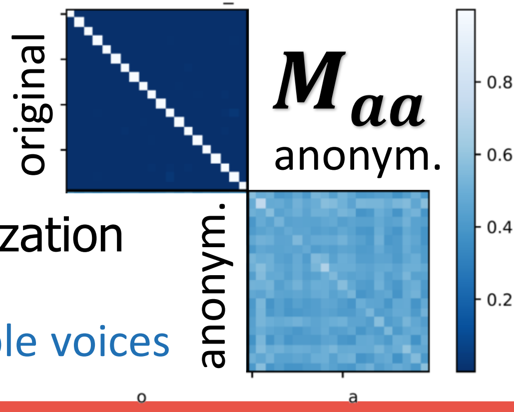
   - higher is better
   - G_{VD}=0 => voice distinctiveness remains the same after anonymization

M_{oo}

original

M_{aa}

anonym.

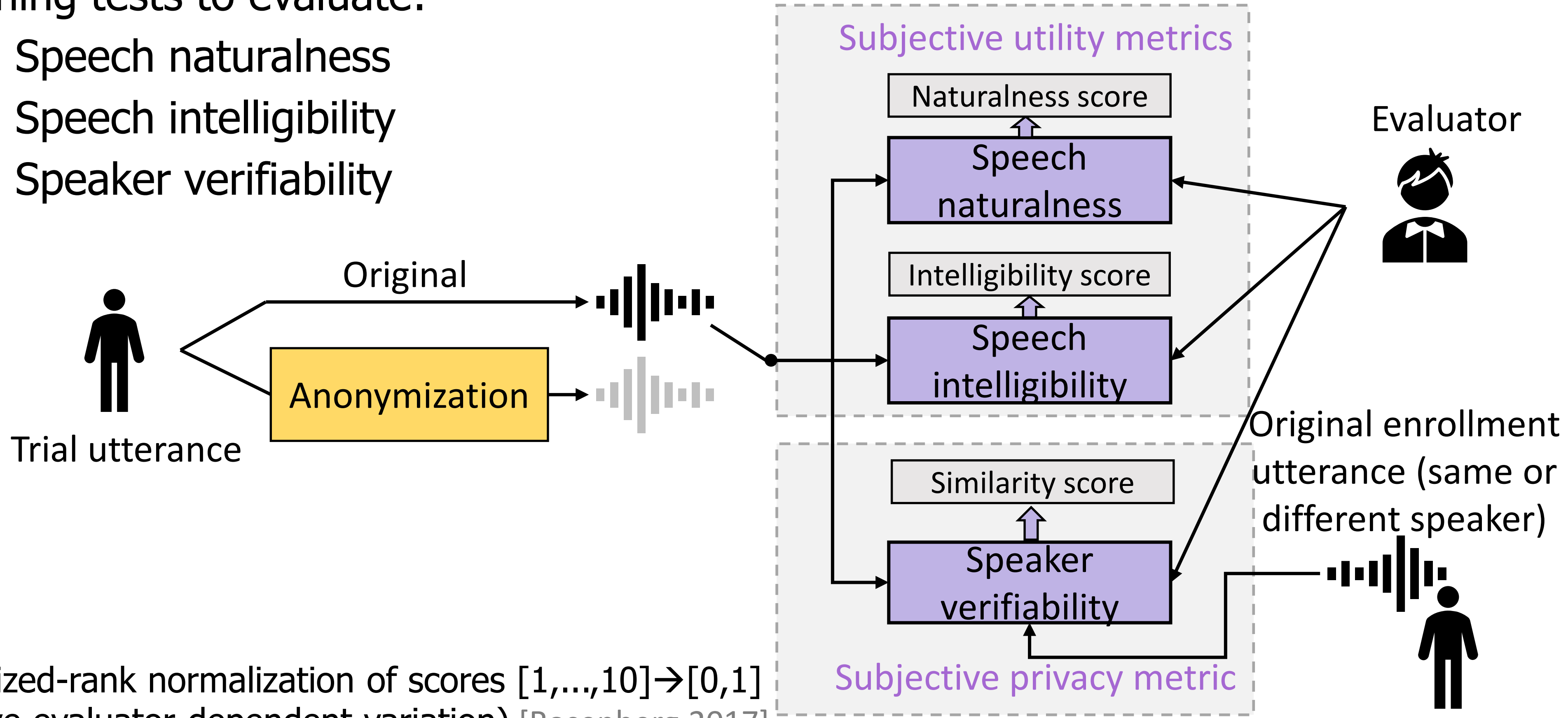


Clear diagonal⇔distinguishable voices

Subjective evaluation design

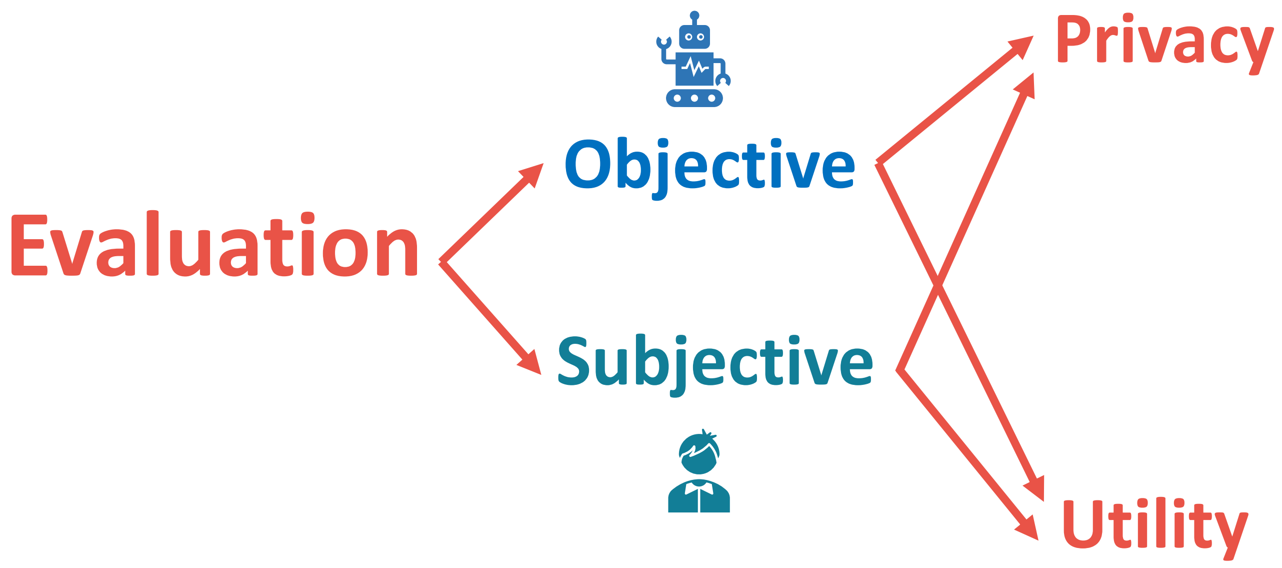
Listening tests to evaluate:

- ✓ Speech naturalness
- ✓ Speech intelligibility
- ✓ Speaker verifiability



+ normalized-rank normalization of scores [1,...,10]→[0,1] (to remove evaluator-dependent variation) [Rosenberg 2017]

Evaluation metrics summary



Privacy:

- **Equal error rate EER** — primary
- Subjective speaker verifiability

Utility:

- **Word error rate WER** — primary
- Pitch correlation ρ^{F_0}
- Gain of voice distinctiveness G_{VD}
- Subjective speech naturalness
- Subjective speech intelligibility

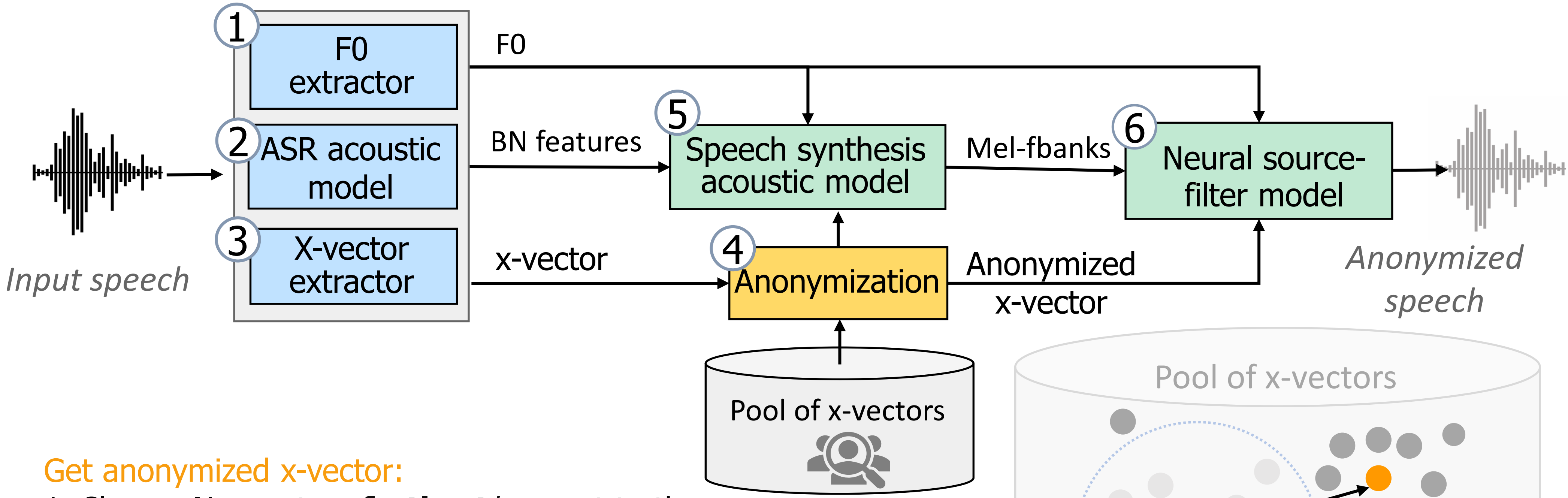
Datasets

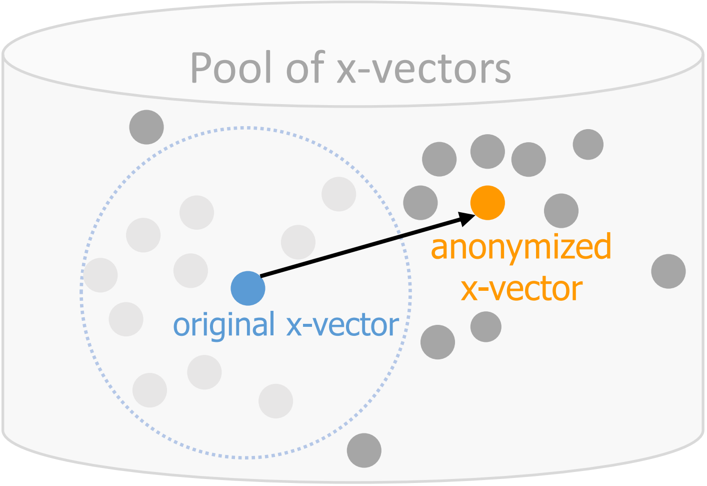
Training	Speakers	Size, h
VoxCeleb-1,2	7363	2794
LibriSpeech: -train-clean-100	251	100
-train-other-500	1166	497
LibriTTS: -train-clean-100	247	54
-train-other-500	1160	310

Development	Speakers	Target trials	Imposter trials
LibriSpeech: -dev-clean	29	1348	27362
VCTK-dev: -common	30	695	9721
VCTK-dev: -different		3796	26204

Evaluation	Speakers	Target trials	Imposter trials
LibriSpeech: -test-clean	29	997	20653
VCTK-test: -common	30	700	9790
VCTK-test: -different		3686	26314

Baseline B1.a: using x-vectors and neural waveform models

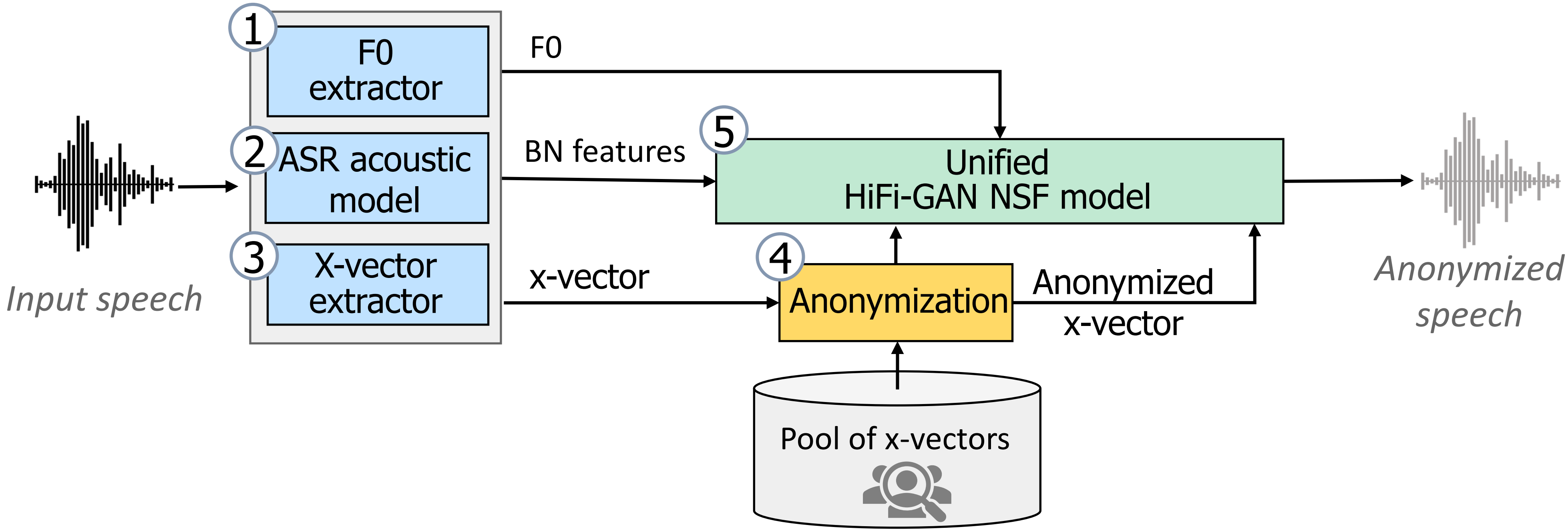




Get anonymized x-vector:

1. Choose N x-vectors **farthest**/nearest to the original one (**PLDA**/cosine)
2. Choose N*<N randomly from them
3. Average N* x-vectors to obtain an anonymized x-vector

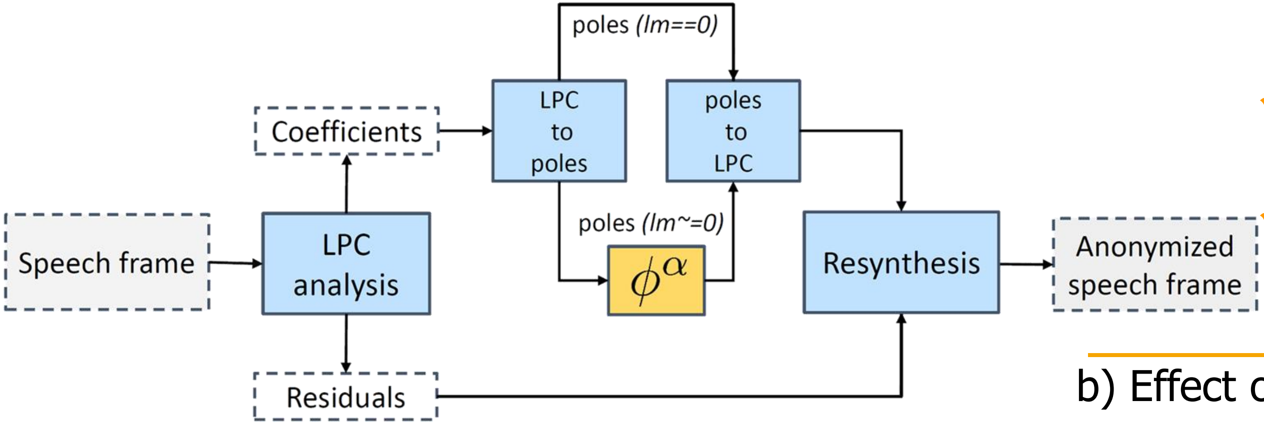
Baseline B1.b: using x-vectors and neural waveform models



- ✓ ⭐ New (2022 edition)
- ✓ Simplified (unified) TTS part
- ✓ Better speech quality

Baseline B2: using McAdams coefficient

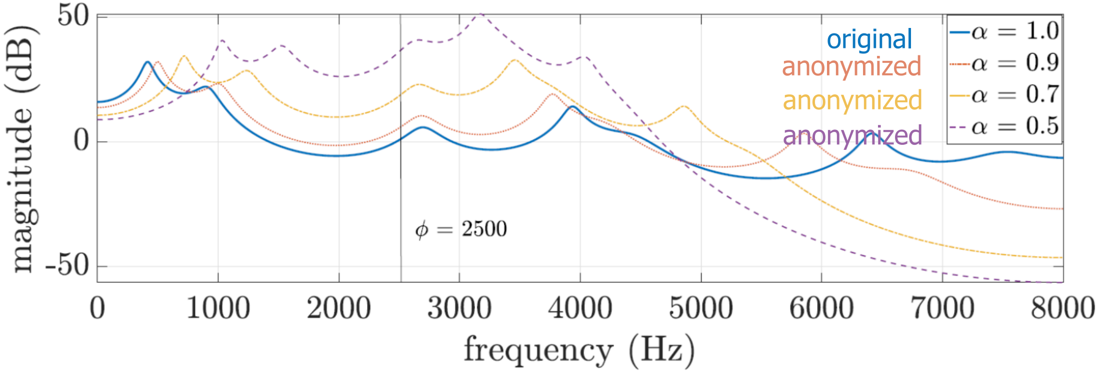
McAdams coefficient α provokes shifts in formants derived from the linear predictive coding (LPC) analysis



a) Overview of the LPC-based pipeline

- ✓ Simple to apply anonymization: single parameter α
- ✓ No training data is required

b) Effect of the McAdams coefficient upon formants

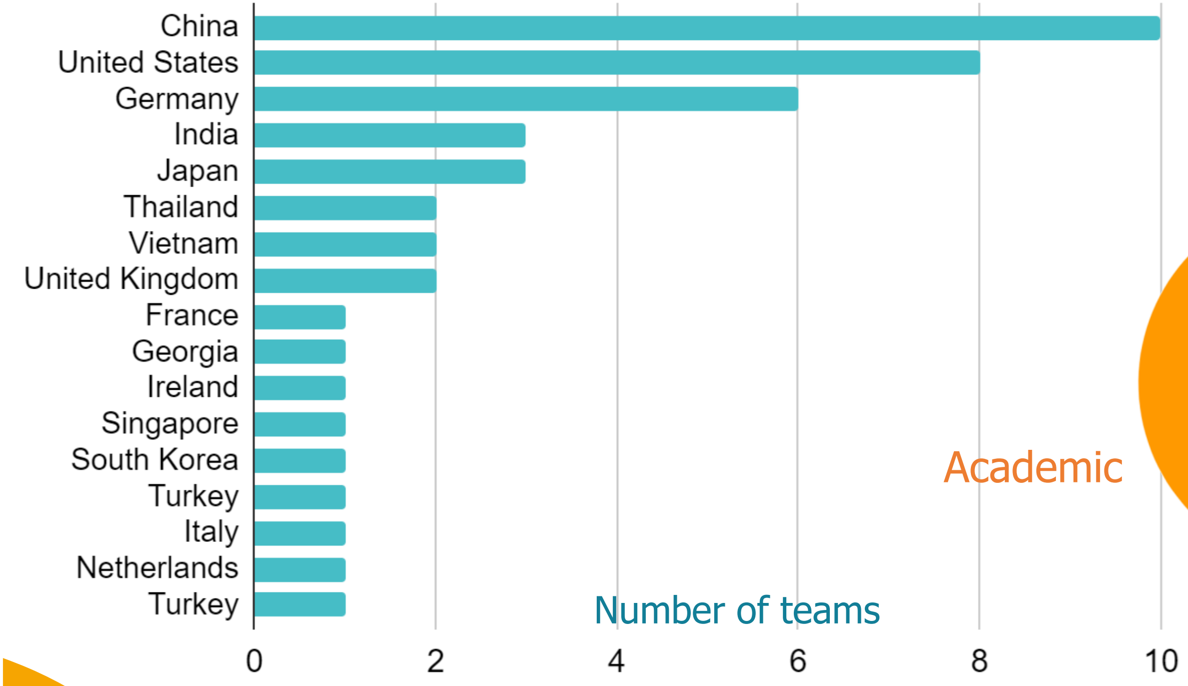


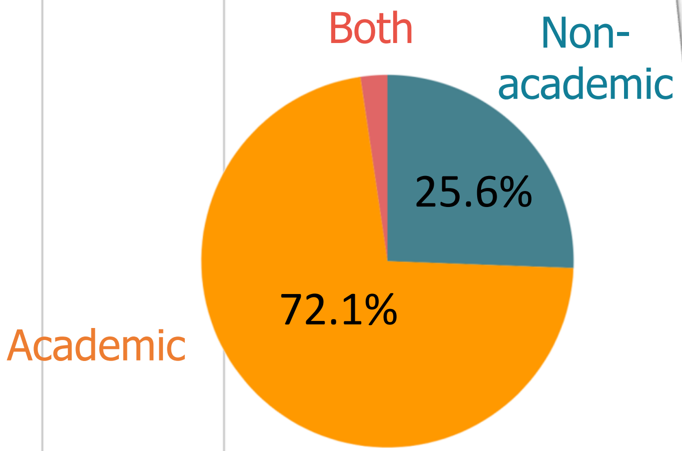
- ✓ 2020: α=0.5
- ✓ 2022: $\alpha \sim U(0.5, 0.9)$ ★

[McAdams 1984] [Patino 2020]

Participants

- Registered teams: **43** (more than **79** participants) from **17** countries
- Teams submitted valid results: **6**
- Submitted anonymization systems: **16**





	Team	Country	Status
1	Hyperconnect	South Korea	Nonacademic
2	SpectrumAI	United States	Nonacademic
3	kuaiyin	China	Nonacademic
4	IMS	Germany	Academic
5	UR_AIR	United States	Academic
6	KGP	India	Academic
7	ElectricSheep	Netherlands	Academic
8	VoiceDenzer	China	Academic
9	Horizon	China	Nonacademic
10	digis-speechlab	Italy	Academic
11	NWPU-ASLP	China	Academic
12	DarkHorse	Turkey	Academic
13	JU UAV Innovator's Lab	India	Nonacademic
14	JAIST-AIS	Japan	Academic
15	NCSU WSPR	United States	Academic
16	DCU	Ireland	Academic
17	OVGU team	Germany	Academic
18	KK (Kyoto-Kwai) team	China, Japan,	Both
19	MIT CCC	United States	Academic
20	N-ICL	United States, U	Nonacademic
21	Metamason	Vietnam	Academic
22	CKC Voice Privacy	China	Academic
23	S3L	China	Academic
24	voID	Thailand	Nonacademic
25	ThinkIT	China	Academic
26	Biometric team	Thailand	Academic
27	STAPRL	Germany	Academic
28	Team one	France	Academic
29	Pattern Recognition Lab	Germany	Academic
30	ningxinhuang	China	Academic
31	CAISA lab	Germany	Academic
32	HIS-JAIST	Japan	Academic
33	ECT team	China	Academic
34	Team	Georgia	Nonacademic
35	VTCC	Vietnam	Nonacademic
36	SPEECH_CSE	India	Academic
37	NUS UbicompLab	Singapore	Academic
38	Mac CAS	Canada	Academic
39	Soton	United Kingdom	Academic
40	Audio Labs Erlangen	Germany	Academic
		United States	Nonacade

Teams and systems

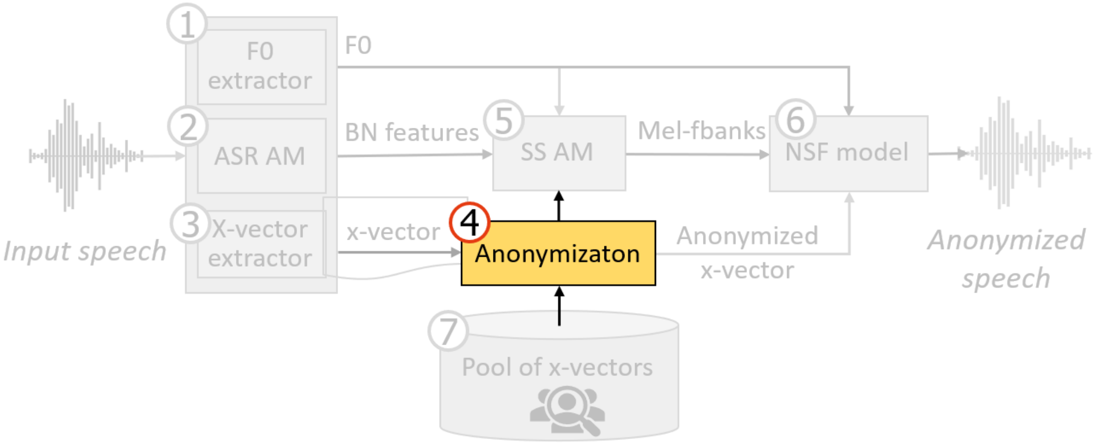
Team	Affiliation(s)	Team notation	Systems	System notation
IMS	- Institute for Natural Language Processing (IMS), University of Stuttgart, Germany	T04	primary.1	T04-p1
Horizon	- N/A	T09	primary.1	T09-p1
			primary.2	T09-p2
			contrastive.1.1	T09-c1
			contrastive.1.2	T09-c2
			contrastive.2.1	T09-c3
			contrastive.2.2	T09-c4
NWPU-ASLP	- Audio, Speech and Langauge Processing Group (ASLP@NPU), School of Computer Science, Northwestern Polytechnical University, China	T11	primary.1	T11-p1
			primary.2	T11-p2
			primary.3	T11-p3
			primary.4	T11-p4
KK team (Kyoto-Kwai team)	- Xinjiang University, Urumqi, China - Kyoto University, Kyoto, Japan - National Institute of Information and Communications Technology (NICT), Kyoto, Japan - Kuaishou Technology, Beijing, China	T18	primary.1	T18-p1
			contrastive.1.1	T18-c1
HIS-JAIST	- Japan Advanced Institute of Science and Technology, Japan	T32	primary.1	T32-p1
			contrastive.1.1	T32-c1
Audio Labs Erlangen	- Friedrich-Alexander-Universitat, International Audio Laboratories Erlangen, Germany - Fraunhofer IIS, Erlangen, Germany	T40	primary.1	T40-p1

Participants' systems

Two types of methods:

1) x-vector / speaker embedding based neural model

~Baseline **B1.a, B1.b**



Systems: **T04, T09, T11, T18, T40**

2) signal-processing

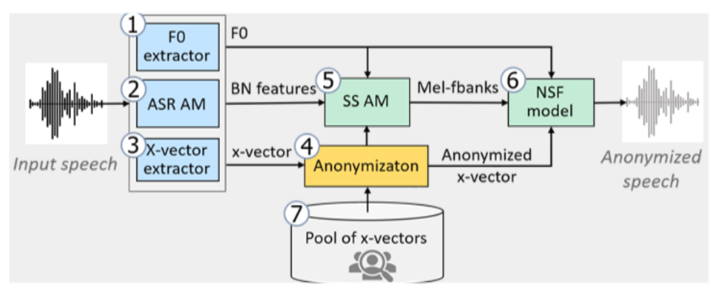
~Baseline **B2**

- modifications in formants, pitch, and speaking rate
- McAdams

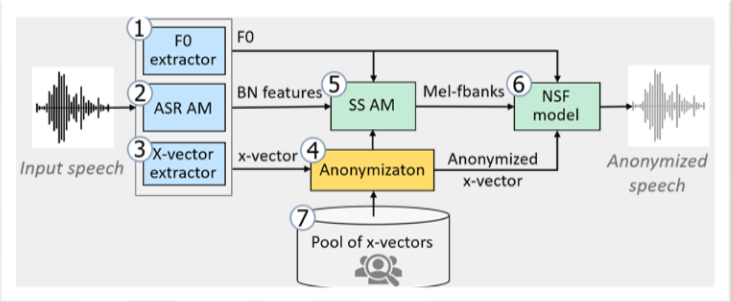
Systems: **T32**

Participants' systems

System	Description			1	2	3	4	5	6	7	Data
				Modified components & Data in B1*							
T04-p1	phonetic speech recognition; speaker embedding anonymization via GAN; multi-speaker SS; no usage of original pitch			+	+	+	+	+	+	+	ASR: LibriTTS-train-clean-100, LibriTTS-train-other-500; VoxCeleb-1,2 data (with ASR output transcripts)
T09-p1	replace architecture for all the models + voice/unvoiced features;	gender selection	same	+	+	+	+	+	+	+	
T09-p2			opposite	+	+	+	+	+	+	+	
T09-c1			random	+	+	+	+	+	+	+	
T09-c2			same	+	+	+	+	+	+	+	Speaker pool: LibriTTS-train-other-500 + VoxCeleb-1,2
T09-c3			opposite	+	+	+	+	+	+	+	
T09-c4			random	+	+	+	+	+	+	+	
T11-p1	replace x-vectors by speaker ids from a look-up table + speaker encoder; replaced architecture for all the models			+	+	+	+	+	+	+	
T11-p2				+	+	+	+	+	+	+	
T11-p3				+	+	+	+	+	+	+	
T11-p4				+	+	+	+	+	+	+	
T18-p1	adding adversarial noise to x-vectors						+				
T18-c1	replace x-vectors by ASR embeddings					+					ASRspk: LibriSpeech-train-clean-100
T40-p1	replace F0 extractor: DNN predicts F0 from x-vectors and BNs			+							F0: LibriSpeech-dev + VCTK-dev
T32-p1	pitch shifting using time-scale modification (TSM): phase vocoder-based TSM (PV-TSM)										
T32-c1											



Participants' systems: 2020 vs 2022



2020

System	Description	Modified components / data in B1							
		1	2	3	4	5	6	7	Data
A2	B1: x-vector anonymization using singular value modification				+			+	Speaker pool: LibriTTS-train-other-500 LibriTTS
A	B1: Different F0 extractors; x-vector anonymization using statistical-based ensemble regression modeling	+			+			+	
O1	B1: x-vector anonymization keeping original distribution of cosine distances between speaker x-vectors; GMM for sampling vectors in a PCA-reduced space with the following reconstruction to the fake x-vectors of the original dimension				+			+	Speaker LibriTTS- VoxCeleb
O1c1	O1: with forced dissimilarity between original and generated x-vectors				+			+	
S2	S2c1: applied on the top of the B1 x-vector anonymization				+				
S2c1	B1: x-vector anonymization using domain-adversarial training, autoencoders: using gender, accent, speaker id outputs corresponding to adversarial branches in ANN for x-vector reconstruction				+				
M1	B1: ASR part to extract BN features for SS models (E2E ASR for BNs)		+			+	+		
M1c1	B1: ASR part to extract BN features for SS models (E2E ASR for BNs; semi-adversarial training to learn linguistic features while masking speaker information)		+			+	+		
M1c2	B1: copy-synthesis (original x-vectors)				+				
M1c3	B1: x-vectors provided to SS AM are anonymized, x-vectors provided to NSF are original				+				
M1c4	B1: x-vectors provided to SS AM are original, x-vectors provided to NSF are anonymized				+				
K2	anonymization using x-vectors and SS models: Voice-Indistinguishability metric; a waveform vocoder based on Griffin-Lim algorithm								Speaker test set i
D1	B2: additional modifications in pole radius								
I1	modifications in formants, F0 and speaking rate								

2022

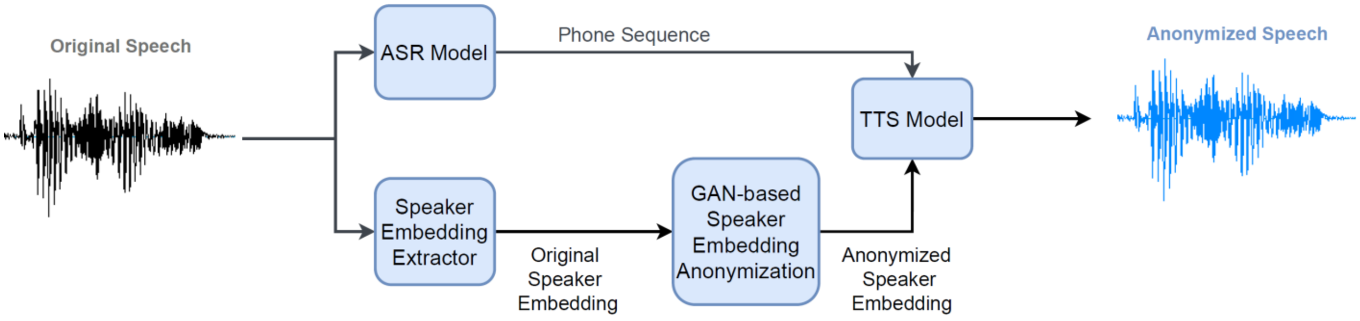
System	Description			Modified components & Data in B1*							
			1	2	3	4	5	6	7	Data
T04-p1	phonetic speech recognition; speaker embedding anonymization via GAN; multi-speaker SS; no usage of original pitch		+	+	+	+	+	+	+	ASR: LibriTTS-train-clean-100, LibriTTS-train-other-500; VoxCeleb-1,2 data (with ASR output transcripts)
T09-p1	replace architecture for all the models + voice/unvoiced features;	gender selection: same	+	+	+	+	+	+	+	
T09-p2		opposite	+	+	+	+	+	+	+	
T09-c1		random	+	+	+	+	+	+	+	
T09-c2		same	+	+	+	+	+	+	+	Speaker pool: LibriTTS-train-other-500 + VoxCeleb-1,2
T09-c3		opposite	+	+	+	+	+	+	+	
T09-c4		random	+	+	+	+	+	+	+	
T11-p1	replace x-vectors by speaker ids from a look-up table + speaker encoder; replaced architecture for all the models		+	+	+	+	+	+	+	
T11-p2			+	+	+	+	+	+	+	
T11-p3			+	+	+	+	+	+	+	
T11-p4			+	+	+	+	+	+	+	
T18-p1	adding adversarial noise to x-vectors					+				
T18-c1	replace x-vectors by ASR embeddings				+					ASRspk: LibriSpeech-train-clean-100
T40-p1	replace F0 extractor: DNN predicts F0 from x-vectors and BNs		+							F0: LibriSpeech-dev + VCTK-dev
T32-p1	pitch shifting using time-scale modification (TSM):									
T32-c1	phase vocoder-based TSM (PV-TSM)									

- **2020**: focus on x-vector anonymization
- **2022**: modifications of all components

Participants' systems T04

System	Description		
T04-p1	phonetic speech recognition; speaker embedding anonymization via GAN; multi-speaker SS; no usage of original pitch		
T09-p1	replace architecture for all the models + voice/unvoiced features;	gender selection	same
T09-p2			opposite
T09-c1			random
T09-c2			same
T09-c3			opposite
T09-c4			random
T11-p1	replace x-vectors by speaker ids from a look-up table + speaker encoder; replaced architecture for all the models		
T11-p2			
T11-p3			
T11-p4			
T18-p1	adding adversarial noise to x-vectors		
T18-c1	replace x-vectors by ASR embeddings		
T40-p1	replace F0 extractor: DNN predicts F0 from x-vectors and BNs		
T32-p1	pitch shifting using time-scale modificatic phase vocoder-based TSM (PV-TSM)		
T32-c1			

[Meyer 2022]



- Phonetic ASR transcriptions
- Speaker embedding anonymization via GAN
- No usage of original pitch (pitch estimation: FastSpeech2 & FastPitch)
- Multi-speaker TTS

Participants' systems T09

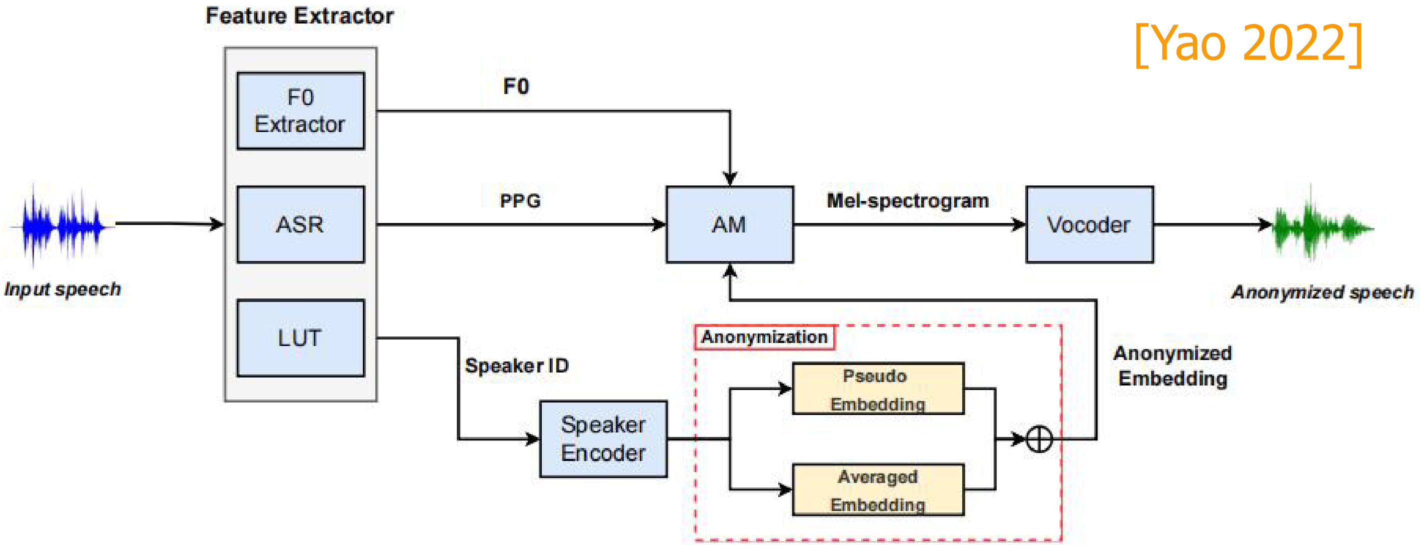
System	Description		
T04-p1	phonetic speech recognition; speaker embedding anonymization via GAN; multi-speaker SS; no usage of original pitch		
T09-p1	replace architecture for all the models + voice/unvoiced features;	gender selection	same
T09-p2			opposite
T09-c1			random
T09-c2			same
T09-c3			opposite
T09-c4			random
T11-p1	replace x-vectors by speaker ids from a look-up table + speaker encoder; replaced architecture for all the models		
T11-p2			
T11-p3			
T11-p4			
T18-p1	adding adversarial noise to x-vectors		
T18-c1	replace x-vectors by ASR embeddings		
T40-p1	replace F0 extractor: DNN predicts F0 from x-vectors and BNs		
T32-p1	pitch shifting using time-scale modificatic phase vocoder-based TSM (PV-TSM)		
T32-c1			

- Replace architecture for all the models (ResNet-34-based x-vector extractor; end-to-end hybrid CTC-attention BN feature extractor; PyWorld toolkit to extract F0;….)
- Voice/unvoiced feature
- 3 gender selection strategies for x-vector anonymization: same, opposite, random

Participants' systems T11

System	Description		
T04-p1	phonetic speech recognition; speaker embedding anonymization via GAN; multi-speaker SS; no usage of original pitch		
T09-p1	replace architecture for all the models + voice/unvoiced features;	gender selection	same
T09-p2			opposite
T09-c1			random
T09-c2			same
T09-c3			opposite
T09-c4			random
T11-p1	replace x-vectors by speaker ids from a look-up table + speaker encoder; replaced architecture for all the models		
T11-p2			
T11-p3			
T11-p4			
T18-p1	adding adversarial noise to x-vectors		
T18-c1	replace x-vectors by ASR embeddings		
T40-p1	replace F0 extractor: DNN predicts F0 from x-vectors and BNs		
T32-p1	pitch shifting using time-scale modificatic phase vocoder-based TSM (PV-TSM)		
T32-c1			

[Yao 2022]



ASV-model-free approach for speaker anonymization:

- Look-up-table (LUT) for speakers in training set as speaker pool
- Reserve a pseudo speaker ID in LUT to generate pseudo speaker embedding
- anonymized embedding: pseudo-speaker embedding + averaged embedding of randomly selected speaker embeddings in LUT

$$\text{anonymized embedding} = \alpha * \text{averaged embedding} \oplus \beta * \text{pseudo embedding}$$

Participants' systems T18

System	Description		
T04-p1	phonetic speech recognition; speaker embedding anonymization via GAN; multi-speaker SS; no usage of original pitch		
T09-p1	replace architecture for all the models + voice/unvoiced features;	gender selection	same
T09-p2			opposite
T09-c1			random
T09-c2			same
T09-c3			opposite
T09-c4			random
T11-p1	replace x-vectors by speaker ids from a look-up table + speaker encoder; replaced architecture for all the models		
T11-p2			
T11-p3			
T11-p4			
T18-p1	adding adversarial noise to x-vectors		
T18-c1	replace x-vectors by ASR embeddings		
T40-p1	replace F0 extractor: DNN predicts F0 from x-vectors and BNs		
T32-p1	pitch shifting using time-scale modificatic phase vocoder-based TSM (PV-TSM)		
T32-c1			

[Chen 2022]

T18-p1: Adding adversarial noise to x-vectors

$$Y_i = X_i + noise_{adv}$$

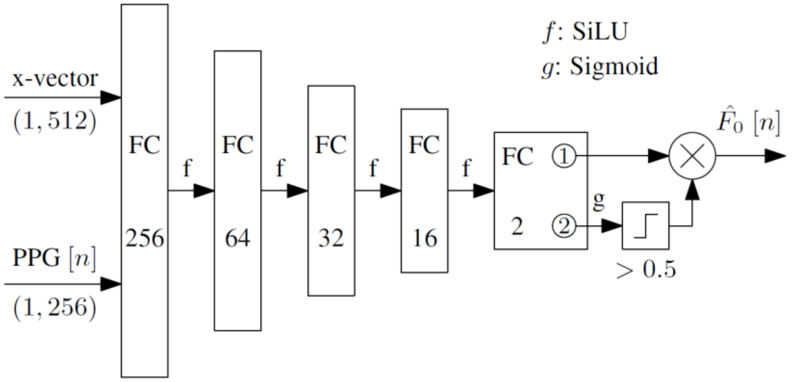
T18-c1: Replace x-vectors by embeddings extracted from a transformer-based ASR

Participants’ systems T40

System	Description		
T04-p1	phonetic speech recognition; speaker embedding anonymization via GAN; multi-speaker SS; no usage of original pitch		
T09-p1	replace architecture for all the models + voice/unvoiced features;	gender selection	same
T09-p2			opposite
T09-c1			random
T09-c2			same
T09-c3			opposite
T09-c4			random
T11-p1	replace x-vectors by speaker ids from a look-up table + speaker encoder; replaced architecture for all the models		
T11-p2			
T11-p3			
T11-p4			
T18-p1	adding adversarial noise to x-vectors		
T18-c1	replace x-vectors by ASR embeddings		
T40-p1	replace F0 extractor: DNN predicts F0 from x-vectors and BNs		
T32-p1	pitch shifting using time-scale modificatic phase vocoder-based TSM (PV-TSM)		
T32-c1			

[Gaznepoglu 2022]

Estimate F0 from BN and anonymized x-vector



$$\mathcal{L}(F_0, \hat{F}_0) = \mathrm{MSE}(F_0 - \hat{F}_0)^2 + \alpha \mathrm{BCE}(p_v, v)$$

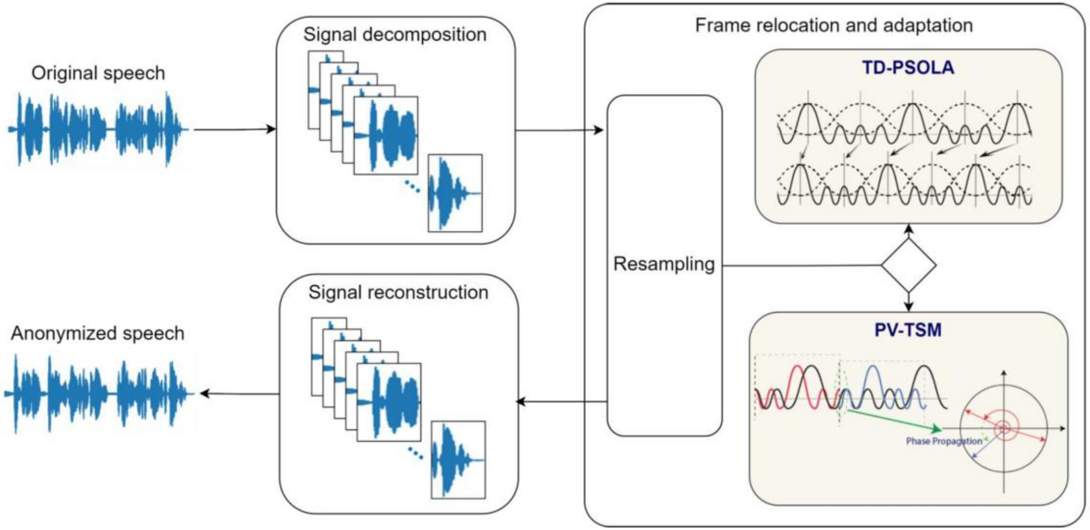
Participants' systems T32

System	Description		
T04-p1	phonetic speech recognition; speaker embedding anonymization via GAN; multi-speaker SS; no usage of original pitch		
T09-p1	replace architecture for all the models + voice/unvoiced features;	gender selection	same
T09-p2			opposite
T09-c1			random
T09-c2			same
T09-c3			opposite
T09-c4			random
T11-p1	replace x-vectors by speaker ids from a look-up table + speaker encoder; replaced architecture for all the models		
T11-p2			
T11-p3			
T11-p4			
T18-p1	adding adversarial noise to x-vectors		
T18-c1	replace x-vectors by ASR embeddings		
T40-p1	replace F0 extractor: DNN predicts F0 from x-vectors and BNs		
T32-p1	pitch shifting using time-scale modificati phase vocoder-based TSM (PV-TSM)		
T32-c1			

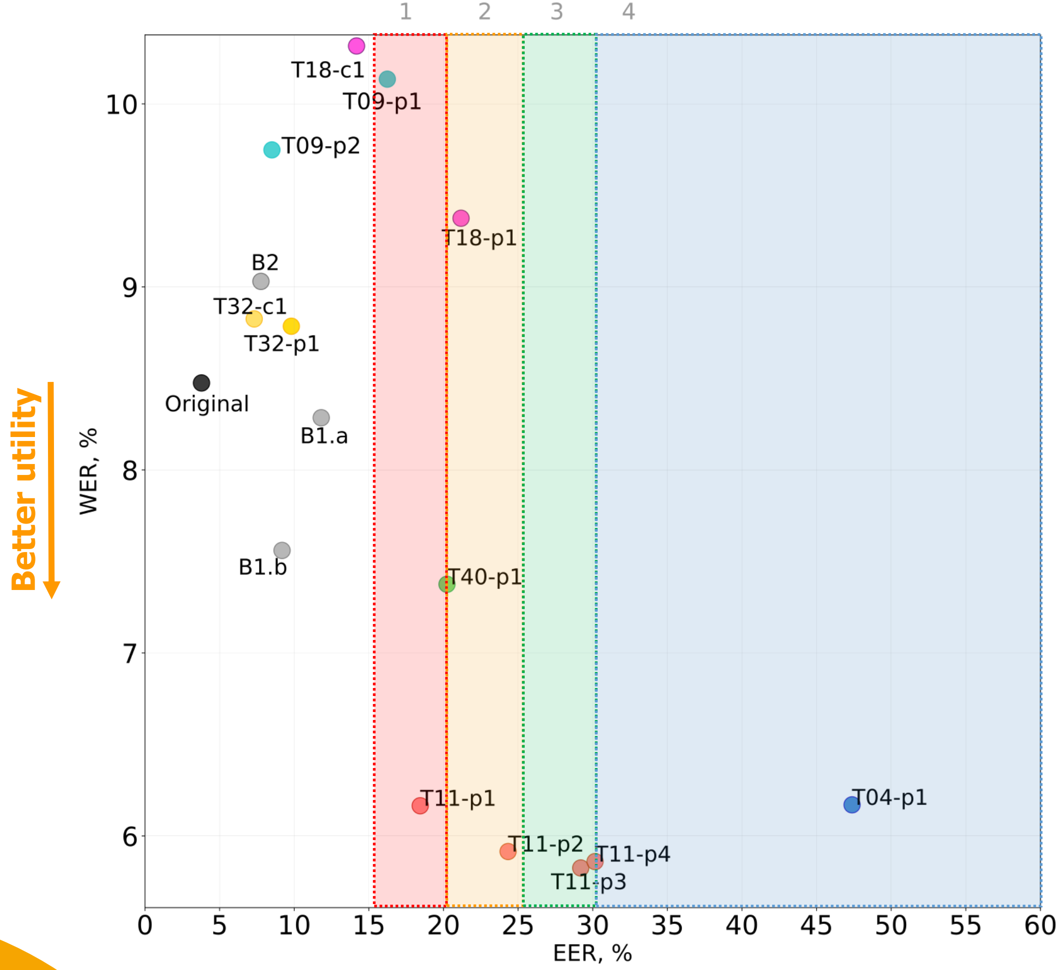
[Mawalim 2022]

Pitch shifting using time-scale modification (TSM):

- phase vocoder-based TSM (PV-TSM)
- time-domain pitch synchronous overlap-add (TD-PSOLA)



Objective evaluation results: EER vs WER



Results on test data

4 privacy protection conditions:

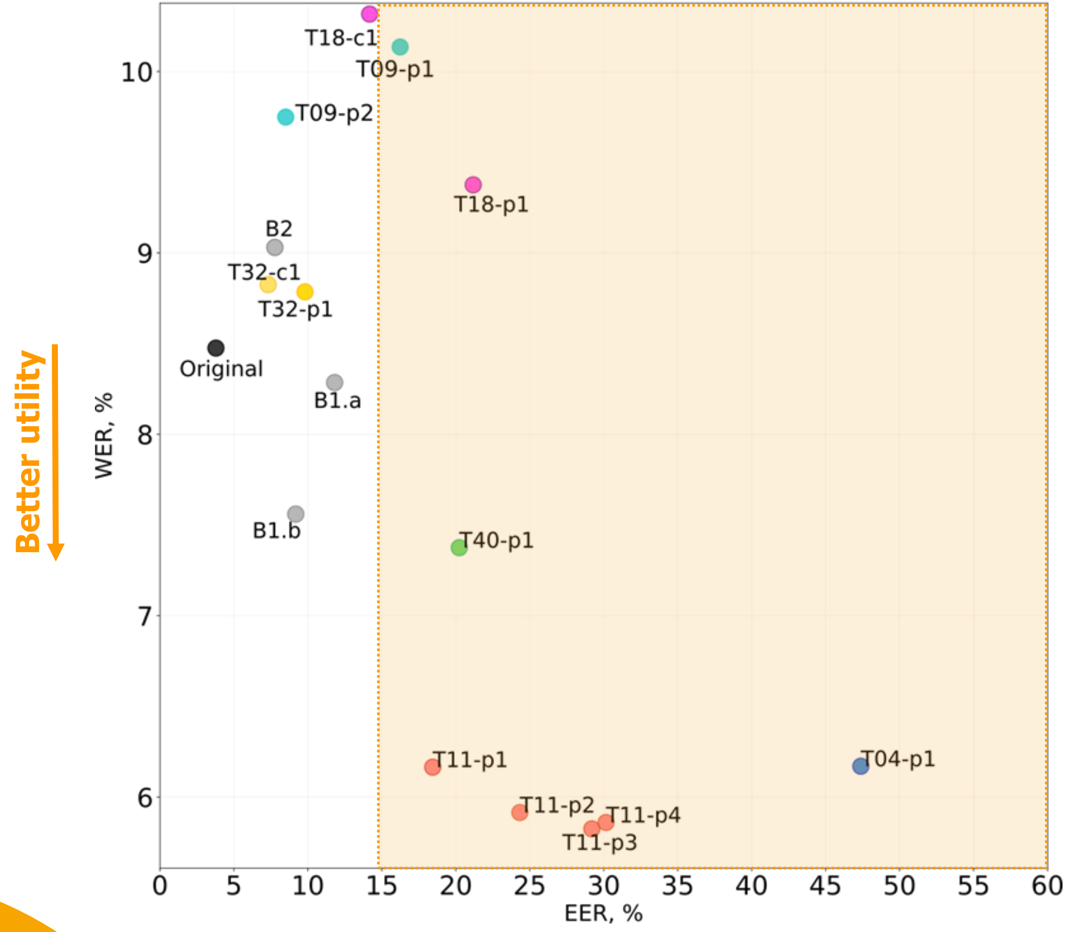
1. EER ≥ 15%
2. EER ≥ 20%
3. EER ≥ 25%
4. EER ≥ 30%

For every condition, rank system by WER

Objective evaluation results: EER vs WER

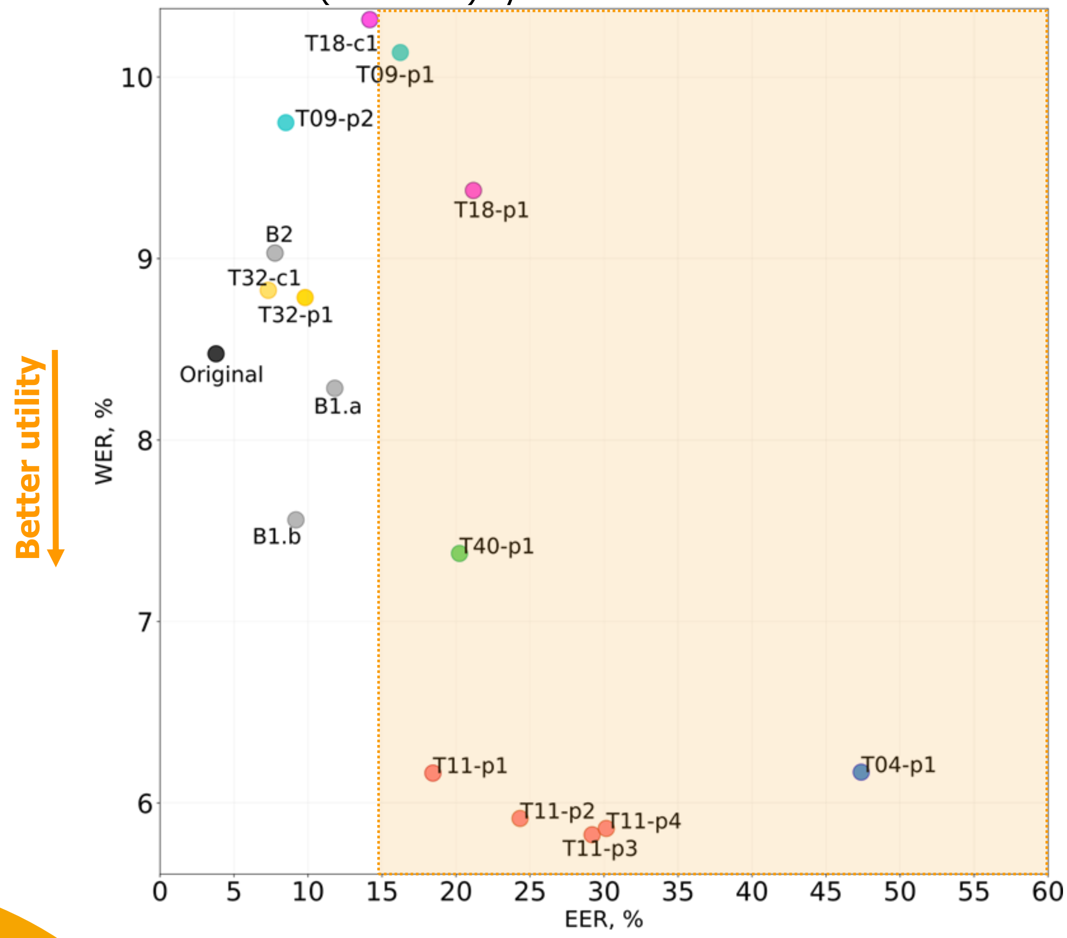
Results on test data: condition **1**: **EER ≥ 15%**

Choose one (best WER) system for each team for this condition

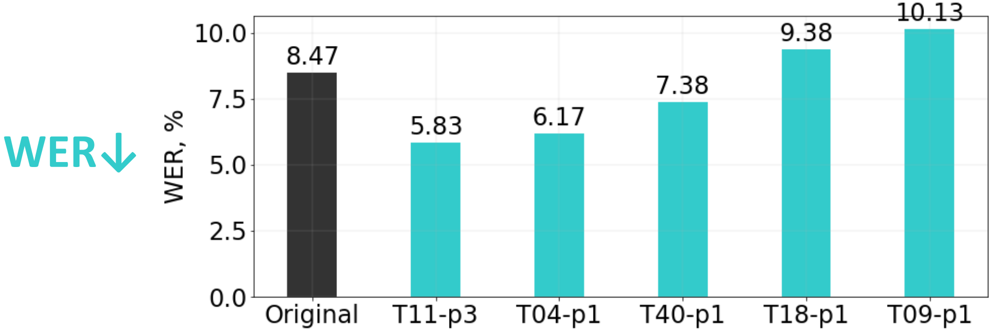


Objective evaluation results: EER vs WER

Choose one (best WER) system for each team for this condition

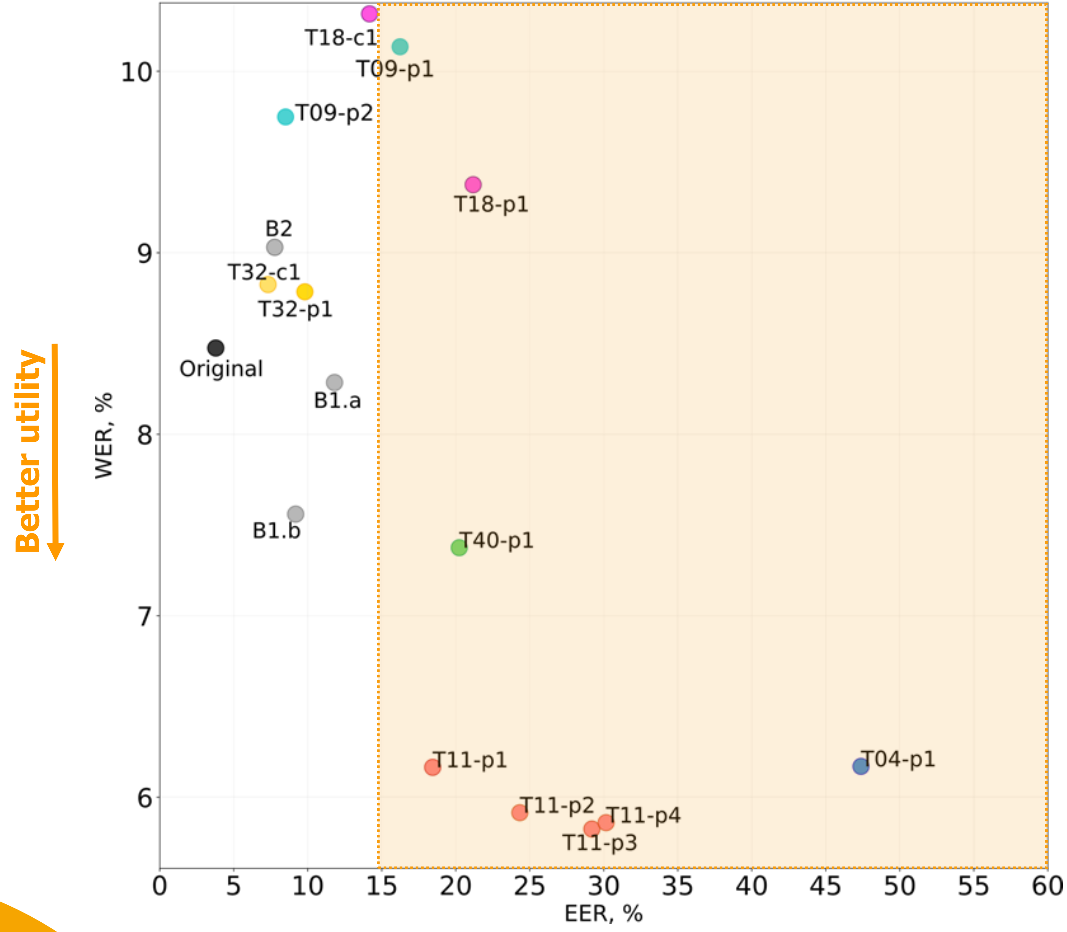


Results on test data: condition **1**: **EER ≥ 15%**

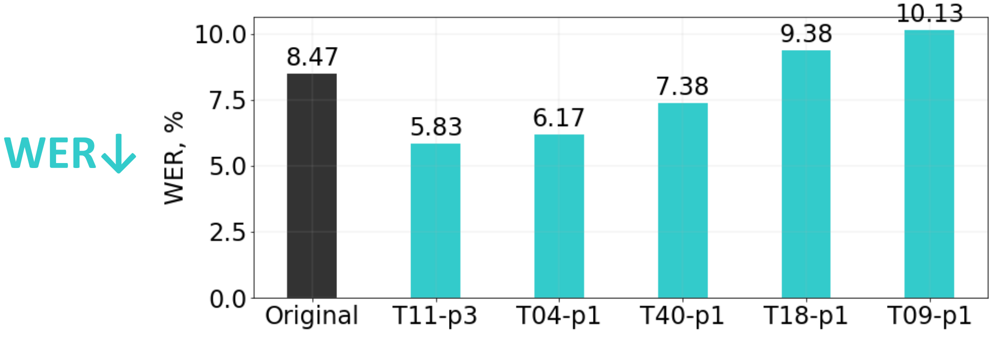


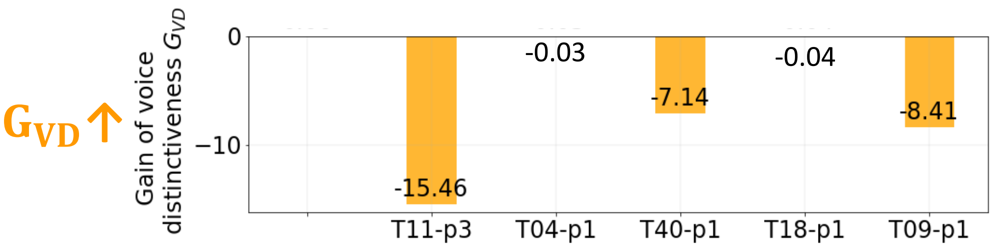
Objective evaluation results: EER vs WER

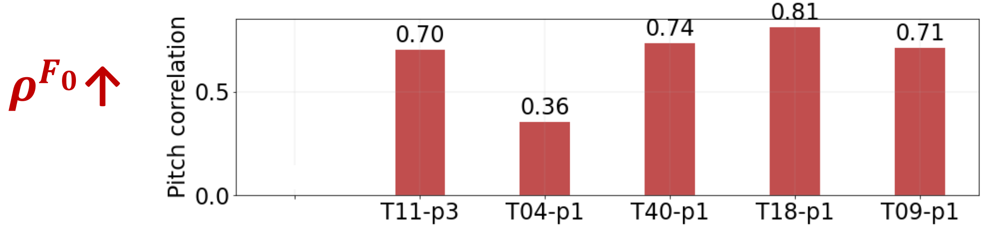
Choose one (best WER) system for each team for this condition



Results on test data: condition **1**: **EER ≥ 15%**

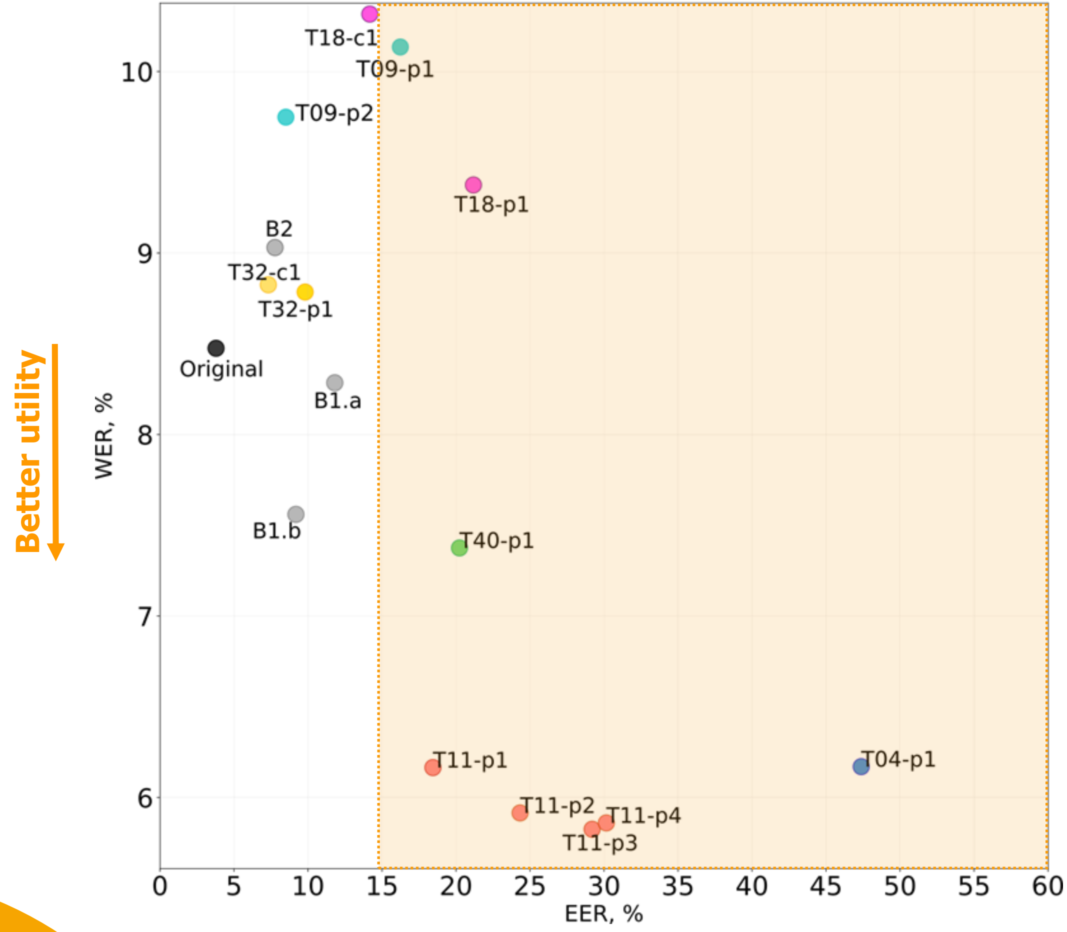




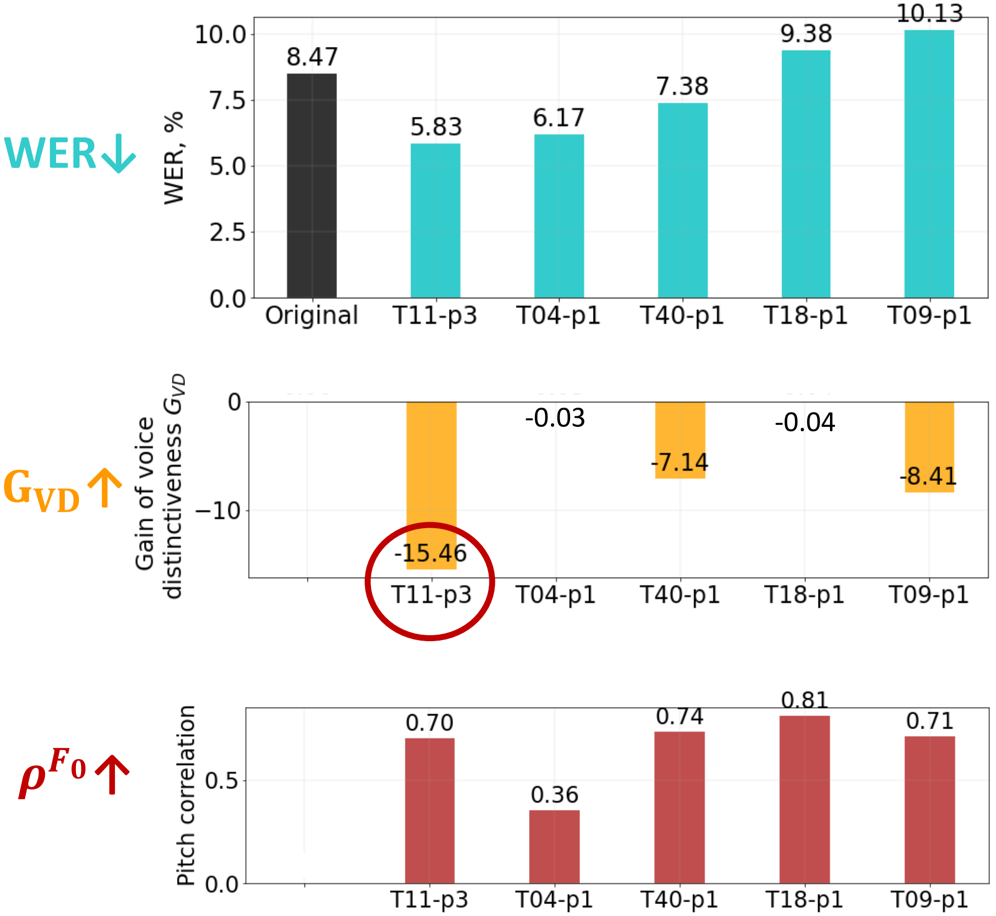


Objective evaluation results: EER vs WER

Choose one (best WER) system for each team for this condition

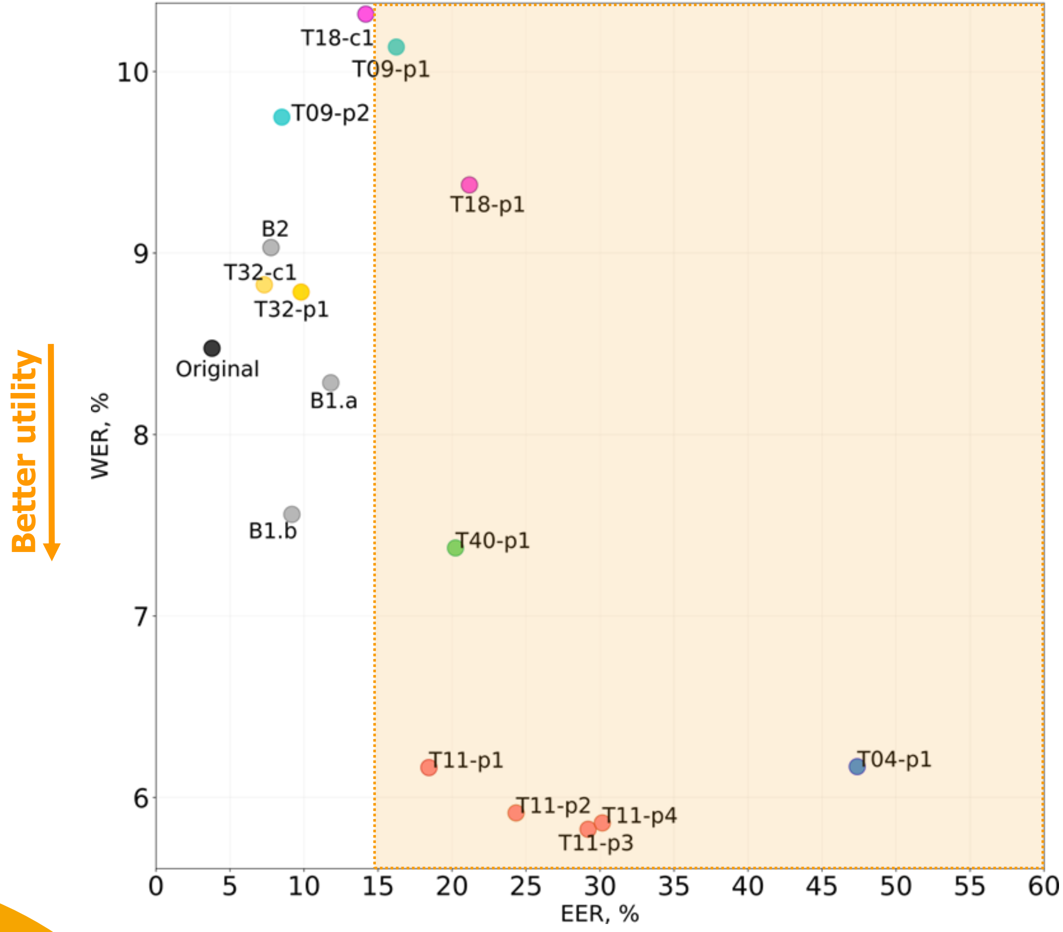


Results on test data: condition **1**: EER ≥ 15%



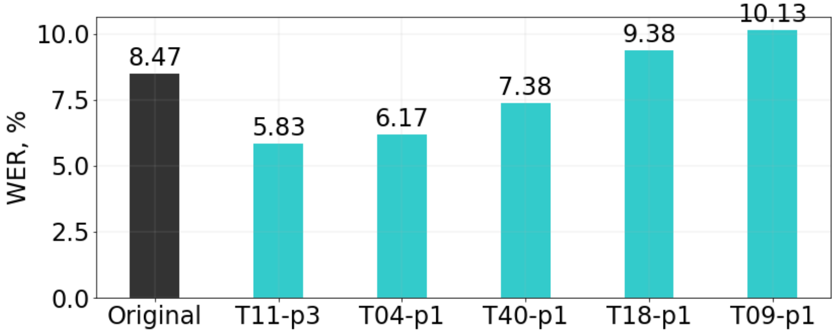
Objective evaluation results: EER vs WER

Choose one (best WER) system for each team for this condition

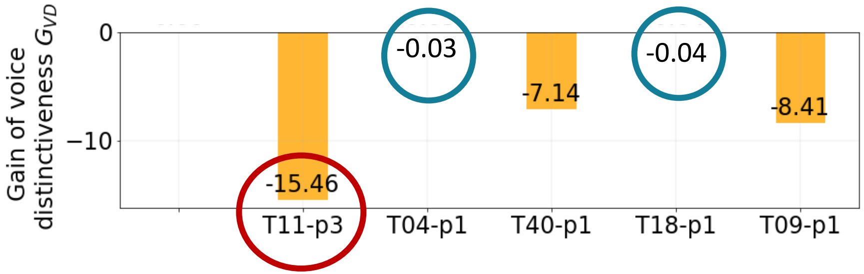


Results on test data: condition **1**: **EER ≥ 15%**

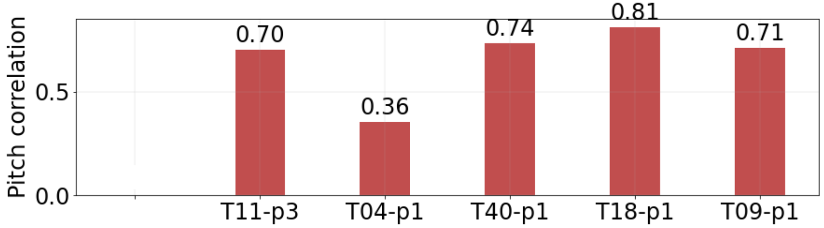
WER↓



G_{VD}↑

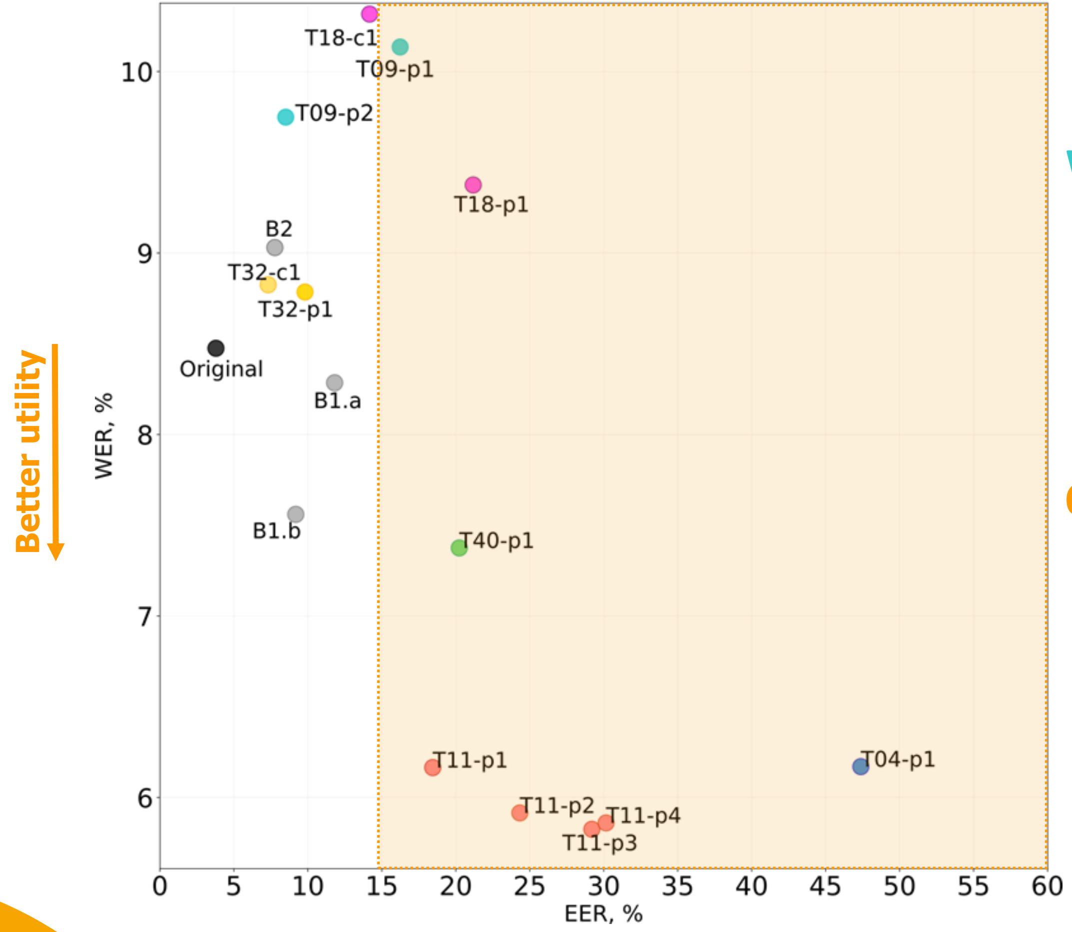


ρ^{F_0}↑

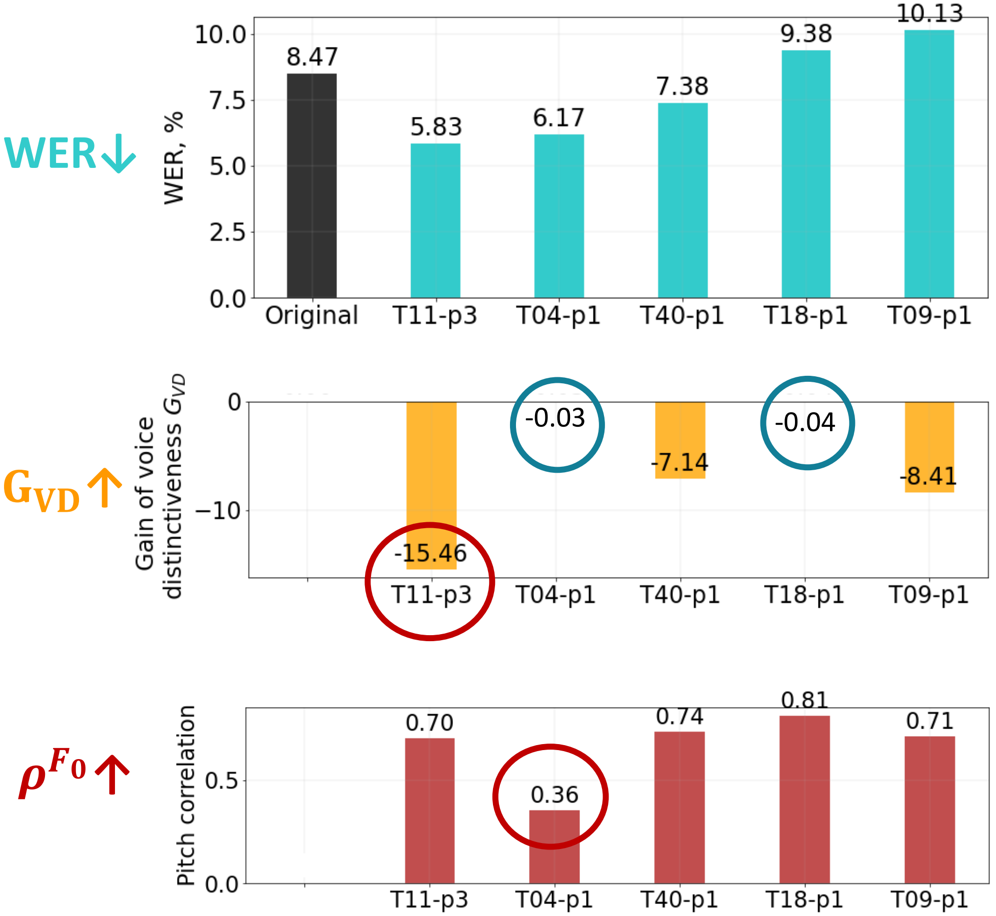


Objective evaluation results: EER vs WER

Choose one (best WER) system for each team for this condition

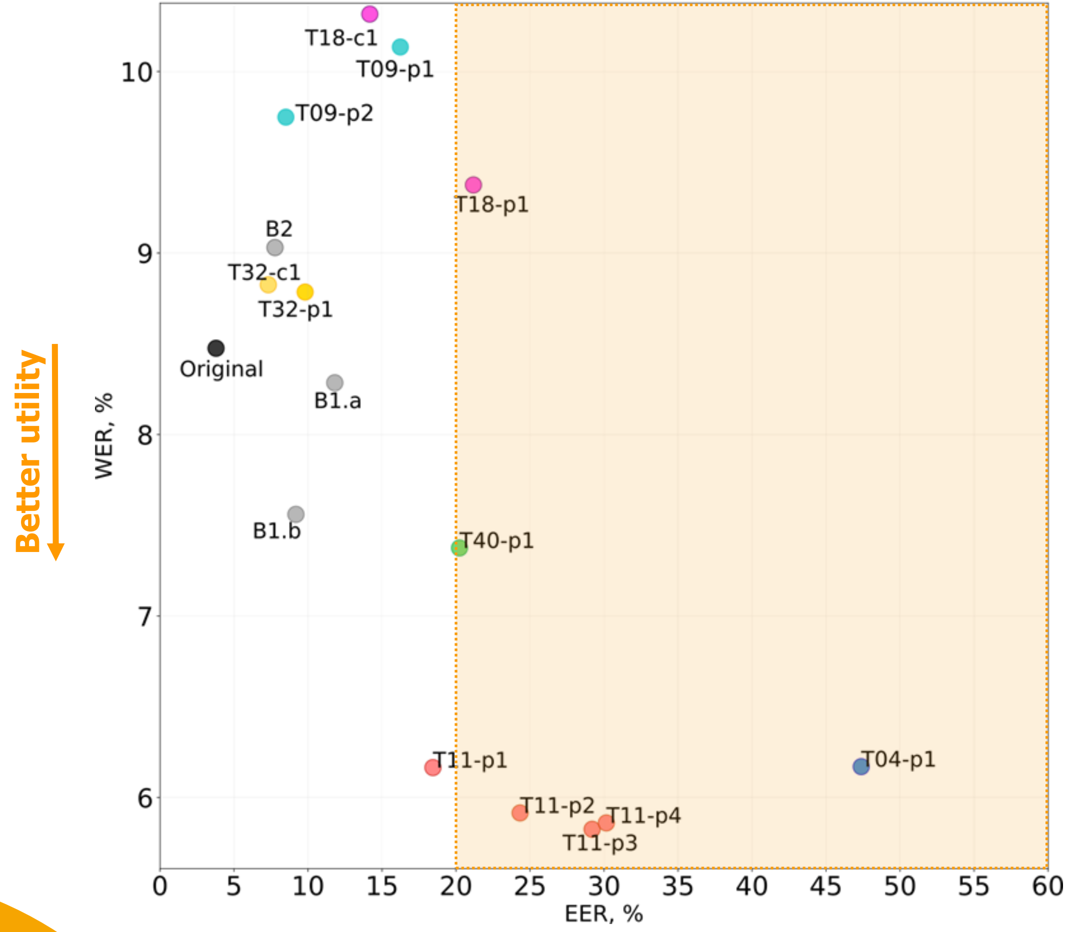


Results on test data: condition **1**: **EER ≥ 15%**

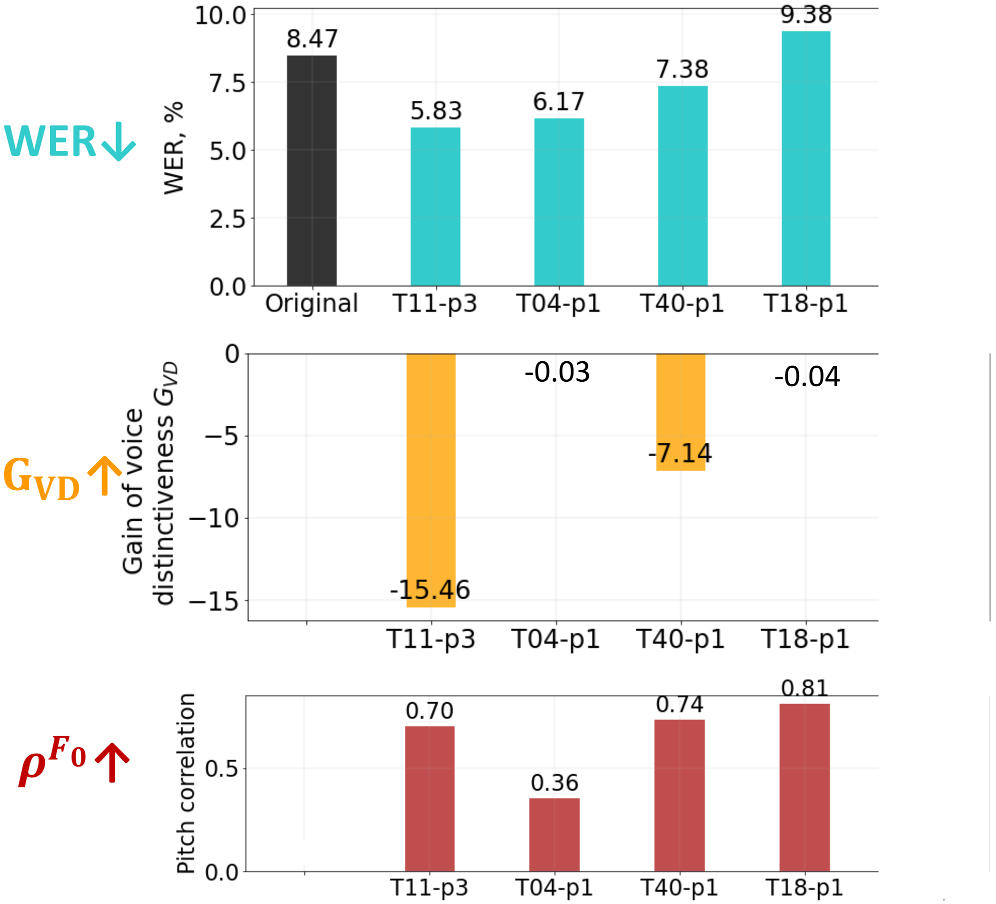


Objective evaluation results: EER vs WER

Choose one (best WER) system for each team for this condition

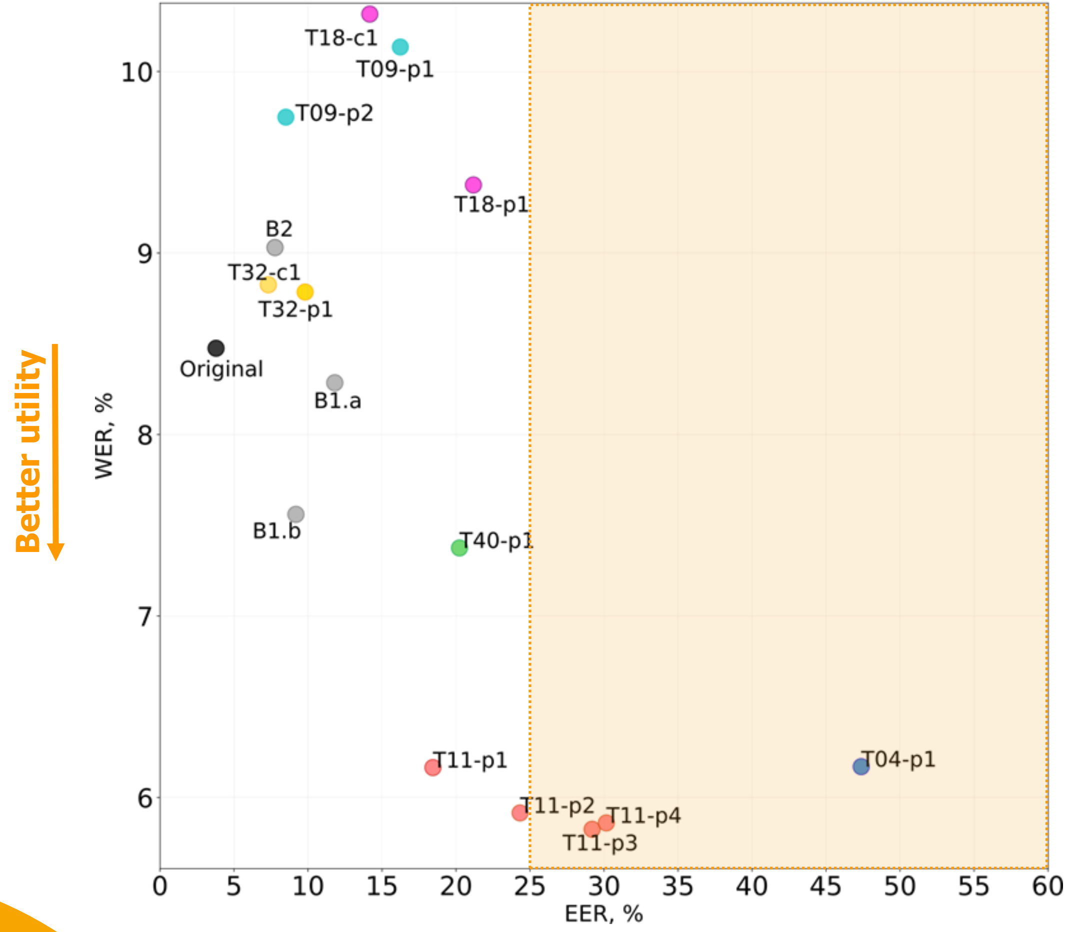


Results on test data: condition **2**: EER ≥ 20%

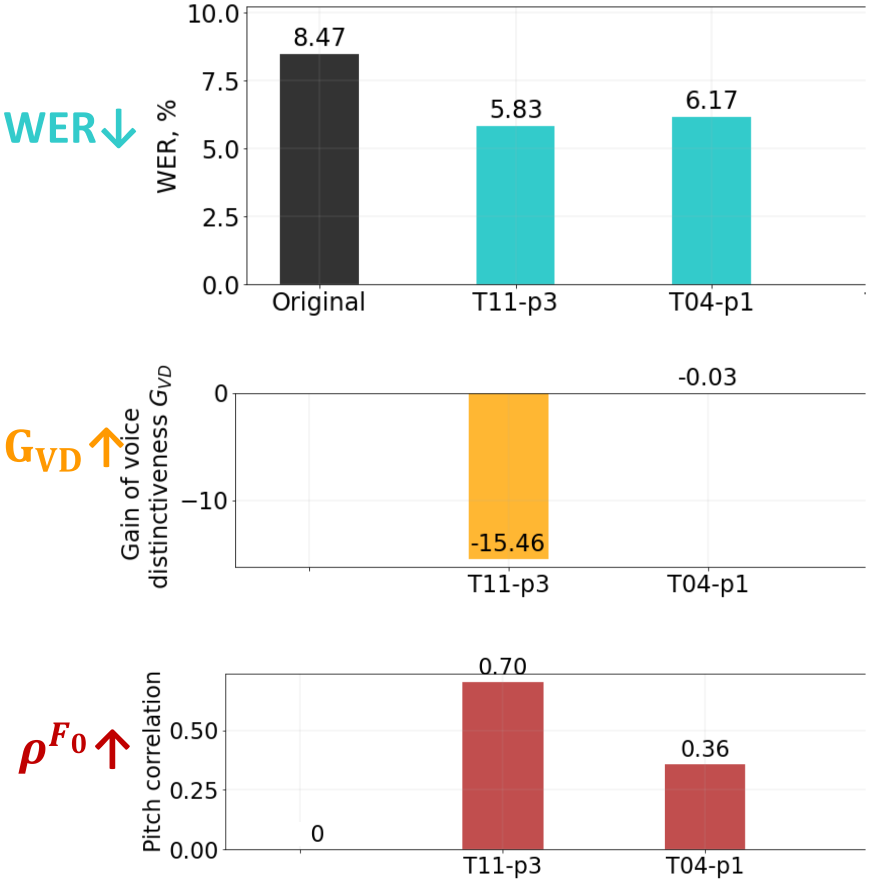


Objective evaluation results: EER vs WER

Choose one (best WER) system for each team for this condition

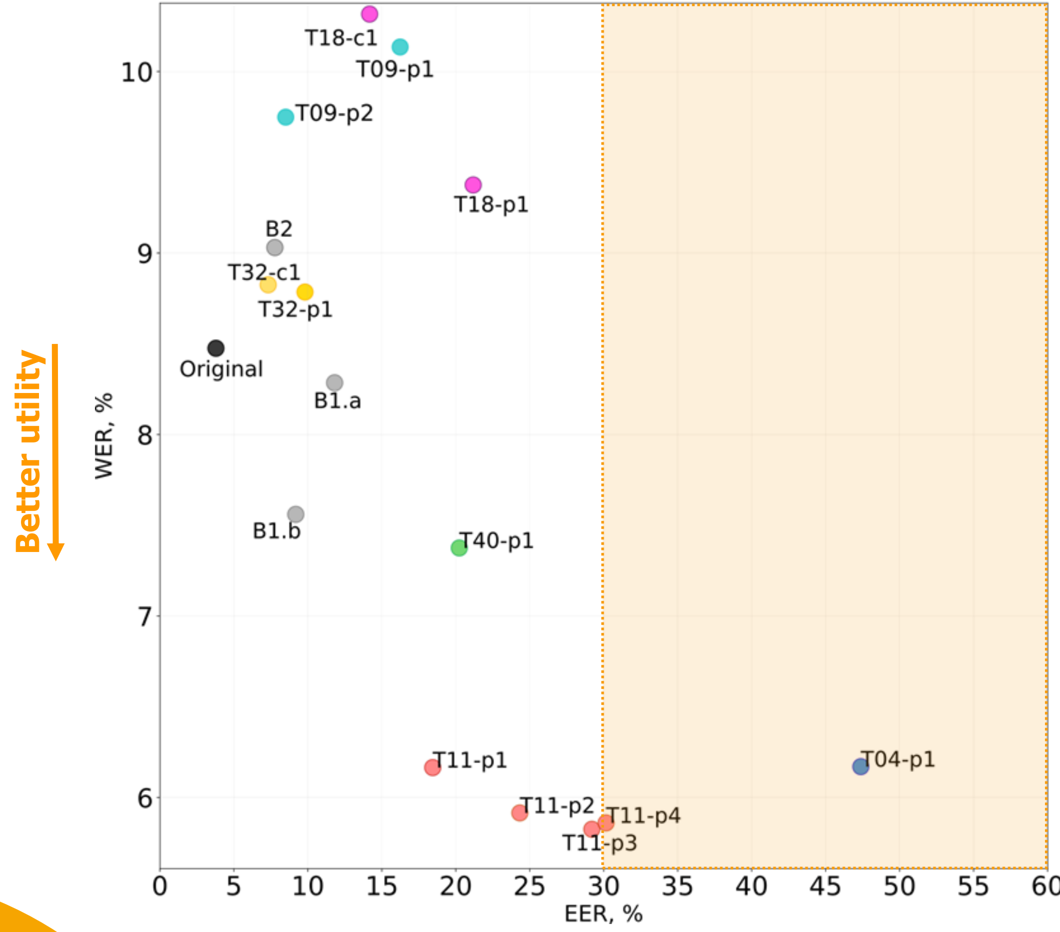


Results on test data: condition **3**: EER ≥ 25%

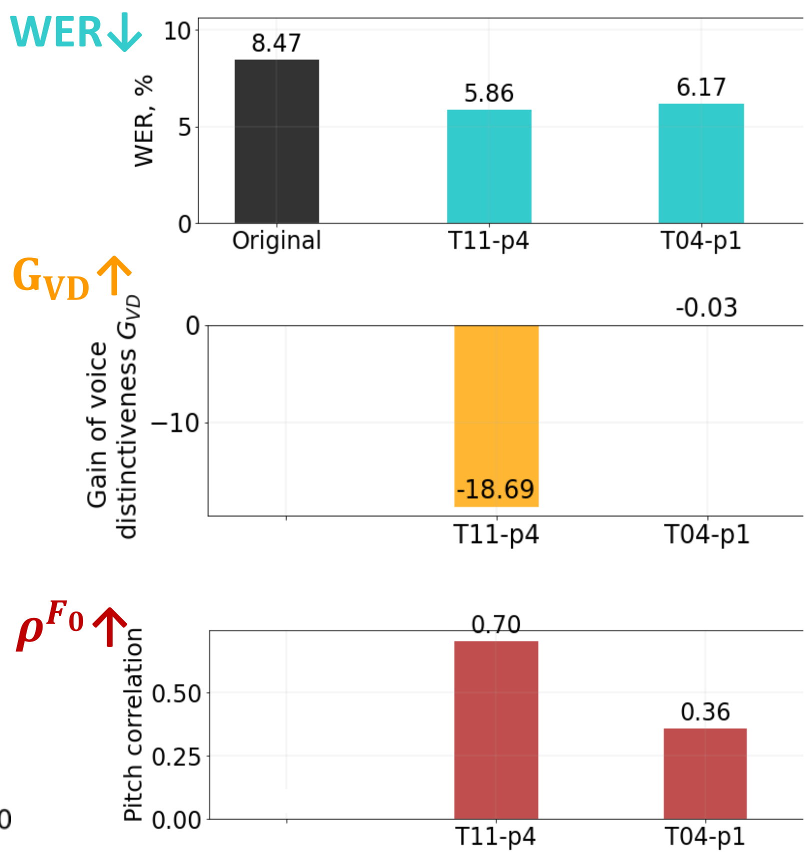


Objective evaluation results: EER vs WER

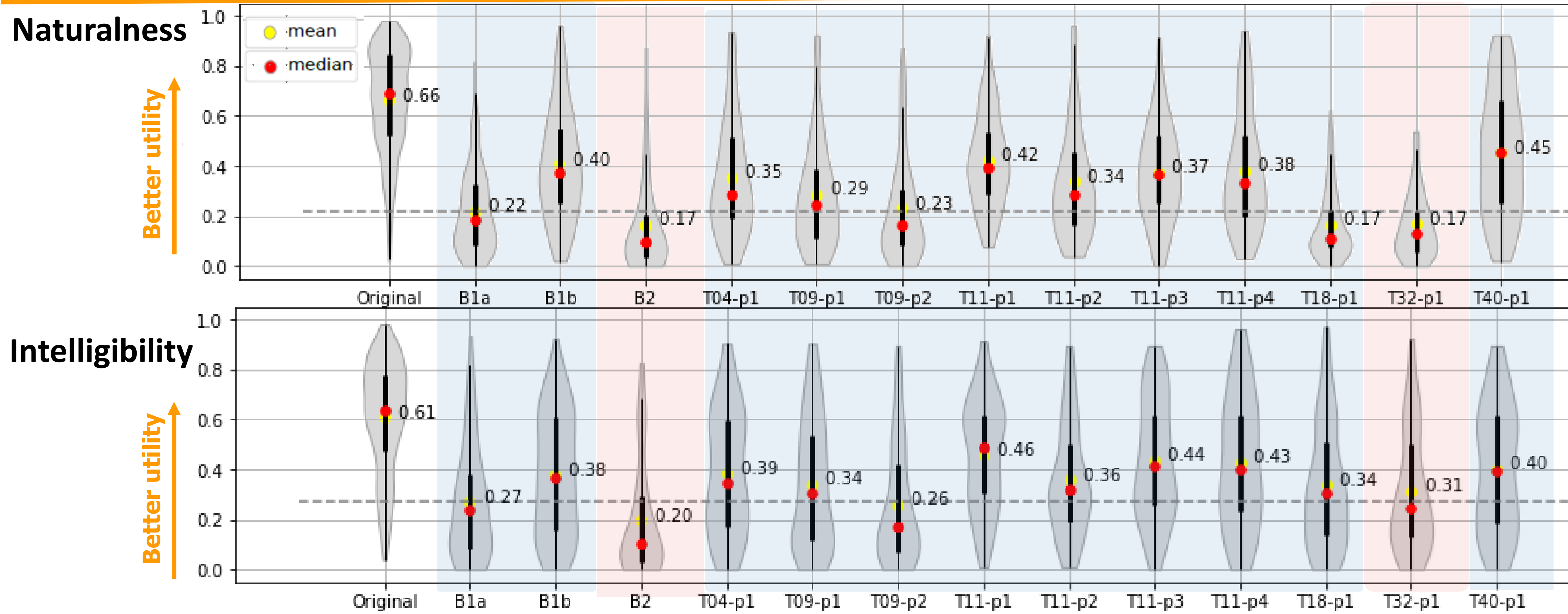
Choose one (best WER) system for each team for this condition



Results on test data: condition **4**: EER ≥ 30%



Subjective evaluation results: utility

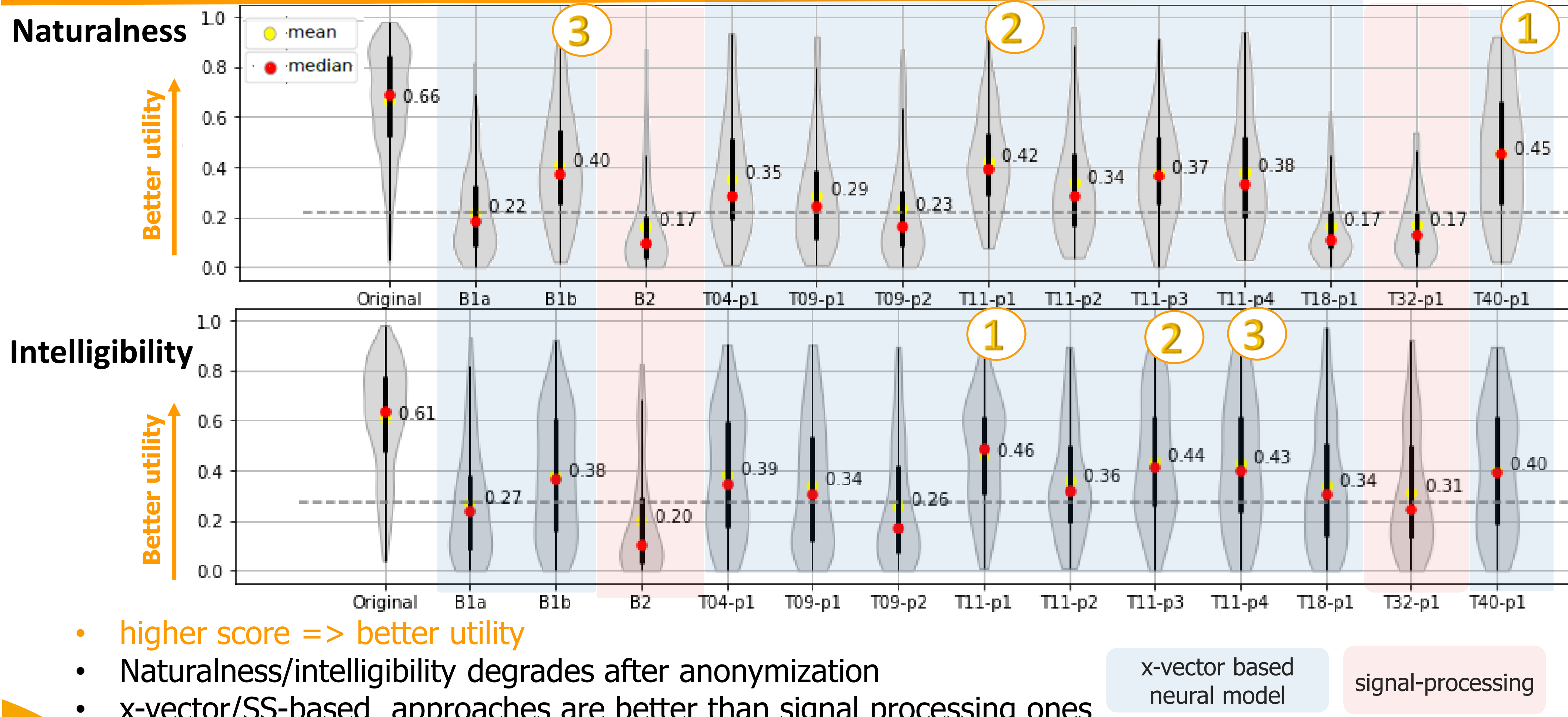


- higher score => better utility
- Naturalness/intelligibility degrades after anonymization
- x-vector/SS-based approaches are better than signal processing ones

x-vector based neural model

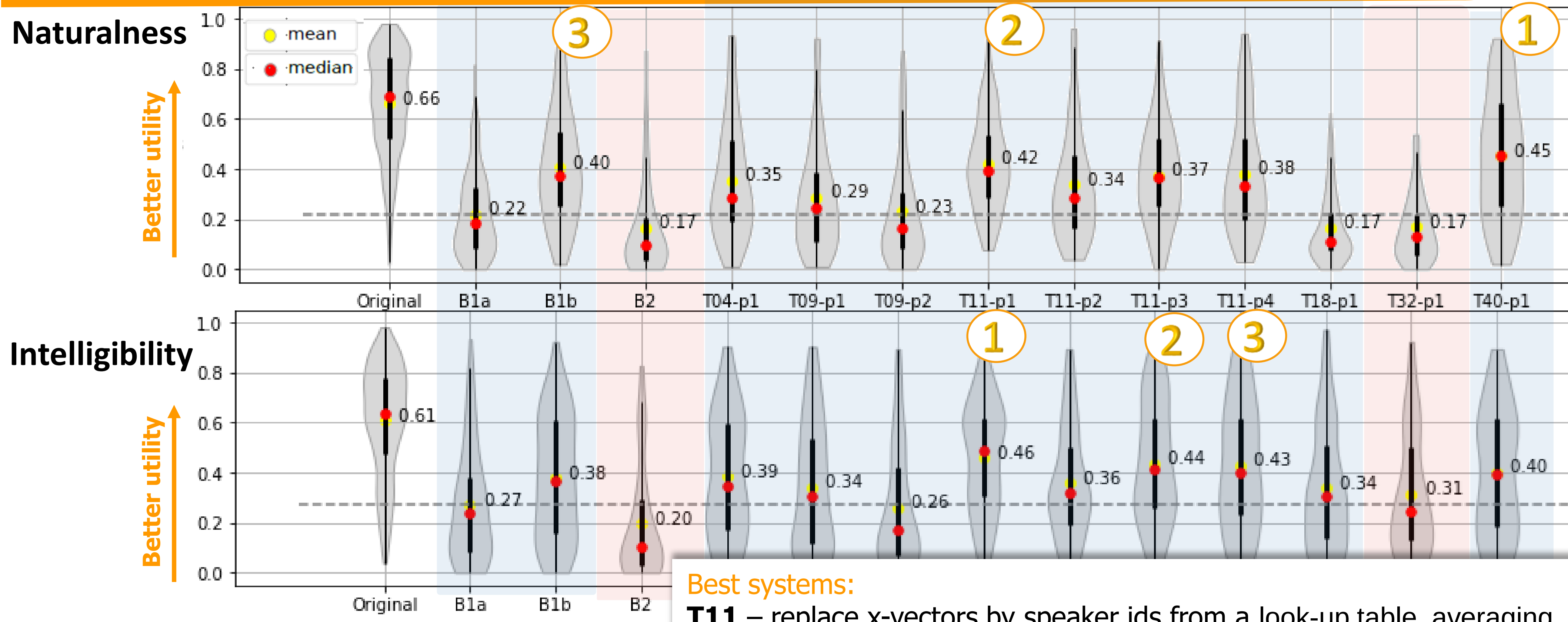
signal-processing

Subjective evaluation results: utility



- higher score => better utility
- Naturalness/intelligibility degrades after anonymization
- x-vector/SS-based approaches are better than signal processing ones

Subjective evaluation results: utility



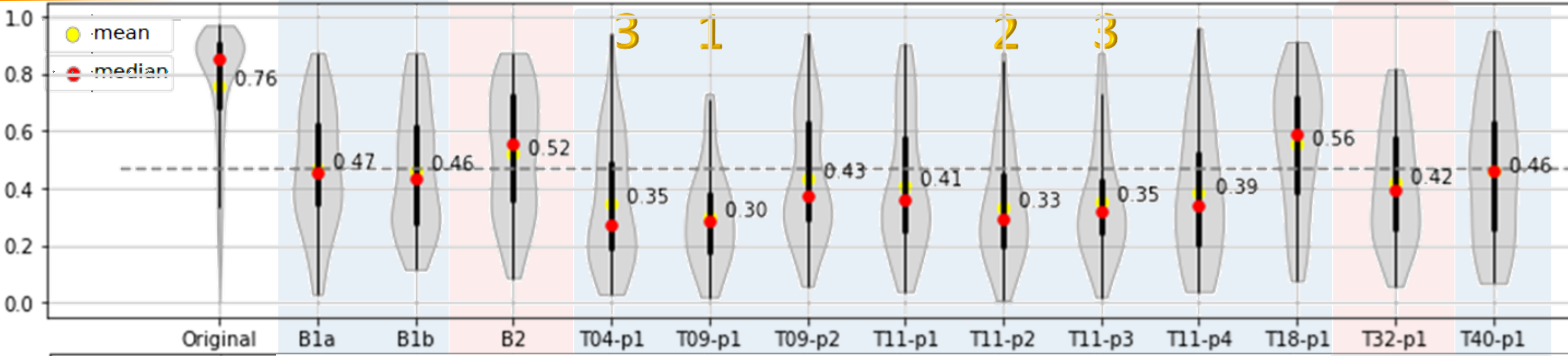
- higher score => better utility

Best systems:
T11 – replace x-vectors by speaker ids from a look-up table, averaging (low voice distinctiveness)
T40-p1 – DNN to predict F0 from x-vectors and BNs
B1.b

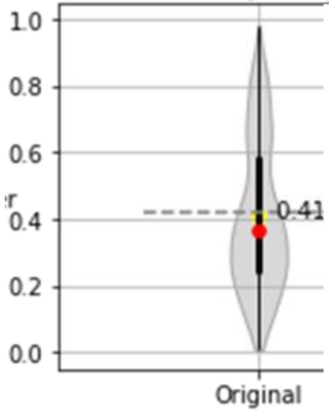
Subjective evaluation results: privacy

Speaker similarity: same speaker

Better privacy



Speaker similarity: different speakers



- lower score => better privacy

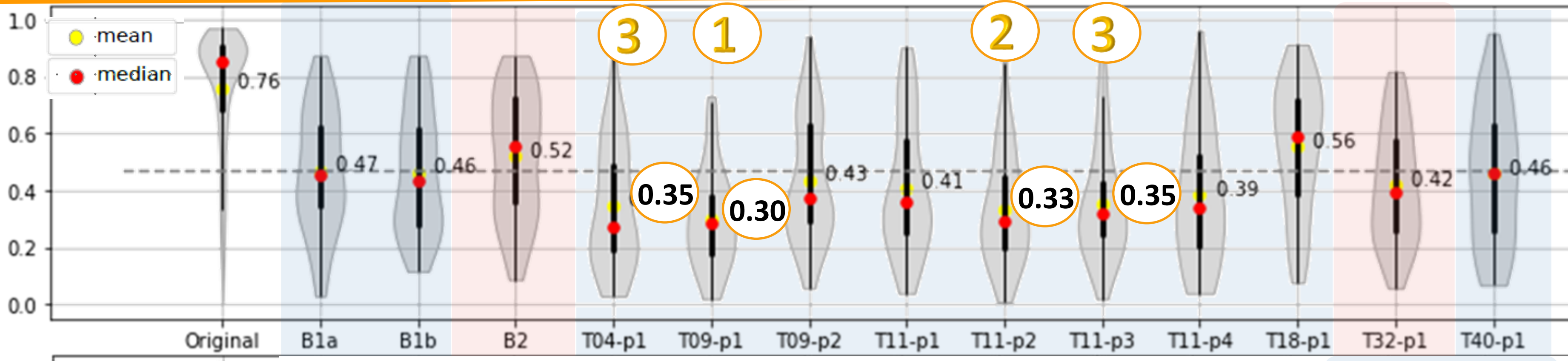
x-vector based neural model

signal-processing

Subjective evaluation results: privacy

Speaker similarity: same speaker

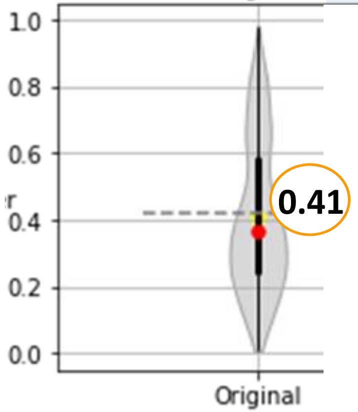
Better privacy



x-vector based neural model

signal-processing

Speaker similarity: different speakers

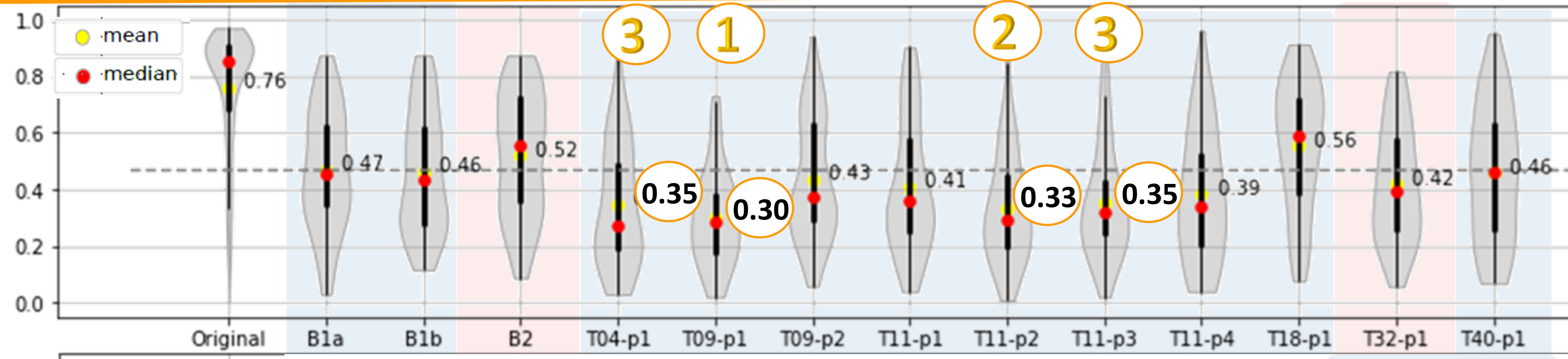


- lower score => better privacy
- Good degree of anonymization, especially for the best systems – lower scores of {original enroll, anonymized trail} comparison of the same speaker than for {original enroll, original trail} for different speakers

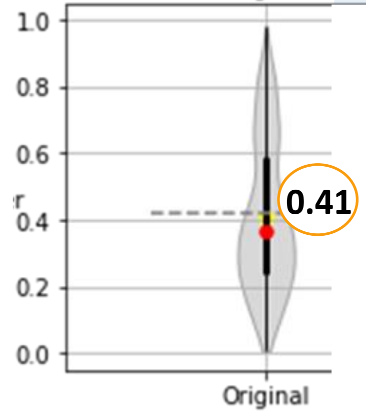
Subjective evaluation results: privacy

Speaker similarity: same speaker

Better privacy



Speaker similarity: different speakers



- lower score => better privacy
- Good degree of anonymization, especially for the best systems – lower scores of {original enroll, anonymized trail} comparison of the same speaker than for {original enroll, original trail} for different speakers

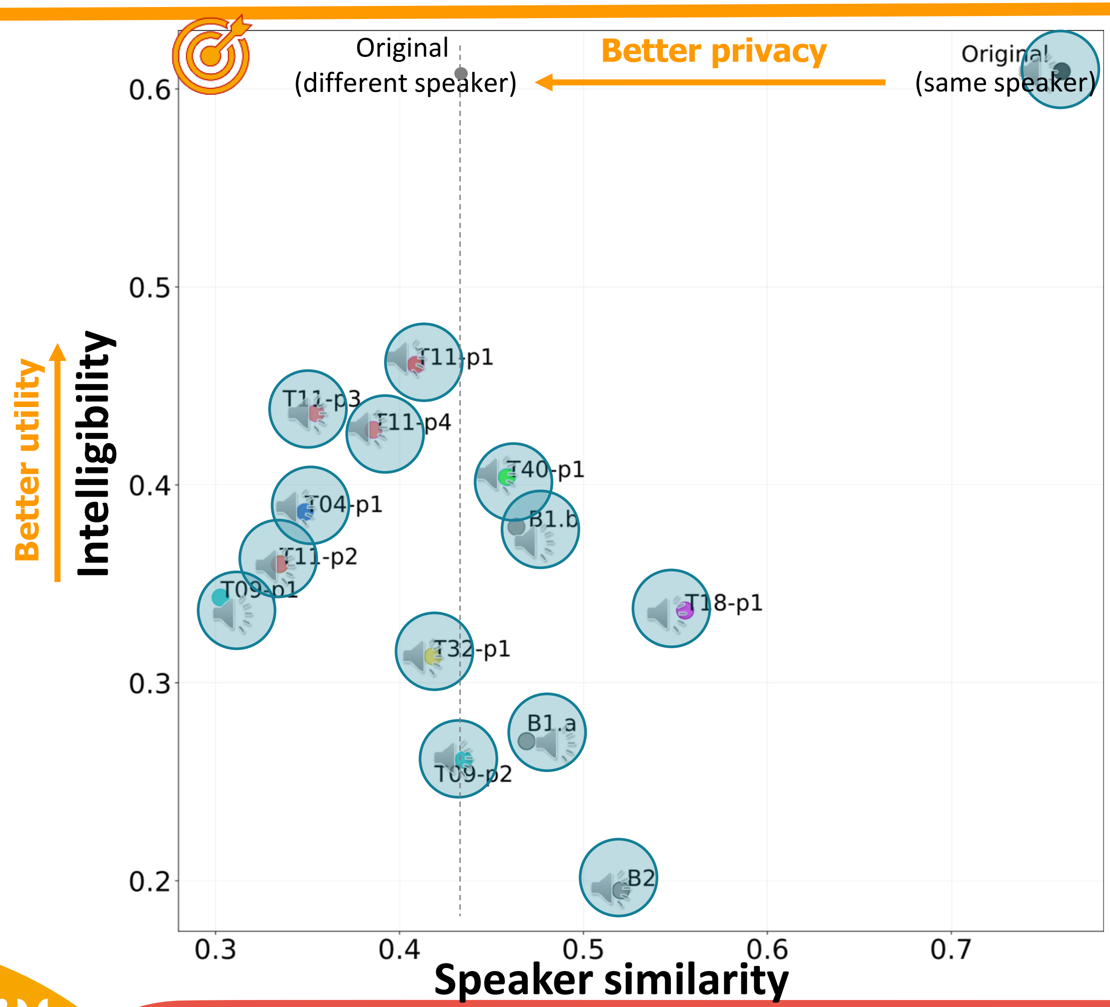
x-vector based neural model

signal-processing

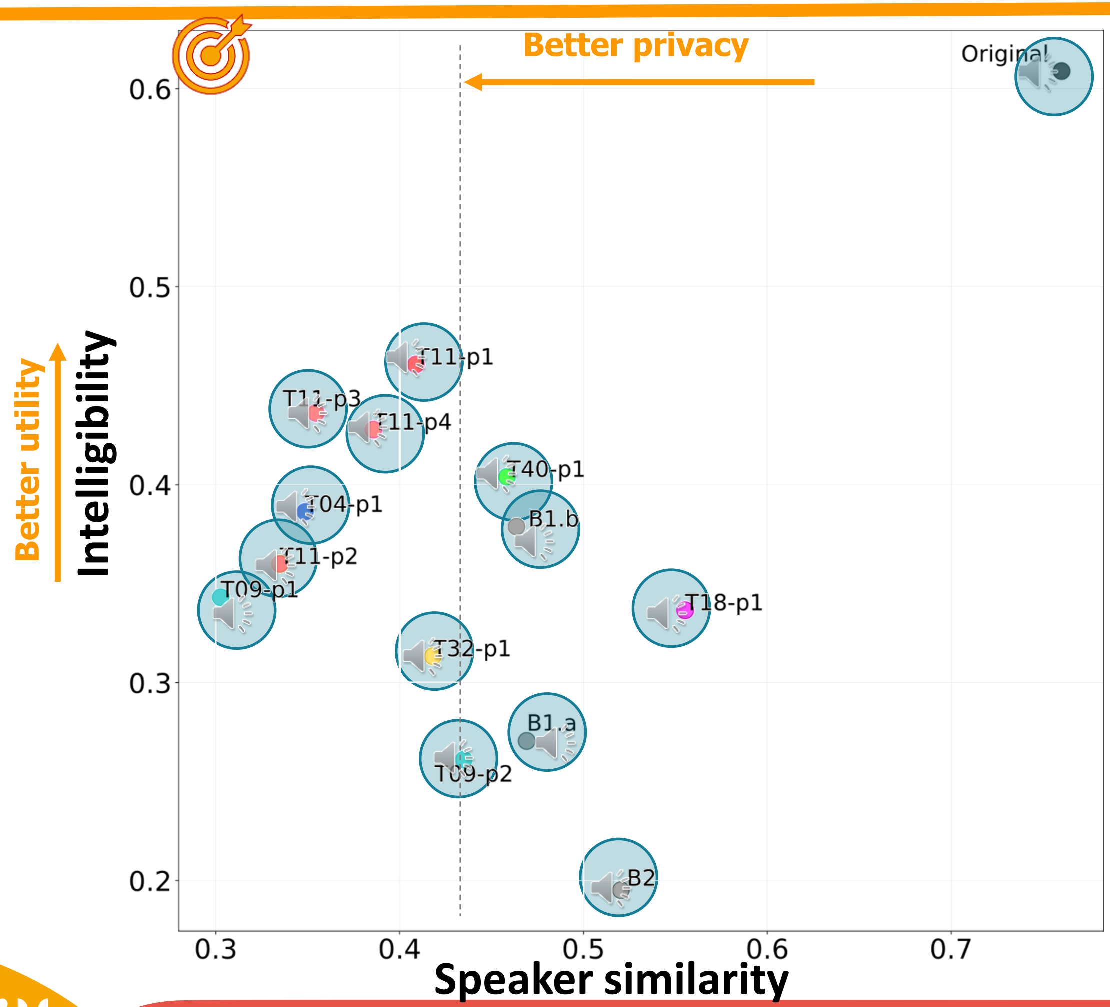
Best systems:

T09-p1 – use opposite gender strategy for x-vector selection
T11-p2, p3 – replace x-vectors by speaker ids from a look-up table, averaging (low voice distinctiveness)
T04-p1 – phonetic ASR transcriptions, no usage of original pitch

Subjective results: privacy vs utility



Subjective results: privacy vs utility

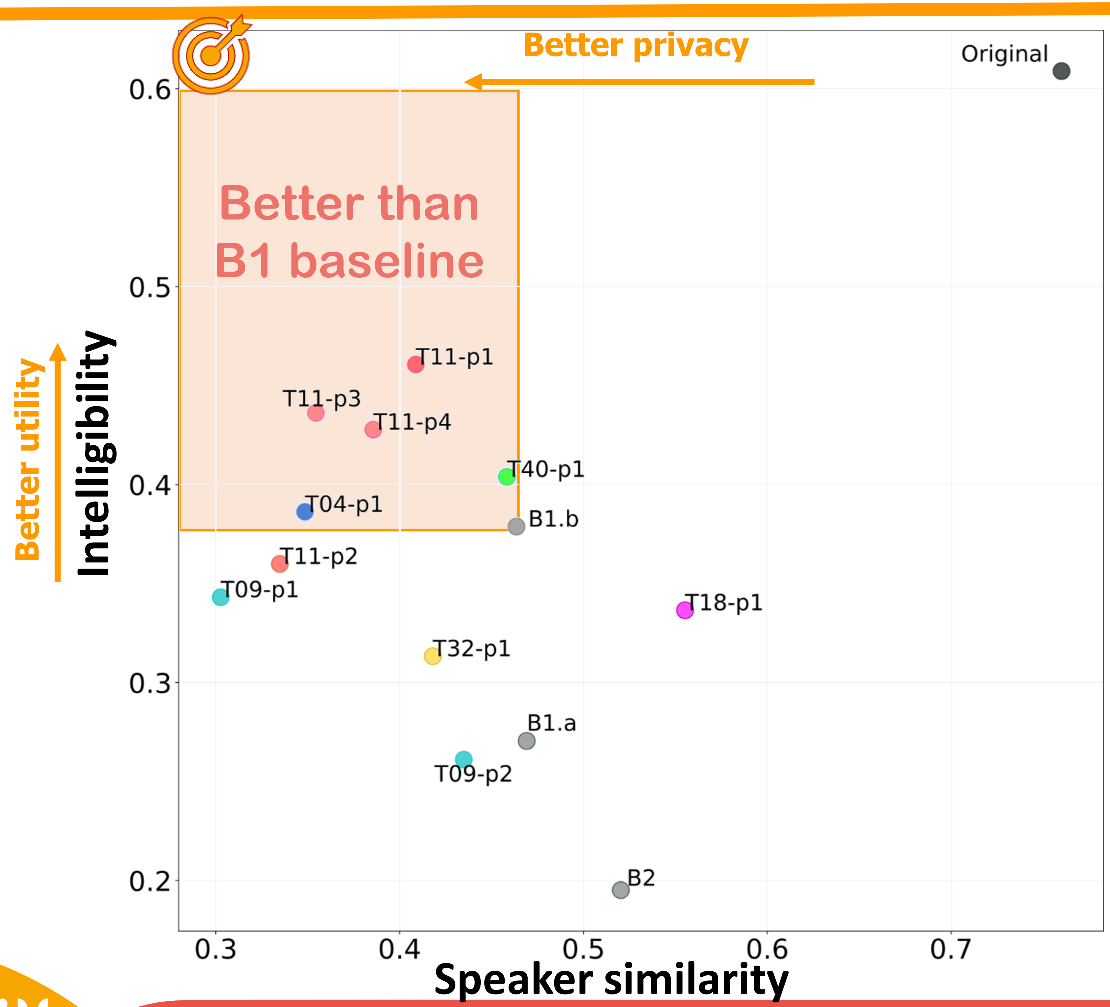


- Similar voice for all systems T11-* (and for all speakers)
- T04-p1 – change speaking rate w.r.t to original
- T09-* - different speaker gender

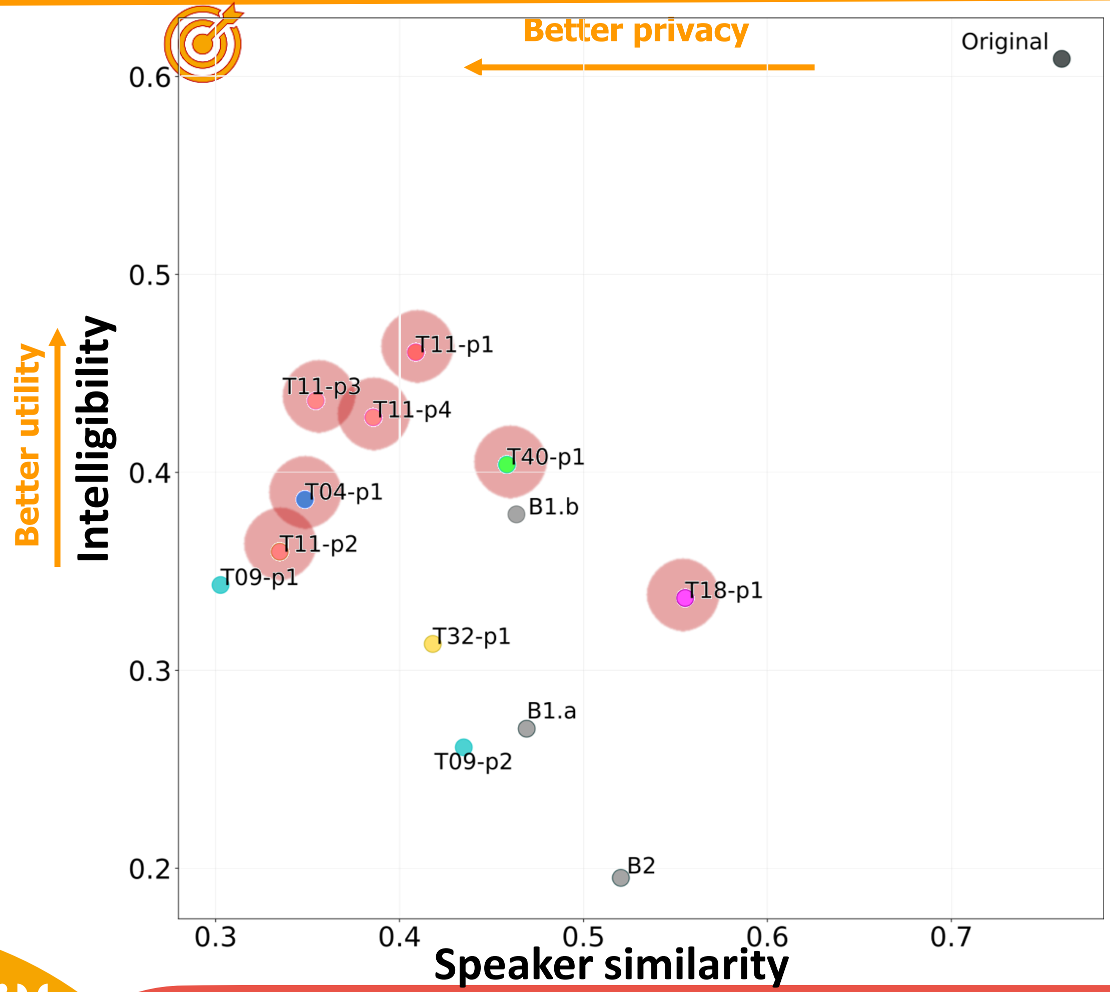
✓ All systems: anonymized speech sounds different from original speakers

! All systems: anonymized speech is less natural and intelligible (the gap decreased w.r.t. 2020)

Subjective results: privacy vs utility



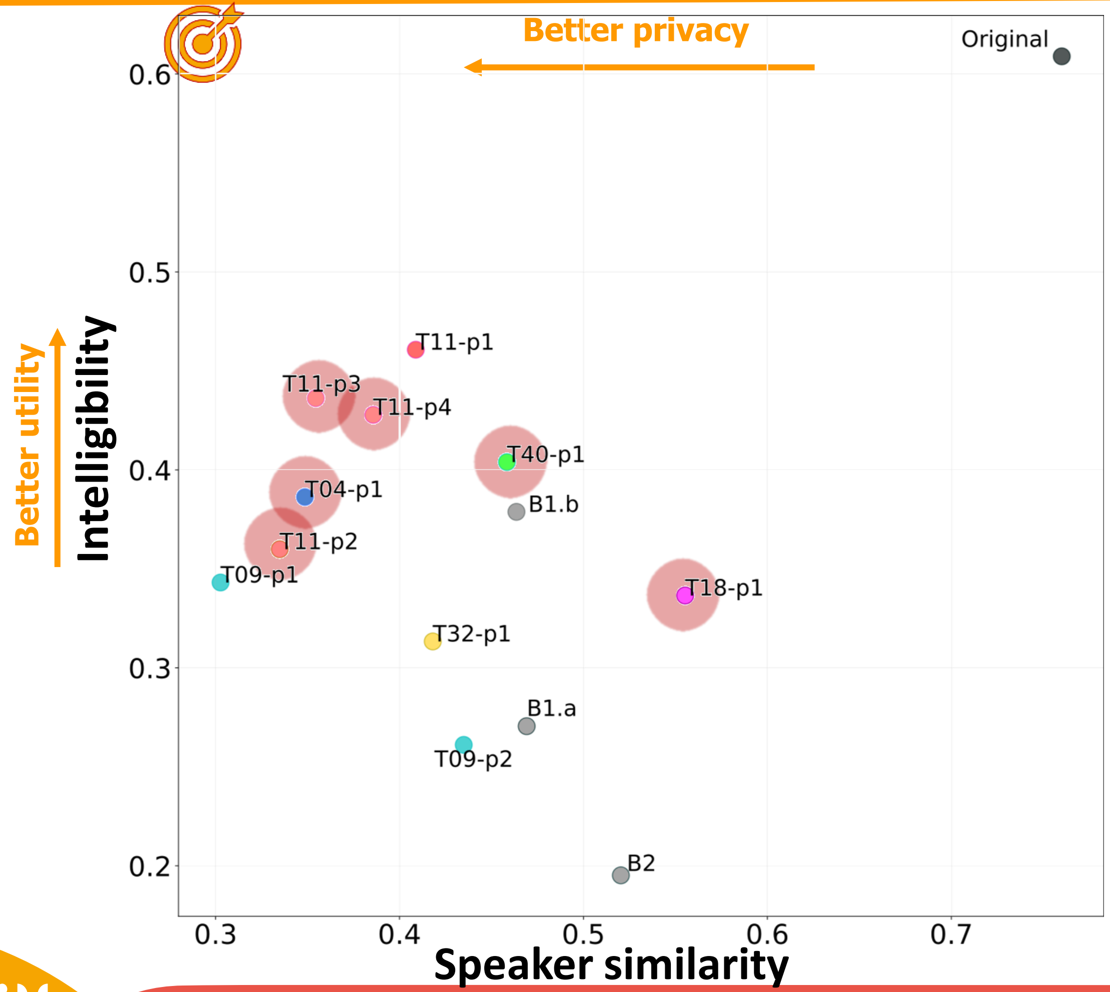
Subjective results: privacy vs utility



Objective privacy conditions:

1. EER ≥ 15%
2. EER ≥ 20%
3. EER ≥ 25%
4. EER ≥ 30%

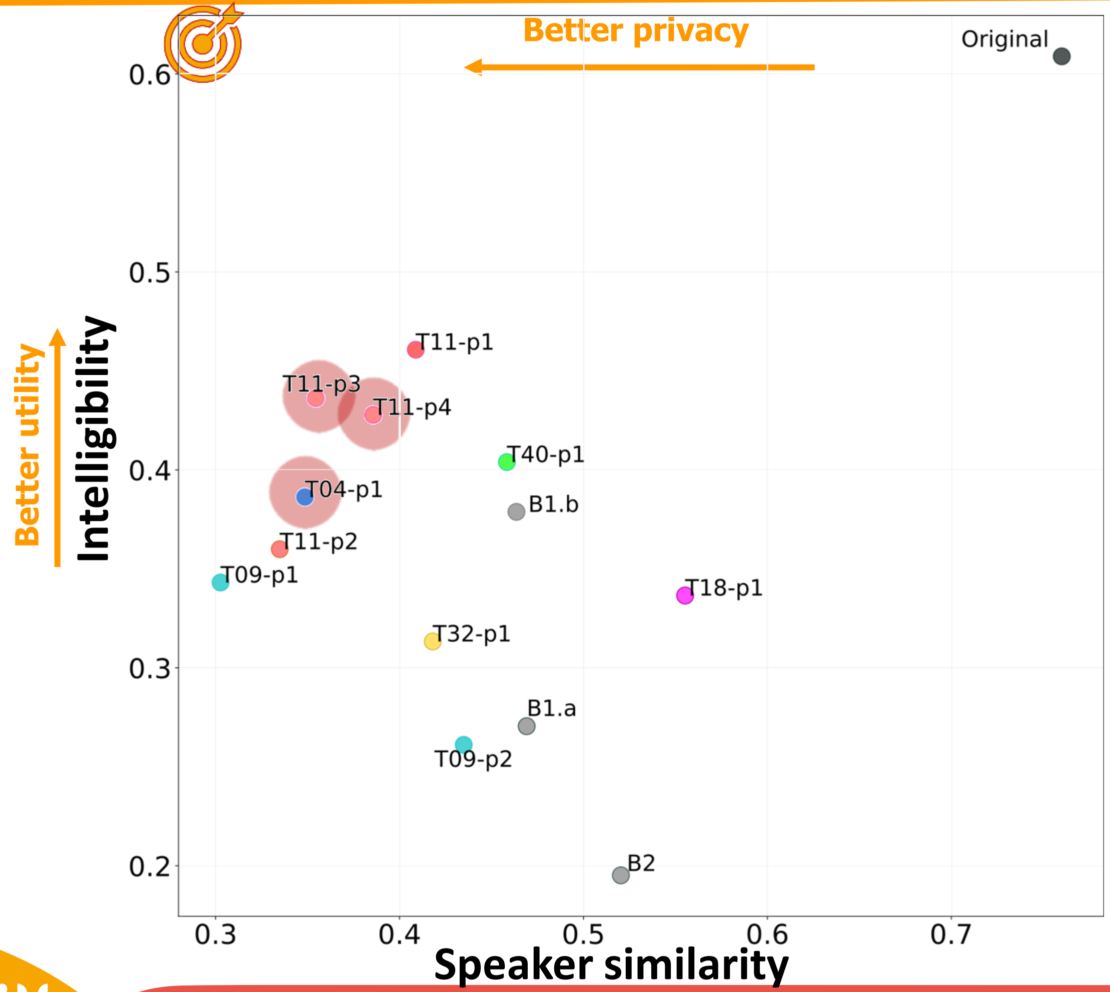
Subjective results: privacy vs utility



Objective privacy conditions:

1. EER ≥ 15%
2. EER ≥ 20%
3. EER ≥ 25%
4. EER ≥ 30%

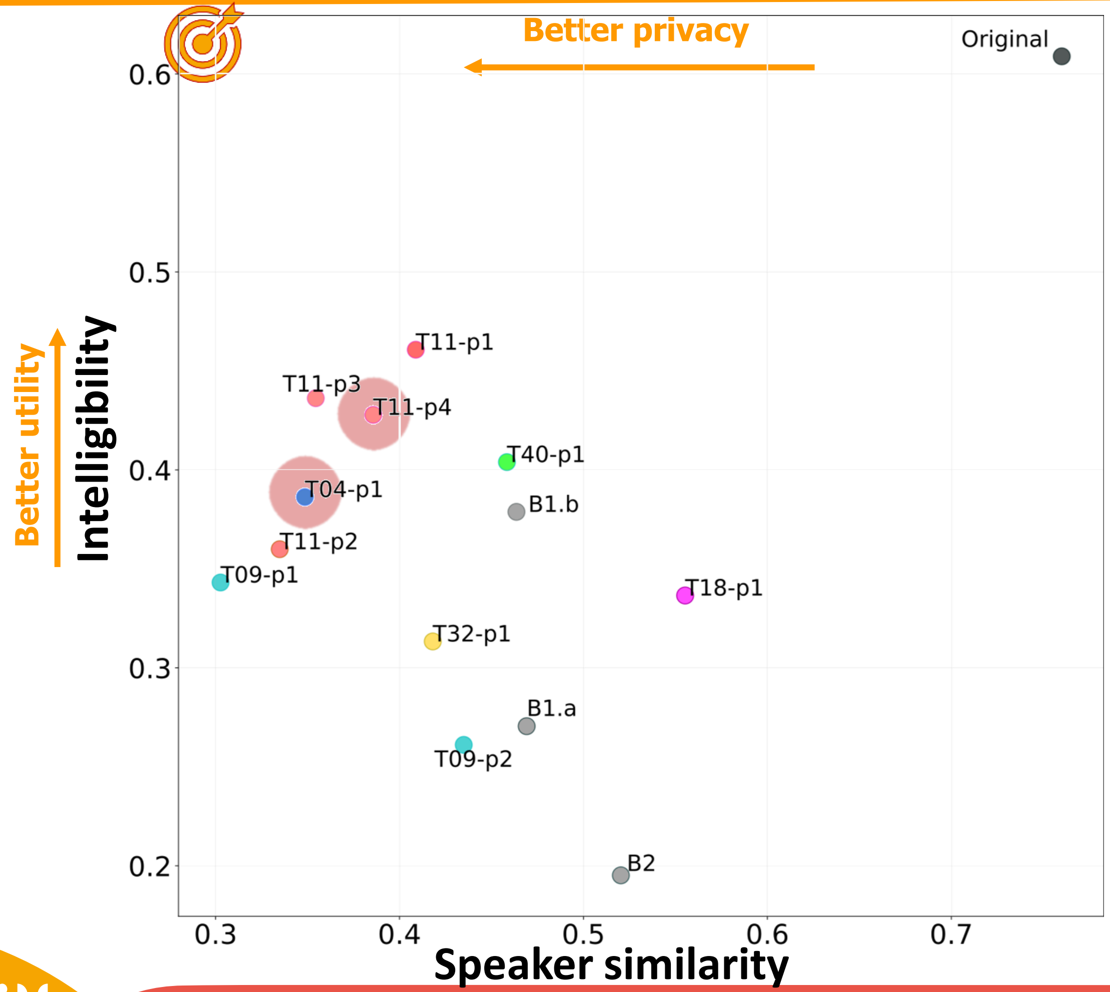
Subjective results: privacy vs utility



Objective privacy conditions:

1. EER ≥ 15%
2. EER ≥ 20%
3. $\boxed{\text{EER} \geq 25\%}$
4. EER ≥ 30%

Subjective results: privacy vs utility



Objective privacy conditions:

1. EER ≥ 15%
2. EER ≥ 20%
3. EER ≥ 25%
4. EER ≥ 30%

Summary and conclusions

Progress in anonymization 2020→2022:

- **Challenge setup**:
  - Stronger attacker for objective evaluation
  - Improved (in utility and computational efficiency) **B1.b** baseline

- **Participants**:
  - Many effective systems (different from the baselines)
  - 3 teams **T11, T04, T40** developed systems that do **not degrade** (even improve) the average **primary utility** metric (WER) while meeting the minimum target privacy requirements:
    - EER≥20 → {**T11, T04, T40**}
    - EER≥30 → {**T11, T04**}

★ T04: EER>45%

Summary and conclusions

Progress in anonymization 2020→2022:

- **Participants**:
  - Proposed approaches and improvements in different components:
    - GAN-based x-vector anonymization **T04**
    - Pitch:
      estimation from BN-features and (anonymized) x-vectors using DNN **T04**
      removal of original pitch, estimation from content **T40**
    - Speaker embeddings: based on speaker ids from look-up-table **T11**
    - Linguistic content: phonetic speech recognition **T04**
    - ...
  - Overall improvement in privacy & utility for subjective and objective evaluation (i.e. 2020 on semi-informed attacker (speaker-level that is weaker than utterance-level in 2022) EER < 25%)

Summary and conclusions

- **2** classes of anonymization methods:
  - **x-vector-based** with speech synthesis models (B1 and related methods) – more effective
  - **signal-processing** based (B2 and others)
- Limitations of the best systems **T11, T04** according to the secondary metrics:
  - Low pitch correlation (however, we aim to keep the prosody/intonation and not all the information in the pitch curve (i.e. not speaker id))
  - Low voice distinctiveness

Perspectives, questions, and future challenges

- Improve anonymization **methods** for stronger baseline solutions
  - x-vector-based (remove residual speaker information form phonetic features & pitch); adversarial approaches, improved synthesis models, better disentanglement
  - simplified, user-friendly software
  - hybrid approaches with other privacy-preservation methods
- **Attributes** (gender, accent, age, emotion,…): anonymize or preserve depending on the task
- Develop **prosody correlation metric**:
  - Pitch correlation is not a suitable utility metric (pitch contains speaker information thus this metric is too (unnecessary) restrictive) + subjective evaluation?
- Improve **voice distinctiveness metric** for anonymized voices
  - Current G_{VD} metric relies on LLR scores from ASV_{orig} model (not suitable for anonymized data) + subjective evaluation?

Perspectives, questions, and future challenges

- Improve anonymization **methods** for stronger baseline solutions
  - x-vector-based (remove residual speaker information form phonetic features & pitch); adversarial approaches, improved synthesis models, better disentanglement
  - simplified, user-friendly software
  - hybrid approaches with other privacy-preservation methods
- **Attributes** (gender, accent, age, emotion,…): anonymize or preserve depending on the task
- Develop **prosody correlation metric**:
  - Pitch correlation is not a suitable utility metric (pitch contains speaker information thus this metric is too (unnecessary) restrictive) + subjective evaluation?
- Improve **voice distinctiveness metric** for anonymized voices
  - Current G_{VD} metric relies on LLR scores from ASV_{orig} model (not suitable for anonymized data) + subjective evaluation?

Perspectives, questions, and future challenges

- Improve anonymization **methods** for stronger baseline solutions
  - x-vector-based (remove residual speaker information form phonetic features & pitch); adversarial approaches, improved synthesis models, better disentanglement
  - simplified, user-friendly software
  - hybrid approaches with other privacy-preservation methods
- **Attributes** (gender, accent, age, emotion,…): anonymize or preserve depending on the task
- Develop **prosody correlation metric**:
  - Pitch correlation is not a suitable utility metric (pitch contains speaker information thus this metric is too (unnecessary) restrictive) + subjective evaluation?
- Improve **voice distinctiveness metric** for anonymized voices
  - Current G_{VD} metric relies on LLR scores from ASV_{orig} model (not suitable for anonymized data) + subjective evaluation?

Perspectives, questions, and future challenges

- Improve anonymization **methods** for stronger baseline solutions
  - x-vector-based (remove residual speaker information form phonetic features & pitch); adversarial approaches, improved synthesis models, better disentanglement
  - simplified, user-friendly software
  - hybrid approaches with other privacy-preservation methods
- **Attributes** (gender, accent, age, emotion,…): anonymize or preserve depending on the task
- Develop **prosody correlation metric**:
  - Pitch correlation is not a suitable utility metric (pitch contains speaker information thus this metric is too (unnecessary) restrictive) + subjective evaluation?
- Improve **voice distinctiveness metric** for anonymized voices
  - Current G_{VD} metric relies on LLR scores from ASV_{orig} model (not suitable for anonymized data) + subjective evaluation?

Perspectives, questions, and future challenges

- **Privacy** vs **utility trade-off**
  - Better ranking policy?
  - Incorporate into system development
- Using other open resources to develop anonymization and attack models (i.e. SSL models, other languages)
- Develop **stronger** and **more realistic attack models**:

VoicePrivacy Attacker Challenge

Perspectives, questions, and future challenges

- **Privacy** vs **utility trade-off**
  - Better ranking policy?
  - Incorporate into system development
- Using other open resources to develop anonymization and attack models (i.e. SSL models, other languages)
- Develop **stronger** and **more realistic attack models**:

VoicePrivacy Attacker Challenge

Perspectives, questions, and future challenges

- **Privacy** vs **utility trade-off**
  - Better ranking policy?
  - Incorporate into system development
- Using other open resources to develop anonymization and attack models (i.e. SSL models, other languages)
- Develop **stronger** and **more realistic attack models**:

VoicePrivacy Attacker Challenge

References: participants' papers

- **T04:** [Meyer 2022] Cascade of Phonetic Speech Recognition, Speaker Embeddings GAN and Multispeaker Speech Synthesis for the VoicePrivacy 2022 Challenge. Sarina Meyer, Pascal Tilli, Florian Lux, Pavel Denisov, Julia Koch, Ngoc Thang Vu
- **T11:** [Yao 2022] NWPU-ASLP System for the VoicePrivacy 2022 Challenge. Jixun Yao, Qing Wang, Li Zhang, Pengcheng Guo, Yuhao Liang, Lei Xie
- **T18:** [Chen 2022] System Description for Voice Privacy Challenge 2022. Xiaojiao Chen, Guangxing Li, Hao Huang, Wangjin Zhou, Sheng Li, Yang Cao, Yi Zhao
- **T32:** [Mawalim 2022] System Description: Speaker Anonymization by Pitch Shifting Based on Time-Scale Modification (PV-TSM). Candy Olivia Mawalim, Shogo Okada, Masashi Unoki
- **T40:** [Gaznepoglu 2022] VoicePrivacy 2022 System Description: Speaker Anonymization with Feature-matched F0 Trajectories. Unal Ege Gaznepoglu, Anna Leschanowsky, Nils Peters
- [Khamsehashari 2022] Voice Privacy Challenge - Rethinking the Baseline. Razieh Khamsehashari, Yamini Sinha, Jan Hintz, Suhita Ghosh, Tim Polzehl, Clarlos Franzreb and Ingo Siegert

The VoicePrivacy Challenge: participants' talks

24th September 9:00-11:00

	<u>VoicePrivacy Challenge</u> • **Speaker Anonymization by Pitch Shifting Based on Time-Scale Modification** Candy Olivia Mawalim, Shogo Okada and Masashi Unoki • **Voice Privacy Challenge - Rethinking the Baseline** Razieh Khamsehashari, Yamini Sinha, Jan Hintz, Suhita Ghosh, Tim Polzehl, Clarlos Franzreb and Ingo Siegert
9:00 - 11:00	• **Cascade of Phonetic Speech Recognition, Speaker Embeddings GAN and Multispeaker Speech Synthesis for the VoicePrivacy 2022 Challenge** Sarina Meyer, Pascal Tilli, Florian Lux, Pavel Denisov, Julia Koch, Ngoc Thang Vu • **NWPU-ASLP System for the VoicePrivacy 2022 Challenge** Jixun Yao, Qing Wang, Li Zhang, Pengcheng Guo, Yuhao Liang, Lei Xie • **System Description for Voice Privacy Challenge 2022** Xiaojiao Chen, Guangxing Li, Hao Huang, Wangjin Zhou, Sheng Li, Yang Cao, Yi Zhao • **VoicePrivacy 2022 System Description: Speaker Anonymization with Feature-matched F0 Trajectories** Unal Ege Gaznepoglu, Anna Leschanowsky, Nils Peters

The VoicePrivacy 2022 Challenge

Thank you!

organisers@lists.voiceprivacychallenge.org

https://www.voiceprivacychallenge.org