The VoicePrivacy 2020 Challenge

Odyssey 2020

Subjective evaluation-1

Presenter: **Xin Wang**

Natalia Tomashenko [1]
Brij M.L. Srivastava [2]
Xin Wang [3]
Emmanuel Vincent [2]
Andreas Nautsch [4]
Junichi Yamagishi [3,5]
Nicholas Evans [4]
Jose Patino [4]
Jean-François Bonastre [1]
Paul-Gauthier Noé [1]
Massimiliano Todisco [4]
Mohamed Maouche [2]
Benjamin O'Brien [6]
Anais Chanclu [1]

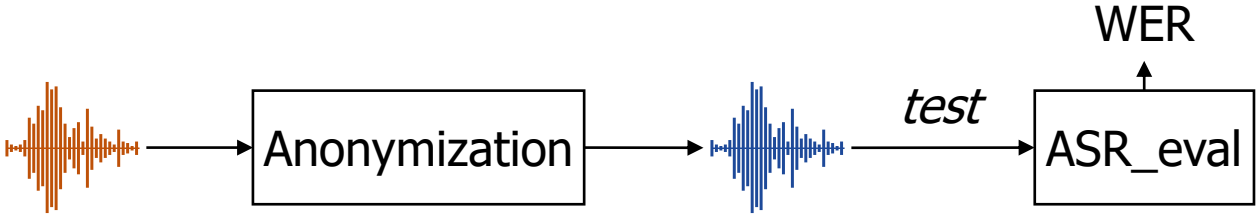
[1] LIA – University of Avignon – France
[2] Inria – France
[3] NII – Tokyo – Japan
[4] Audio Security and Privacy Group, EURECOM – France
[5] University of Edinburgh – UK
[6] LPL – Aix-Marseille University – France

4th November 2020

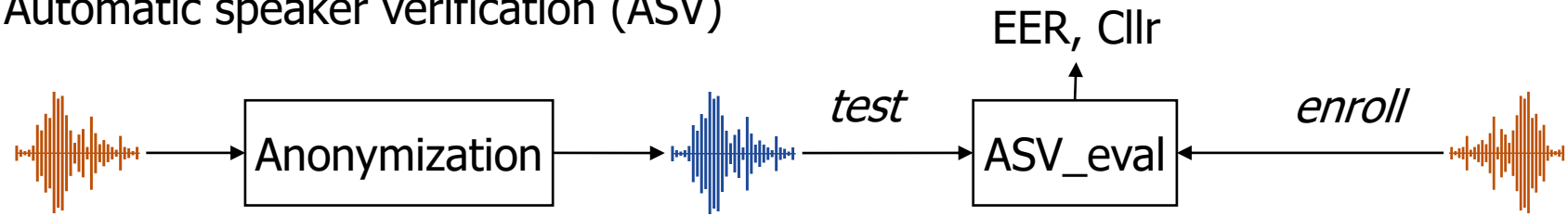
Subjective evaluation design

Objective metrics in evaluation plan

- Automatic speech recognition (ASR)



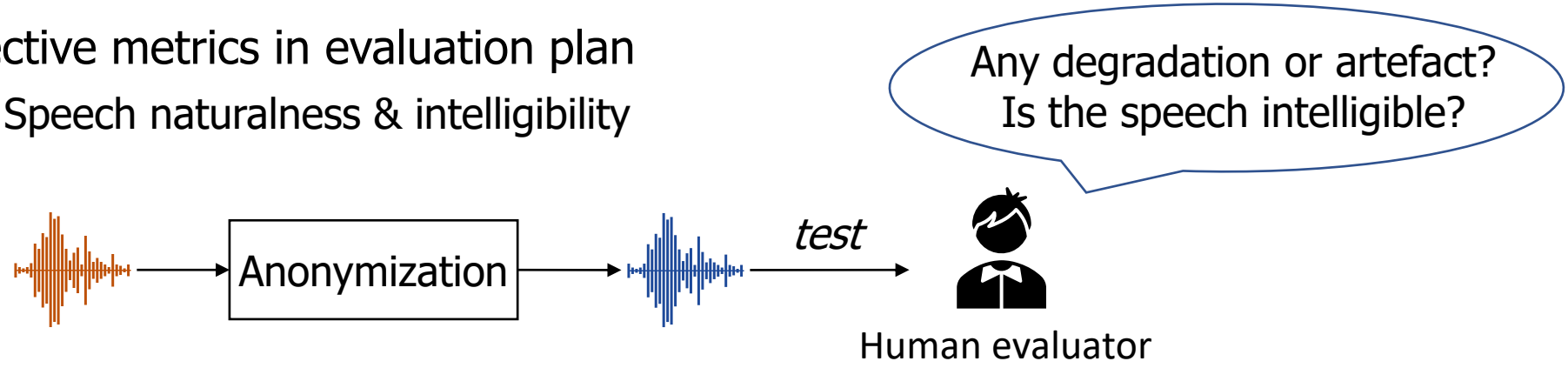
- Automatic speaker verification (ASV)



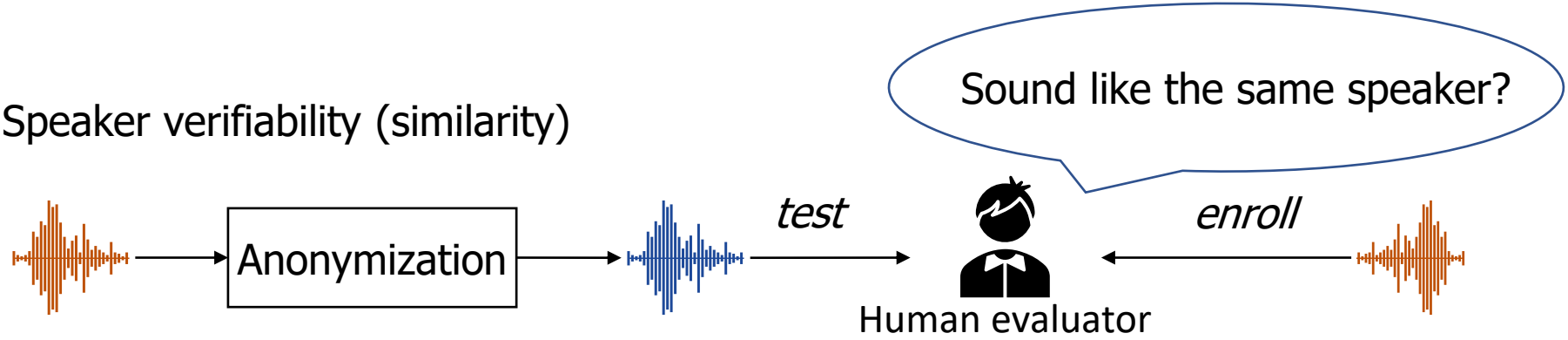
Subjective evaluation design

Subjective metrics in evaluation plan

- Speech naturalness & intelligibility



- Speaker verifiability (similarity)

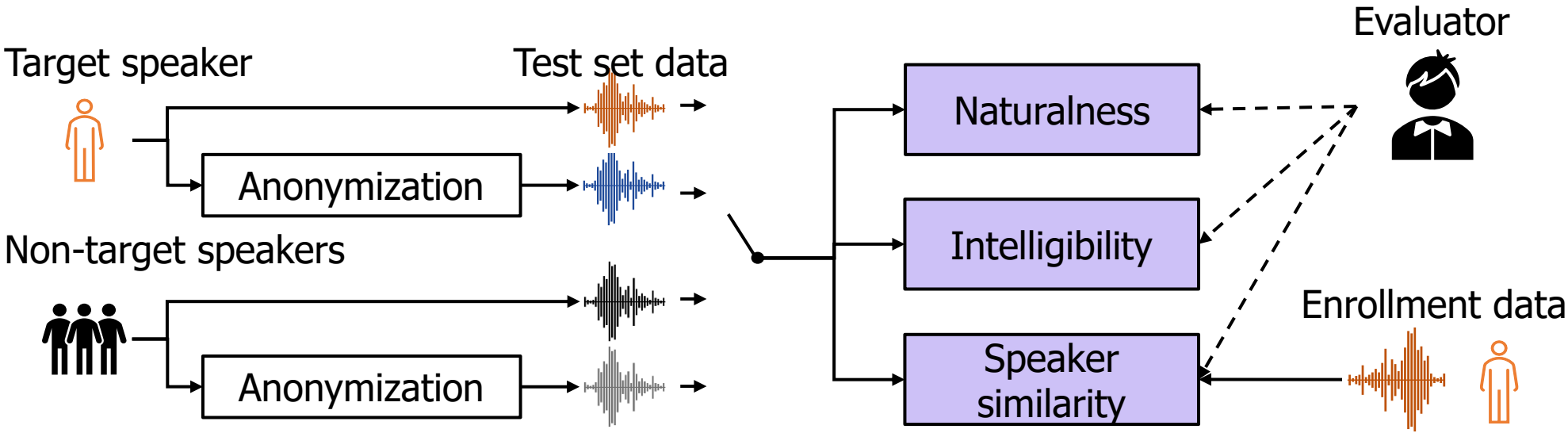


- Speaker linkability (see Part 2 in next presentation)

Subjective evaluation design – Part 1

Listening test with four types of trials

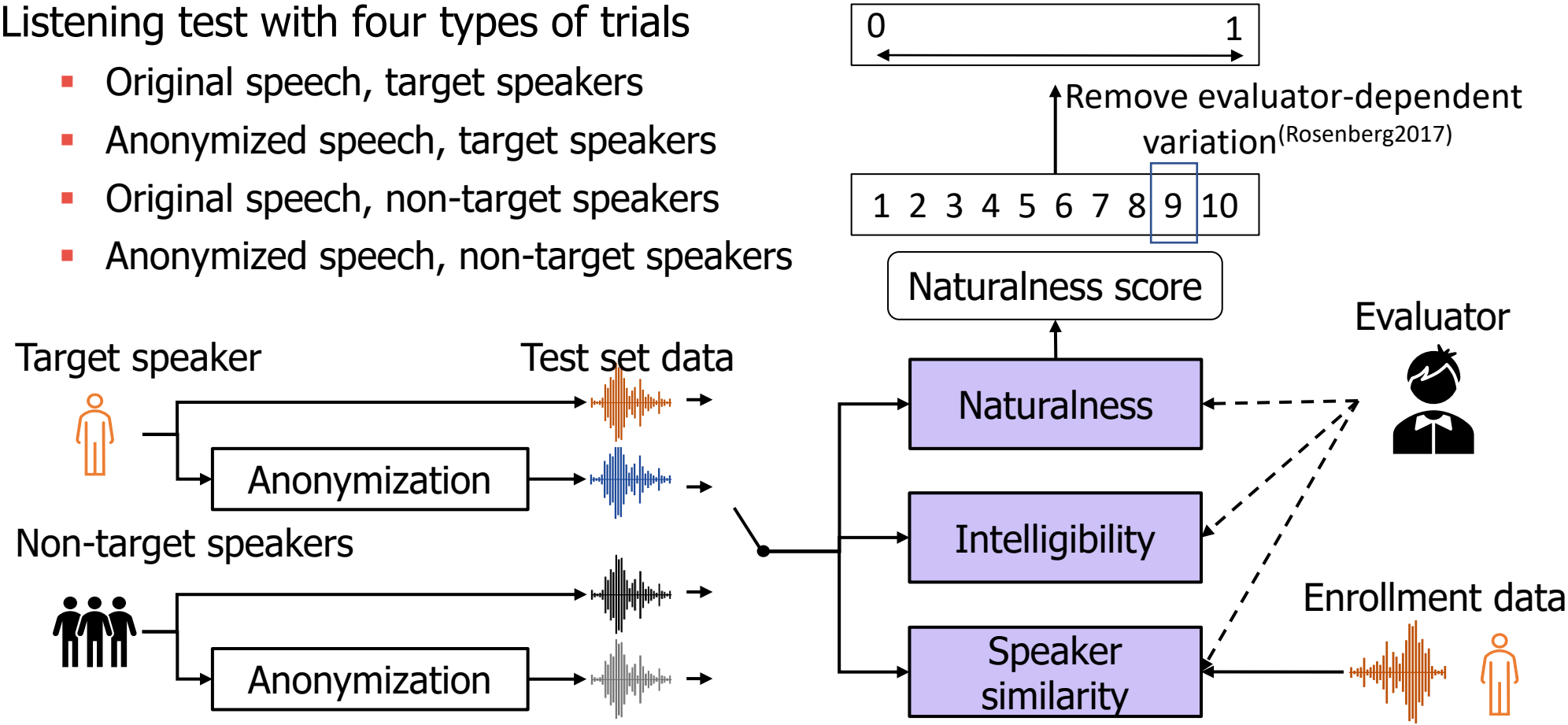
- Original speech, target speakers
- Anonymized speech, target speakers
- Original speech, non-target speakers
- Anonymized speech, non-target speakers



Subjective evaluation design – Part 1

Listening test with four types of trials

- Original speech, target speakers
- Anonymized speech, target speakers
- Original speech, non-target speakers
- Anonymized speech, non-target speakers



Rosenberg, A. et al. Bias and Statistical Significance in Evaluating Speech Synthesis with Mean Opinion Scores. in *Proc. Interspeech* 3976–3980 (2017).

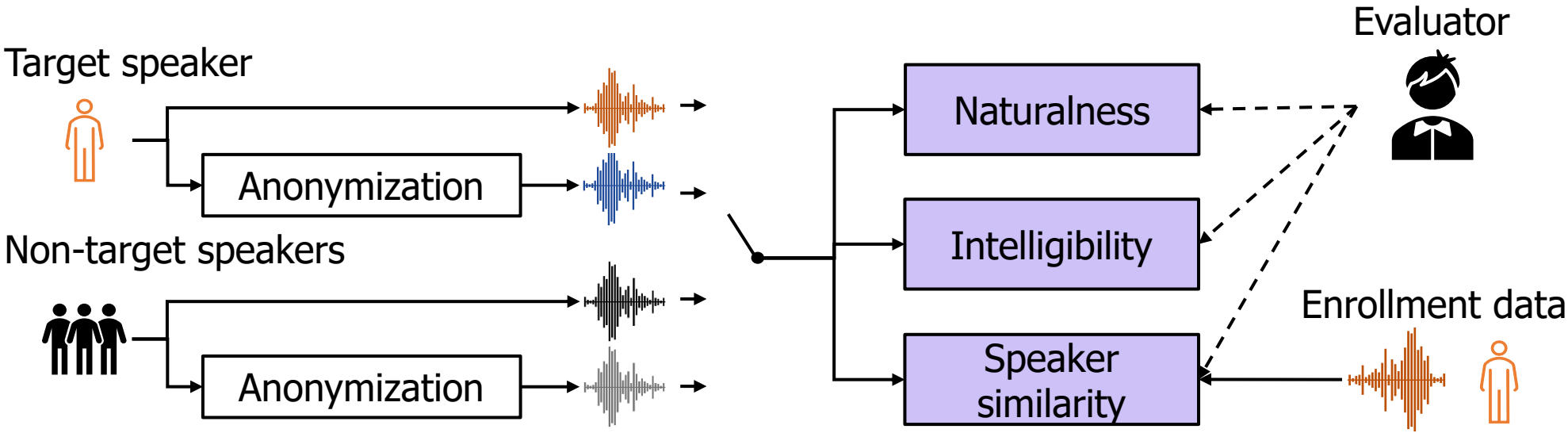
Subjective evaluation results – Part 1

Evaluated trials: 16,200 in total

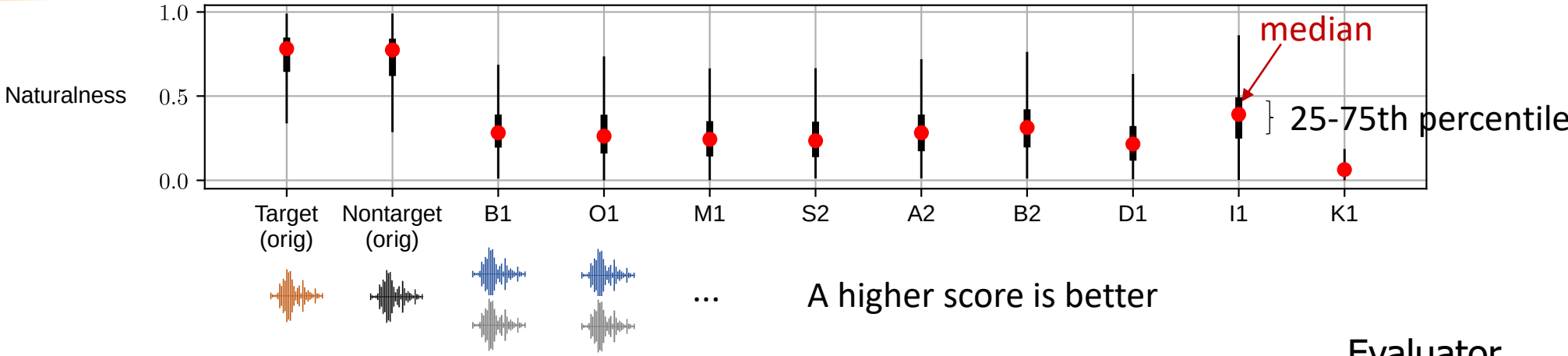
- Trials cover all test speakers & anonymization systems
- Four types of trials 1:1:1:1
- 36 trials in one listening set

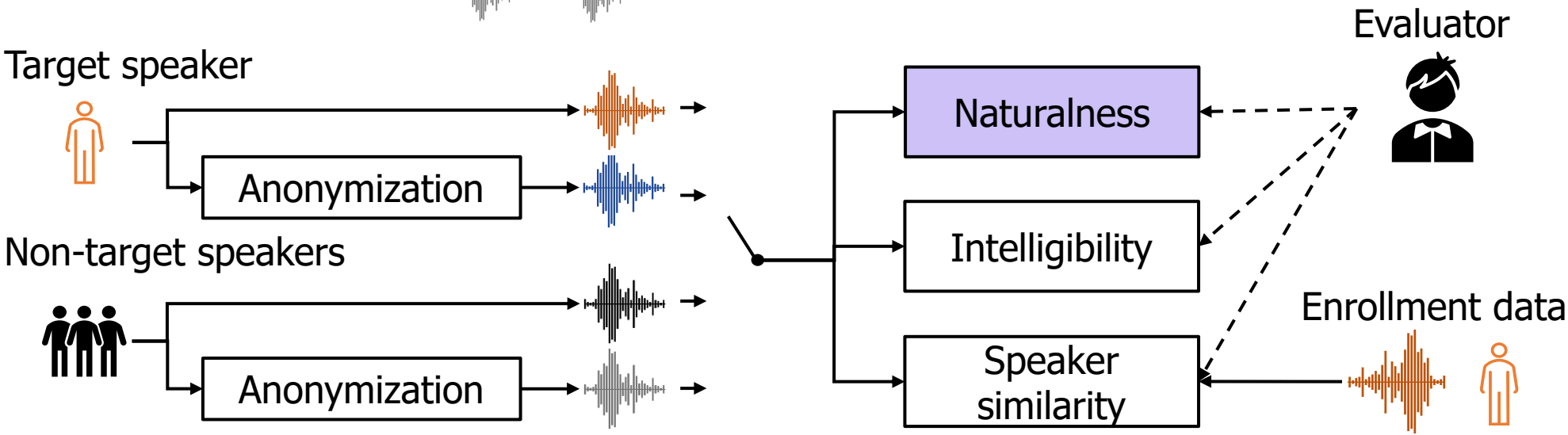
Evaluators: 47

- English as mother tongue: US: 24, UK: 11, ...
- Many of them evaluated 360 trials (10 sets)

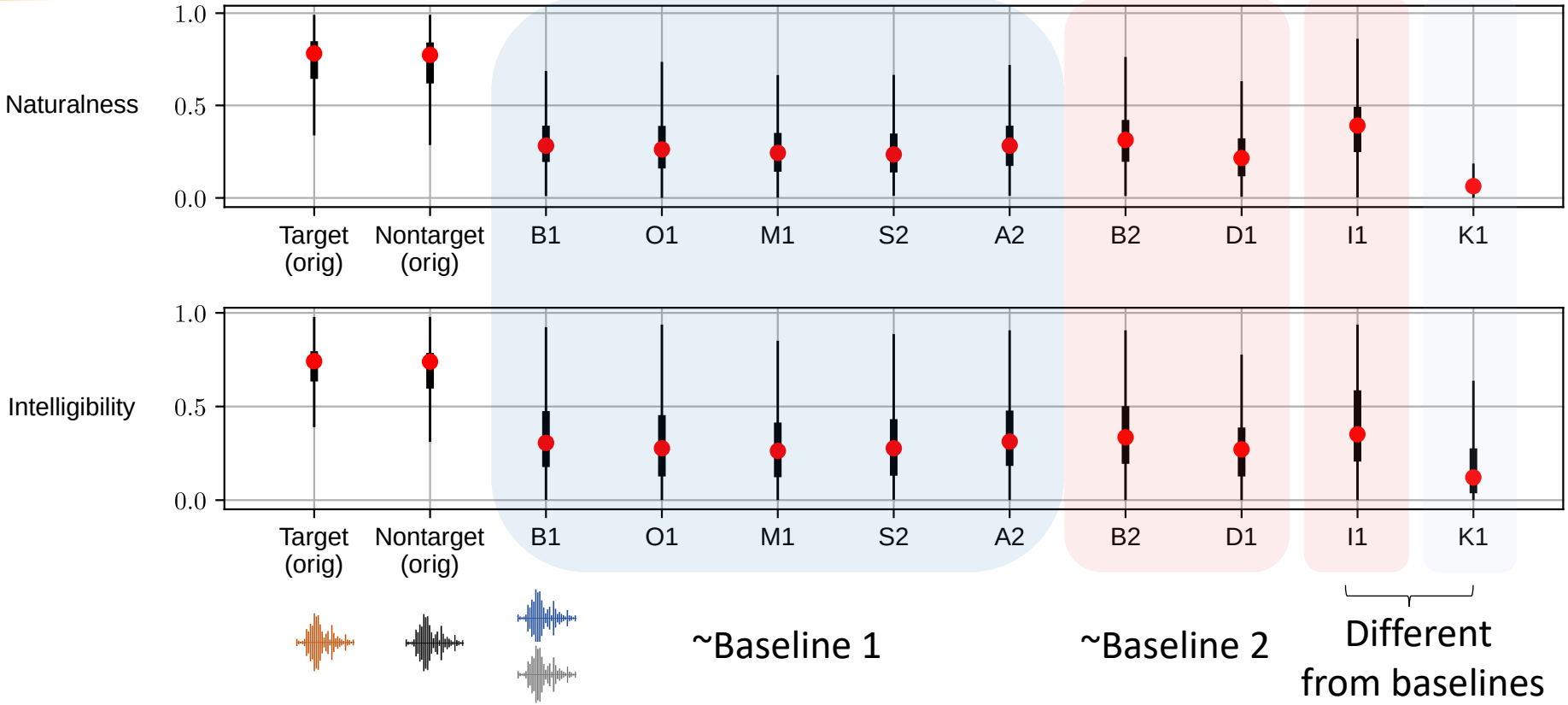


Subjective evaluation results – Part 1





Subjective evaluation results – Part 1

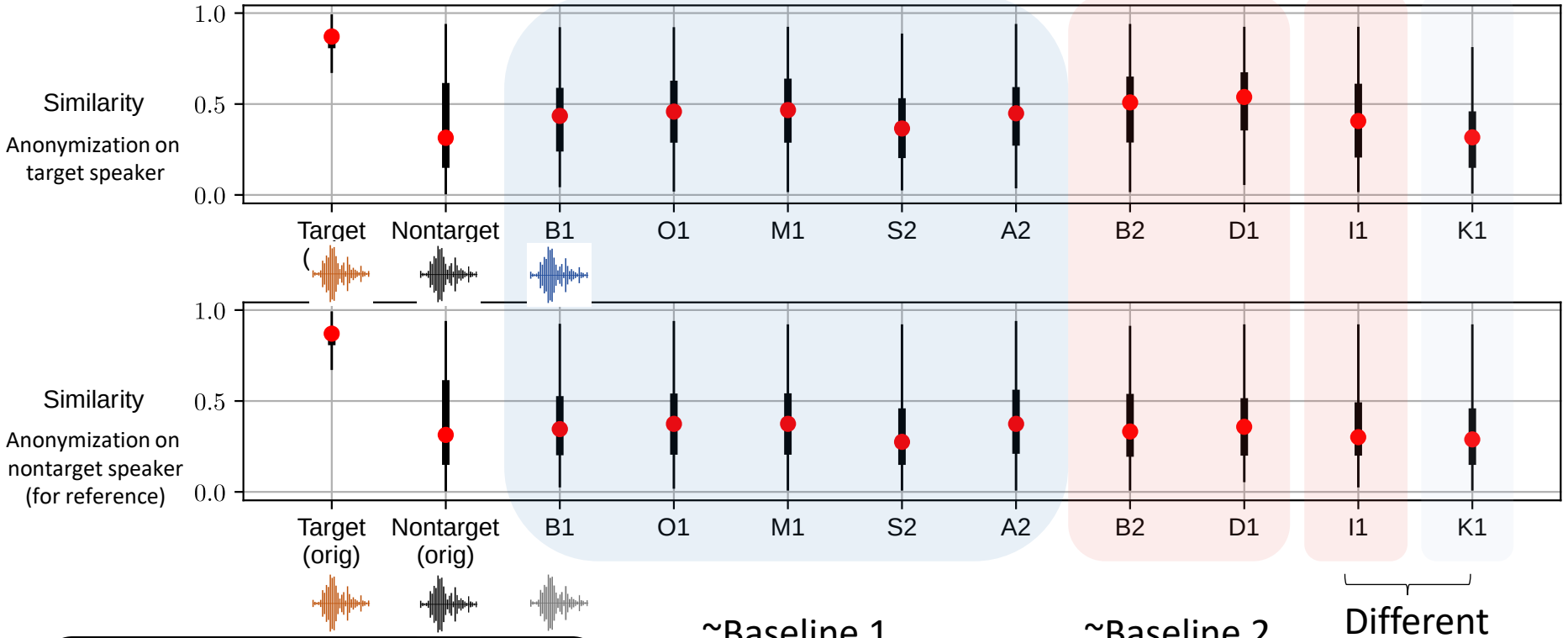


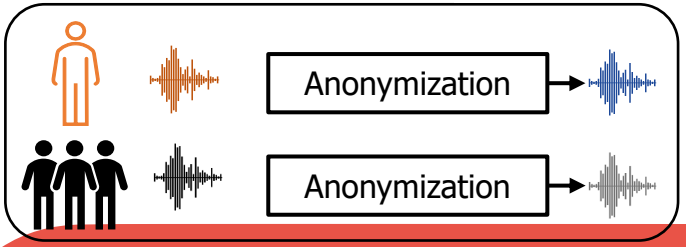
A higher score -> better utility

x-vector based neural model

signal-processing

Subjective evaluation results – Part 1





~Baseline 1

~Baseline 2

Different from baselines

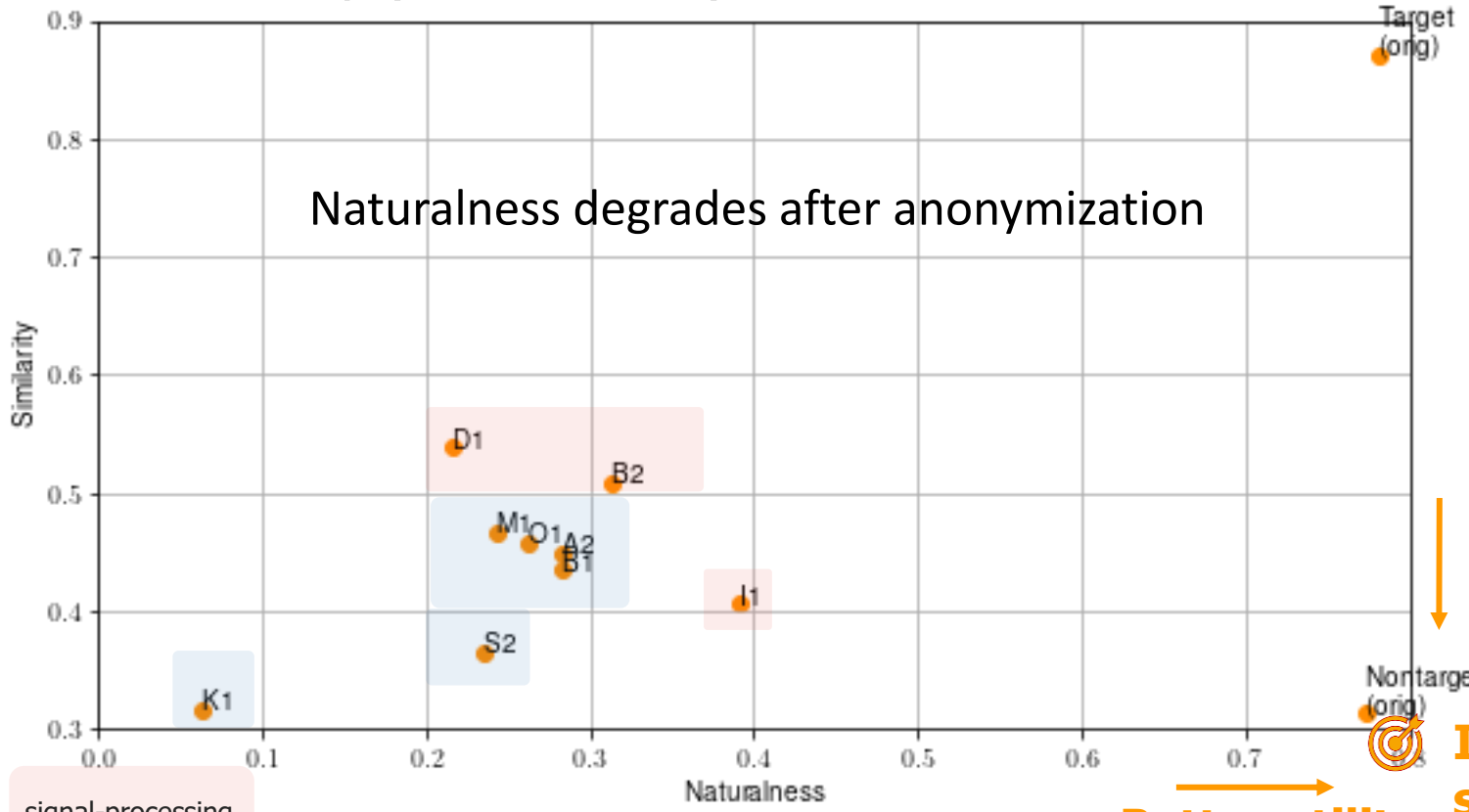
A lower score -> better privacy

x-vector based neural model

signal-processing

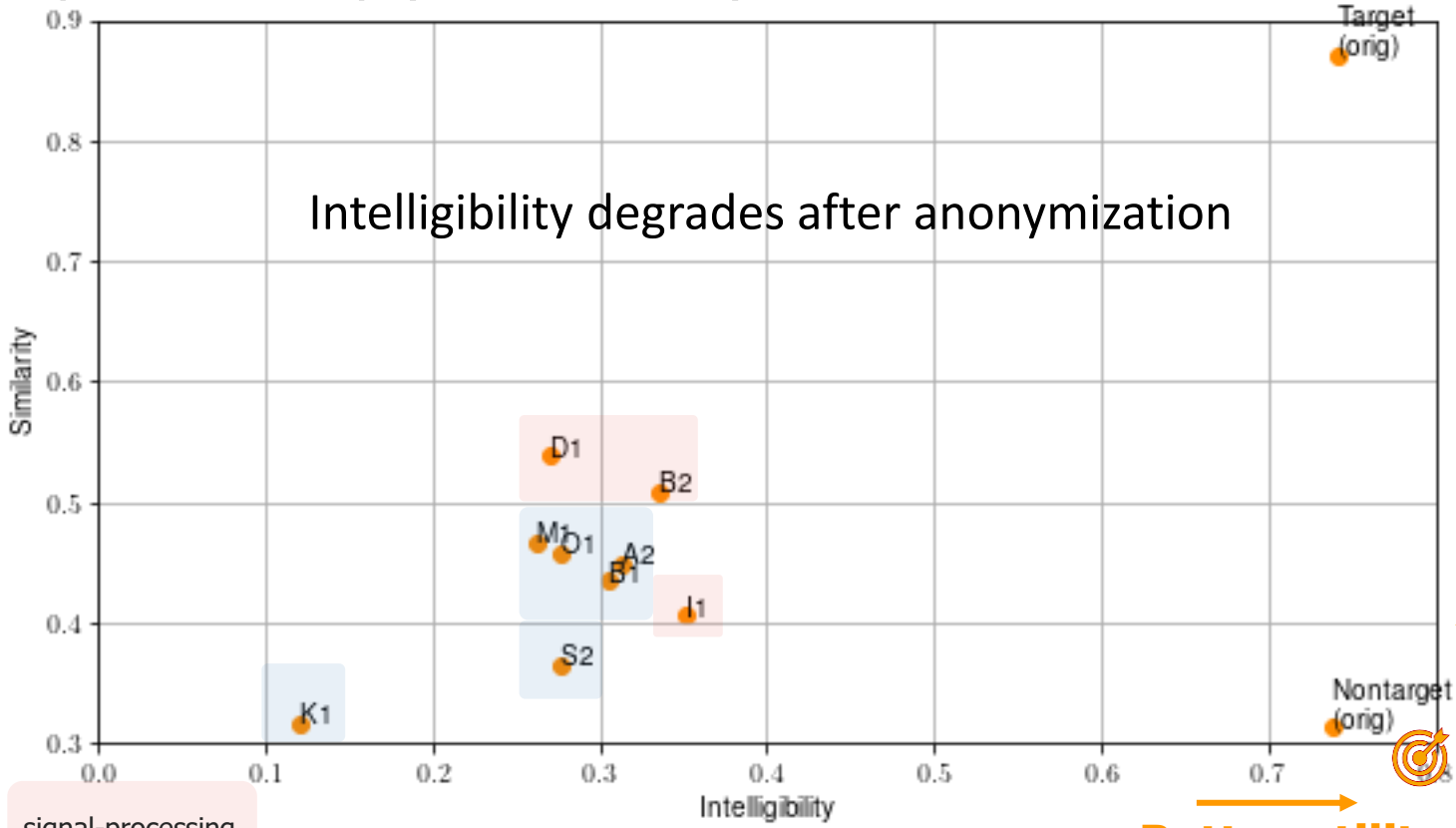
Subjective evaluation results – Part 1

Naturalness vs Similarity (median score)



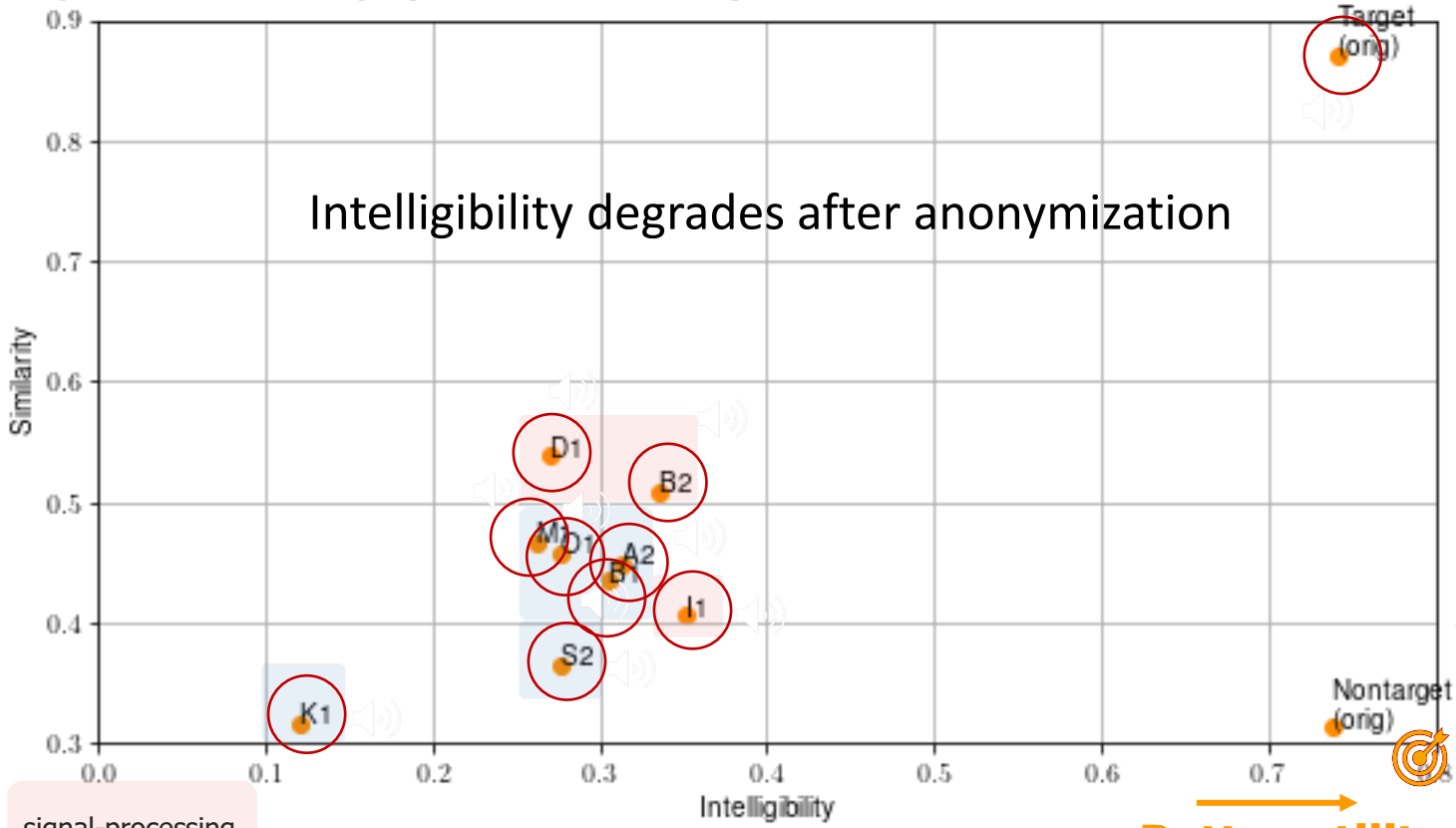
Subjective evaluation results – Part 1

Intelligibility vs Similarity (median score)



Subjective evaluation results – Part 1

Intelligibility vs Similarity (median score)



Better privacy

Ideal system

Better utility

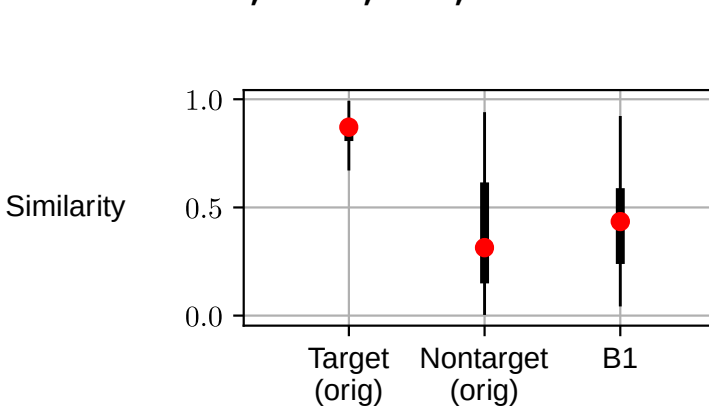
x-vector based neural model

signal-processing

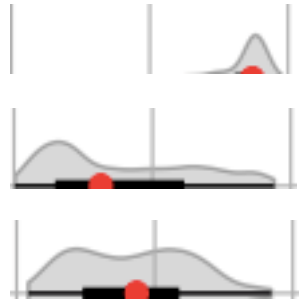
Subjective evaluation results – Part 1

More analysis were conducted (☞ appendix)

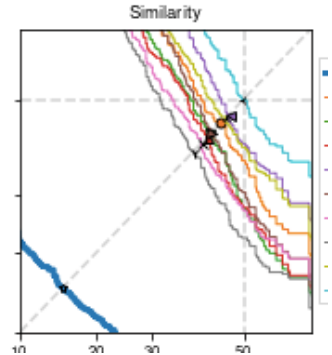
- DET curves, EER, Cllr, ROCCH-EER



Score distribution



DET curve



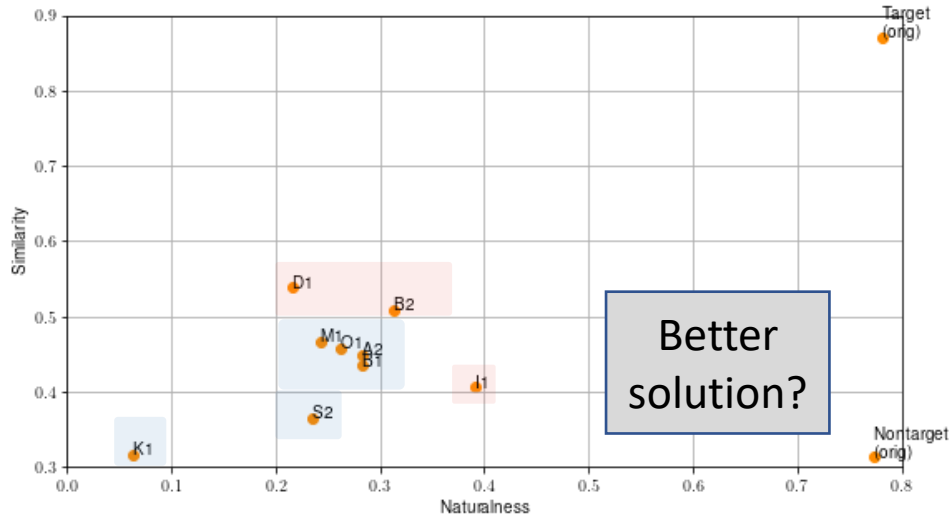
- Correlation between subjective (human perception) and objective (ASR & ASV) scores
  - High correlation

Subjective evaluation summary – Part 1

Messages

- ✓ All systems: anonymized speech sounds different from original speakers
- ! All systems: anonymized speech are less natural and intelligible
- Similar trends in objective results (☞ appendix)
- Further improvement is necessary

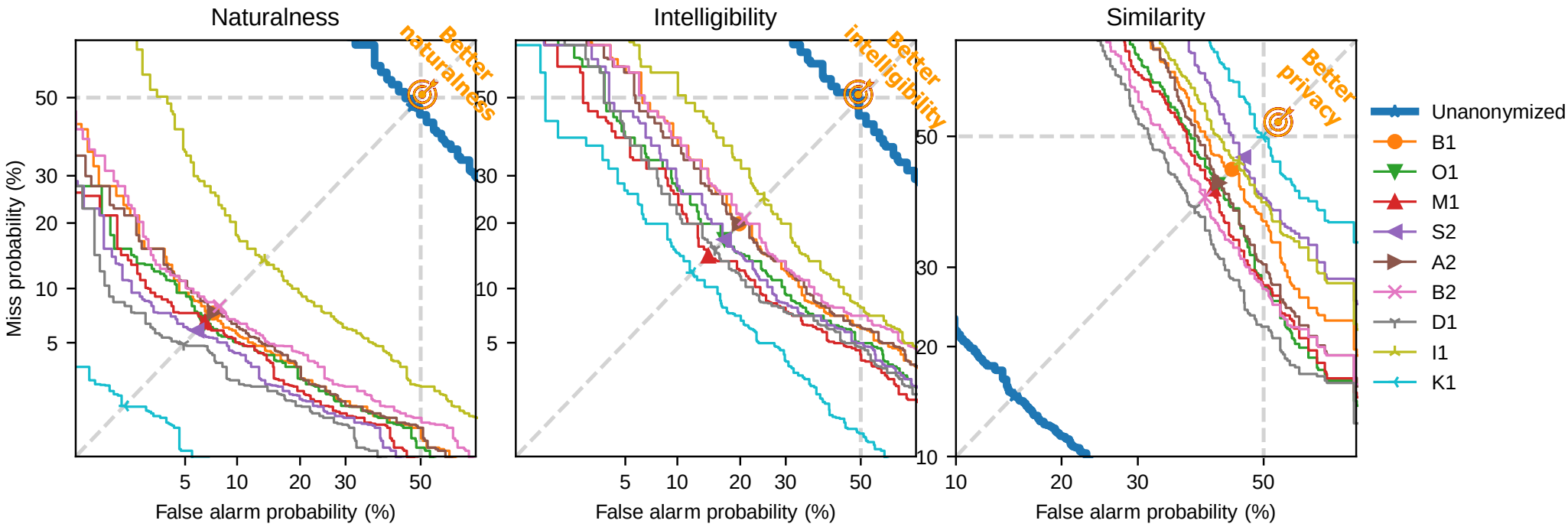
Thank you for your attention



Subjective evaluation results – Part 1

More analysis

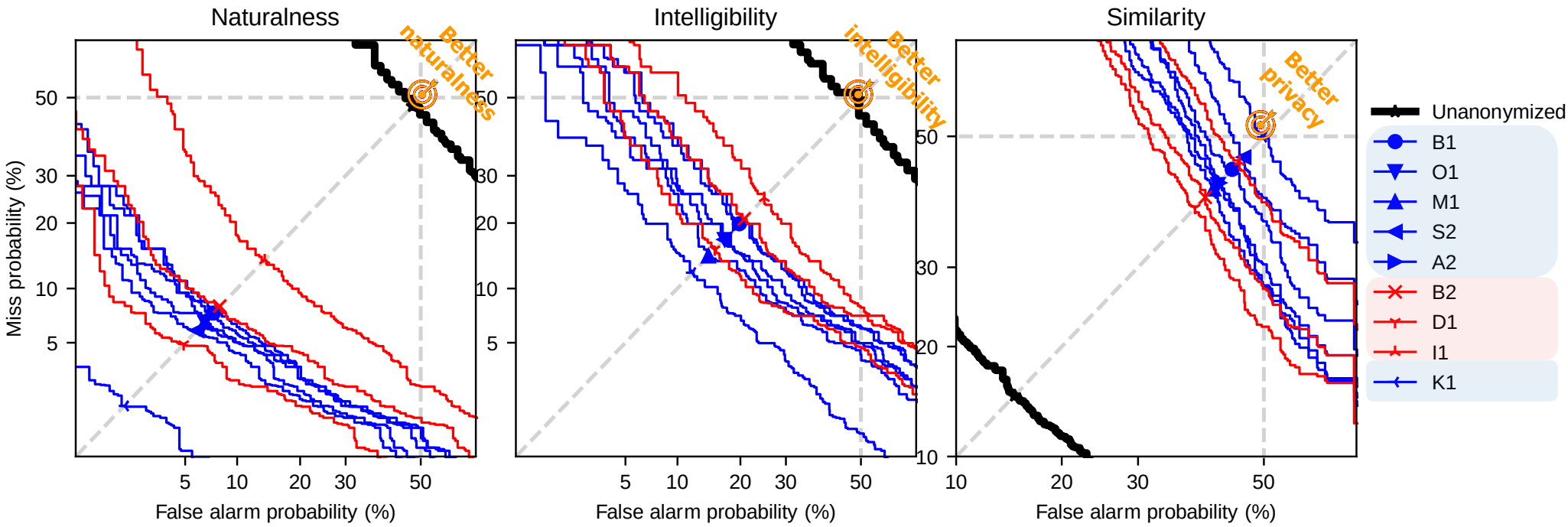
- EER & DET curves



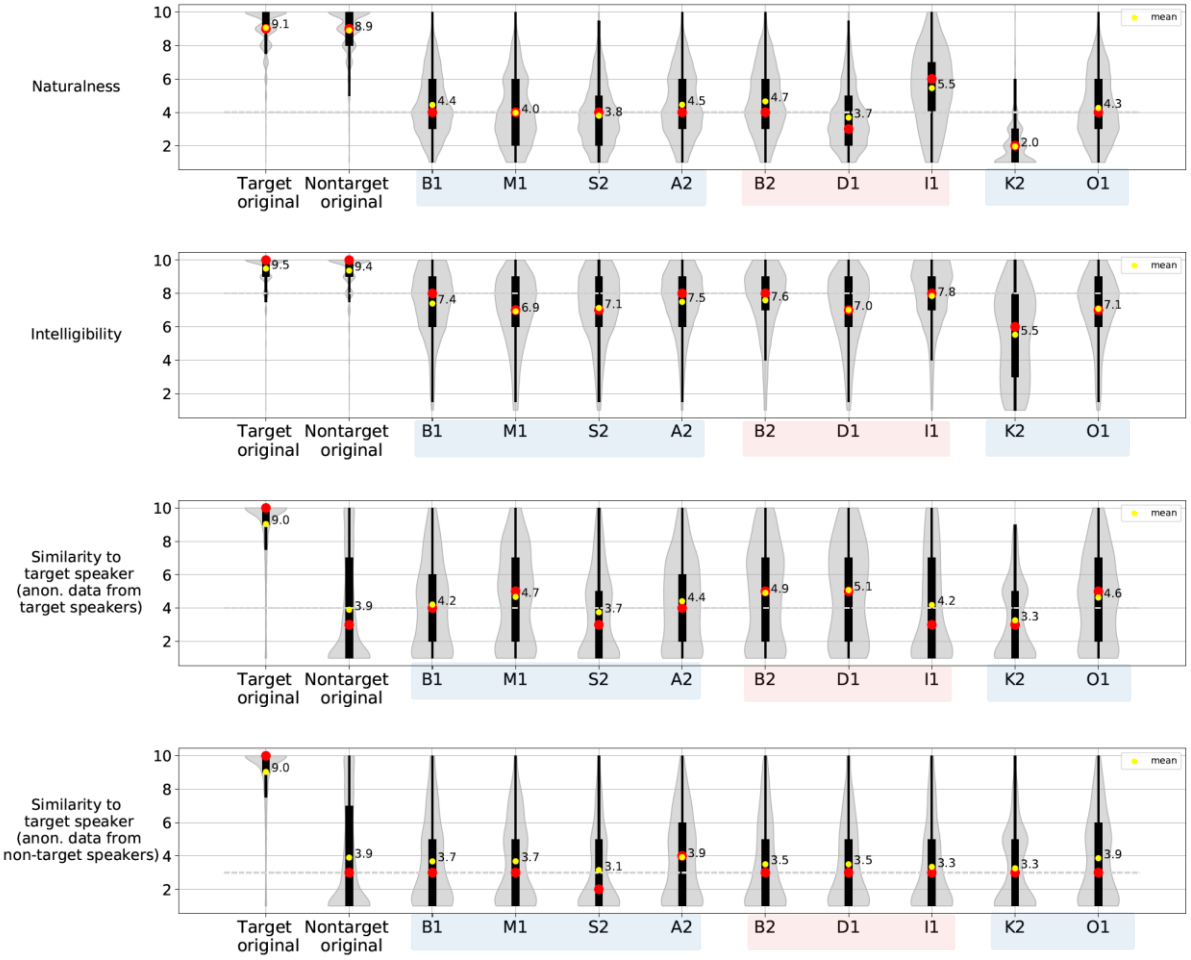
Subjective evaluation results – Part 1

More analysis

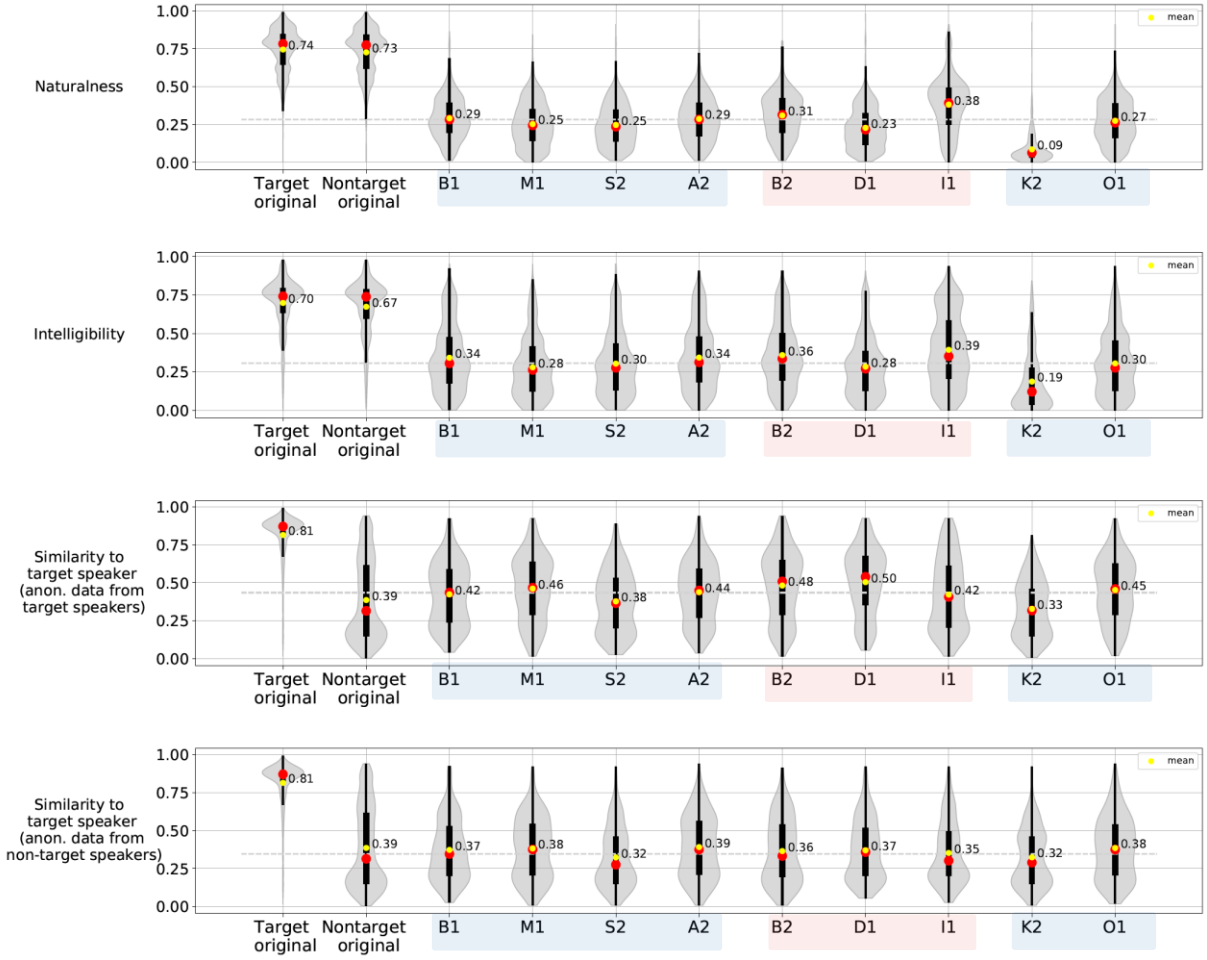
- EER & DET curves



Subjective evaluation results: violin plots



Subjective evaluation results: violin plots



with rank-based per-listener score normalization

signal-processing based

x-vector based related to B1

Subjective evaluation – Part 1

Details about protocol

Trial 1	Target original	Speaker A		Utt. m
Trial 2	Target anony.	Speaker A	Anony. system 1	Utt. n
Trial 3	Non-target original	Speaker L		Utt. k
Trial 4	Non-target anony.	Speaker L	Anony. system 1	Utt. g
Trial 5	Target original	Speaker B		Utt. o
Trial 6	Target anony.	Speaker B	Anony. system 2	Utt. q
Trial 7	Non-target original	Speaker K		Utt. r
Trial 8	Non-target anony.	Speaker K	Anony. system 2	Utt. k
Trial 9	Target original	Speaker C		Utt. w
Trial 10	Target anony.	Speaker C	Anony. system 3	Utt. z
Trial 11	Non-target original	Speaker P		Utt. v
Trial 12	Non-target anony.	Speaker P	Anony. system 3	Utt. u
…				
Trial *N*	…	…		…

Sampling shuffling

Trial 12				
Trial 10				
Trial 9				
Trial 11				
Trial 1				
Trial 4				
Trial 2				
Trial 3				
…				
…				

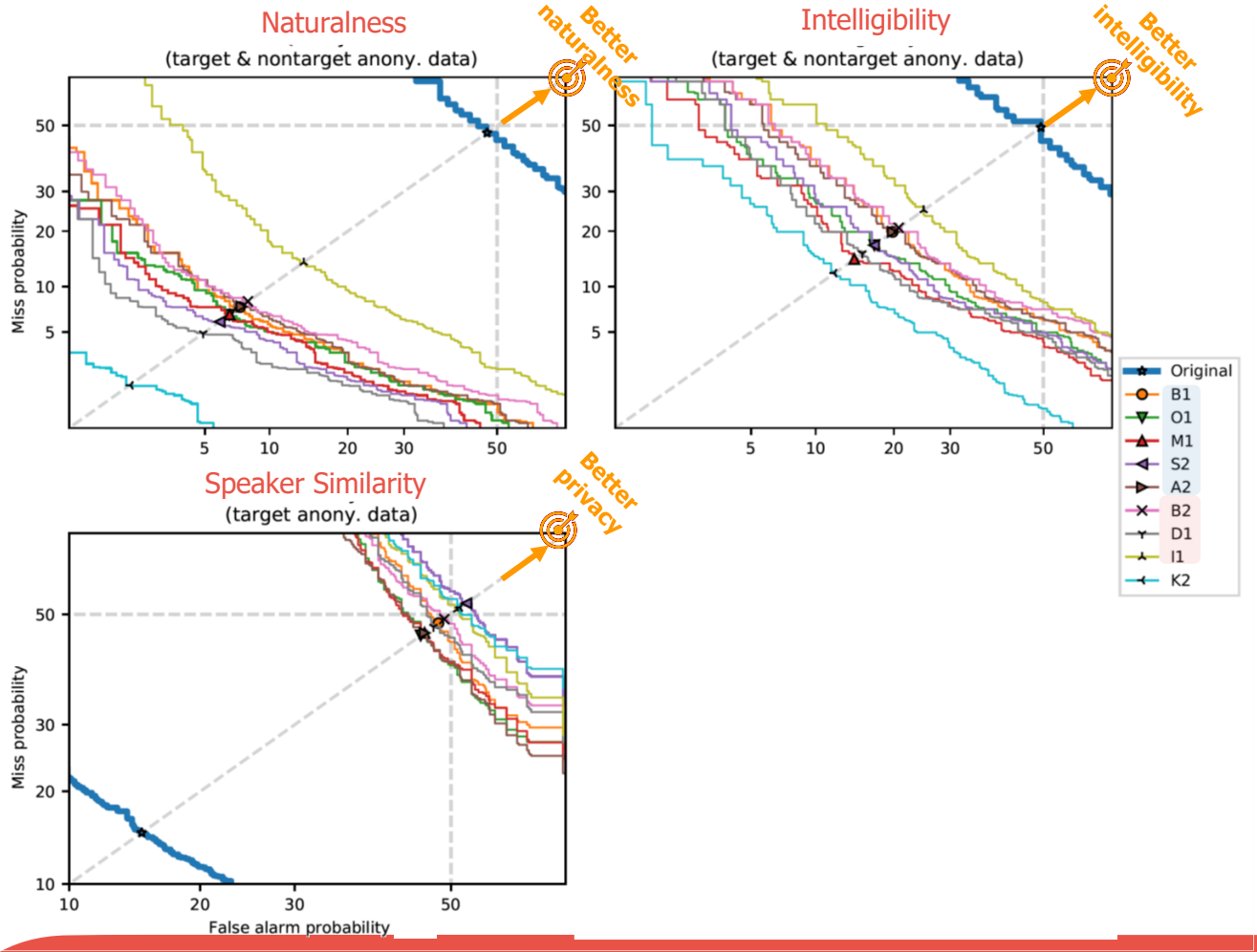
36 trials/ session

N=48,600 4 cases 30 speakers 9 systems 15 uttr.

48,600 trials → 1,350 sessions

3 test sets

Subjective evaluation results: DET-curves

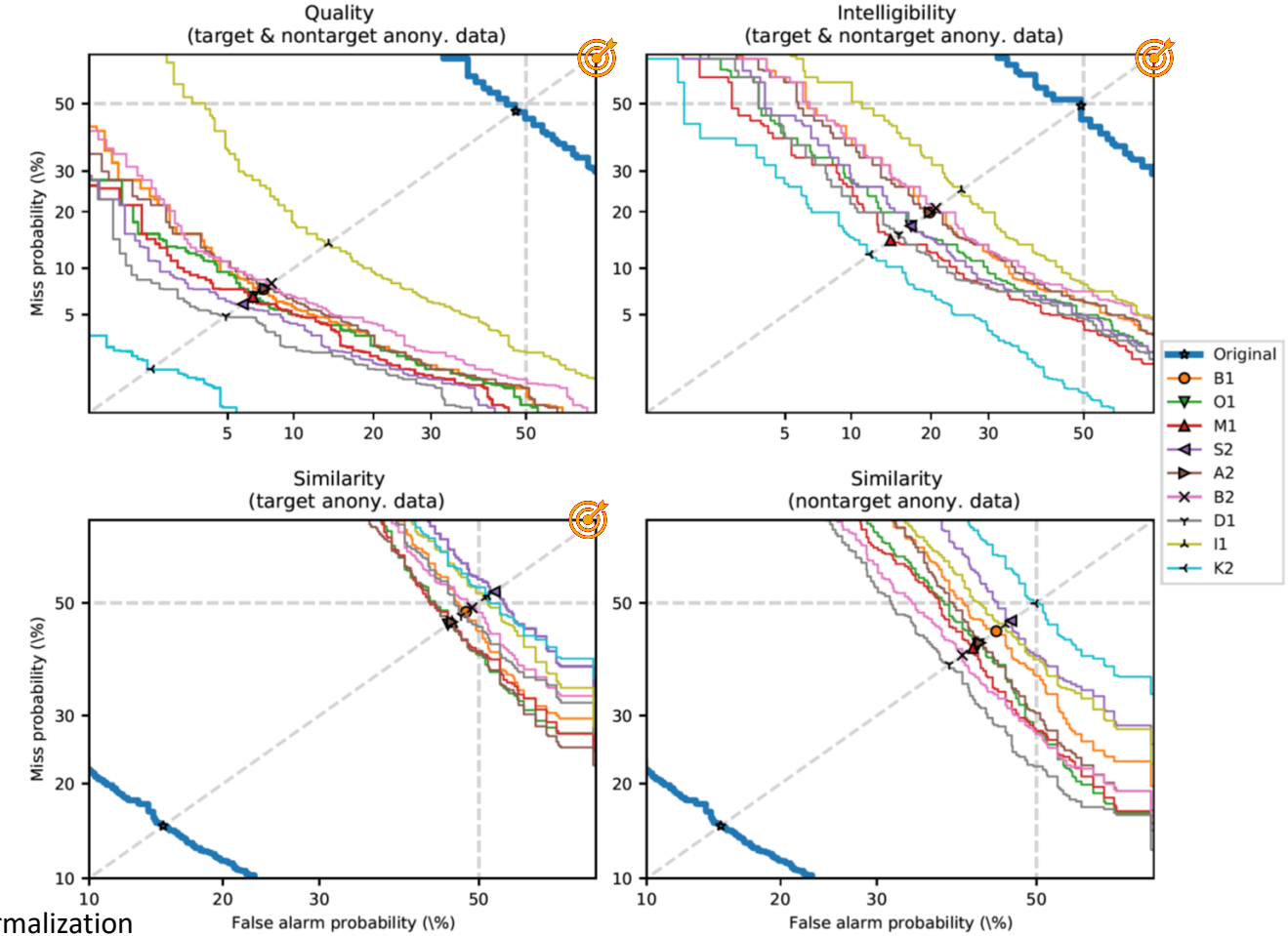


with rank-based per-listener score normalization

signal-processing based

x-vector based related to B1

Subjective evaluation results



* With ranked-based per-listener score normalization

Significance tests: naturalness & intelligibility

	Tar	Non-tar	B1	O1	M1	S2	A2	B2	D1	I1	K2
Tar		$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$
Non-tar	$\ll 0.01$		$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$
B1	$\ll 0.01$	$\ll 0.01$		0.017	$\ll 0.01$	$\ll 0.01$	0.562	0.0027	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$
O1	$\ll 0.01$	$\ll 0.01$	0.017		0.0019	$\ll 0.01$	0.081	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$
M1	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.0019		0.398	$\ll 0.01$	$\ll 0.01$	0.0003	$\ll 0.01$	$\ll 0.01$
S2	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.398		$\ll 0.01$	$\ll 0.01$	0.0026	$\ll 0.01$	$\ll 0.01$
A2	$\ll 0.01$	$\ll 0.01$	0.562	0.081	$\ll 0.01$	$\ll 0.01$		0.0006	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$
B2	$\ll 0.01$	$\ll 0.01$	0.0027	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.0006		$\ll 0.01$	$\ll 0.01$	$\ll 0.01$
D1	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.0003	0.0026	$\ll 0.01$	$\ll 0.01$		$\ll 0.01$	$\ll 0.01$
I1	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$		$\ll 0.01$
K2	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	

Table 3: *Naturalness*

	Tar	Non-tar	B1	O1	M1	S2	A2	B2	D1	I1	K2
Tar		$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$
Non-tar	$\ll 0.01$		$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$
B1	$\ll 0.01$	$\ll 0.01$		0.0003	$\ll 0.01$	$\ll 0.01$	0.866	0.043	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$
O1	$\ll 0.01$	$\ll 0.01$	0.0003		0.0053	0.764	0.0001	$\ll 0.01$	0.048	$\ll 0.01$	$\ll 0.01$
M1	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.0053		0.013	$\ll 0.01$	$\ll 0.01$	0.421	$\ll 0.01$	$\ll 0.01$
S2	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.764	0.013		$\ll 0.01$	$\ll 0.01$	0.101	$\ll 0.01$	$\ll 0.01$
A2	$\ll 0.01$	$\ll 0.01$	0.866	0.0001	$\ll 0.01$	$\ll 0.01$		0.064	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$
B2	$\ll 0.01$	$\ll 0.01$	0.043	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.064		$\ll 0.01$	0.0059	$\ll 0.01$
D1	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.048	0.421	0.101	$\ll 0.01$	$\ll 0.01$		$\ll 0.01$	$\ll 0.01$
I1	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.0059	$\ll 0.01$		$\ll 0.01$
K2	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	

Table 4: *Intelligibility*

Significance tests: similarity

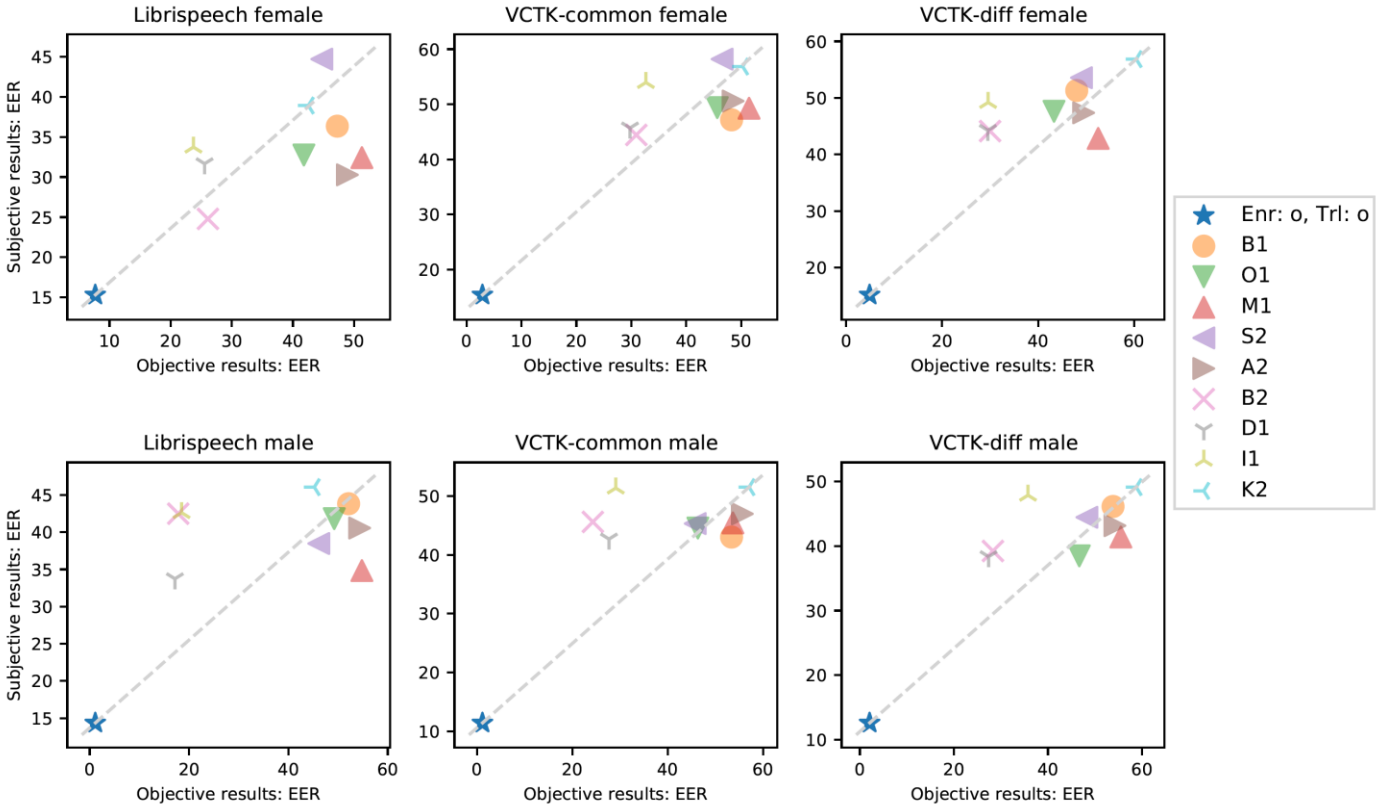
	Tar	Non-tar	B1	O1	M1	S2	A2	B2	D1	I1	K2
Tar		$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$
Non-tar	$\ll 0.01$		$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.510	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.0002	0.0019
B1	$\ll 0.01$	$\ll 0.01$		0.027	0.0057	0.0007	0.258	$\ll 0.01$	$\ll 0.01$	0.660	$\ll 0.01$
O1	$\ll 0.01$	$\ll 0.01$	0.027		0.483	$\ll 0.01$	0.266	0.042	0.0002	0.018	$\ll 0.01$
M1	$\ll 0.01$	$\ll 0.01$	0.0057	0.483		$\ll 0.01$	0.077	0.167	0.0030	0.0047	$\ll 0.01$
S2	$\ll 0.01$	0.510	0.0007	$\ll 0.01$	$\ll 0.01$		$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.0087	0.0004
A2	$\ll 0.01$	$\ll 0.01$	0.258	0.266	0.077	$\ll 0.01$		0.0020	$\ll 0.01$	0.135	$\ll 0.01$
B2	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.042	0.167	$\ll 0.01$	0.0020		0.140	0.0001	$\ll 0.01$
D1	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.0002	0.0030	$\ll 0.01$	$\ll 0.01$	0.140		$\ll 0.01$	$\ll 0.01$
I1	$\ll 0.01$	0.0002	0.660	0.018	0.0047	0.0087	0.135	0.0001	$\ll 0.01$		$\ll 0.01$
K2	$\ll 0.01$	0.0019	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.0004	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	

Table 1: *Similarity target*

	Tar	Non-tar	B1	O1	M1	S2	A2	B2	D1	I1	K2
Tar		$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$
Non-tar	$\ll 0.01$		0.440	0.171	0.241	0.0004	0.038	0.653	0.634	0.411	0.0005
B1	$\ll 0.01$	0.440		0.327	0.379	0.0002	0.119	0.490	0.839	0.102	0.0003
O1	$\ll 0.01$	0.171	0.327		0.943	$\ll 0.01$	0.546	0.122	0.267	0.012	$\ll 0.01$
M1	$\ll 0.01$	0.241	0.379	0.943		$\ll 0.01$	0.478	0.167	0.293	0.018	$\ll 0.01$
S2	$\ll 0.01$	0.0004	0.0002	$\ll 0.01$	$\ll 0.01$		$\ll 0.01$	0.0047	0.0004	0.033	0.977
A2	$\ll 0.01$	0.038	0.119	0.546	0.478	$\ll 0.01$		0.036	0.078	0.0016	$\ll 0.01$
B2	$\ll 0.01$	0.653	0.490	0.122	0.167	0.0047	0.036		0.603	0.467	0.0066
D1	$\ll 0.01$	0.634	0.839	0.267	0.293	0.0004	0.078	0.603		0.187	0.0006
I1	$\ll 0.01$	0.411	0.102	0.012	0.018	0.033	0.0016	0.467	0.187		0.038
K2	$\ll 0.01$	0.0005	0.0003	$\ll 0.01$	$\ll 0.01$	0.977	$\ll 0.01$	0.0066	0.0006	0.038	

Table 2: *Similarity non-target*

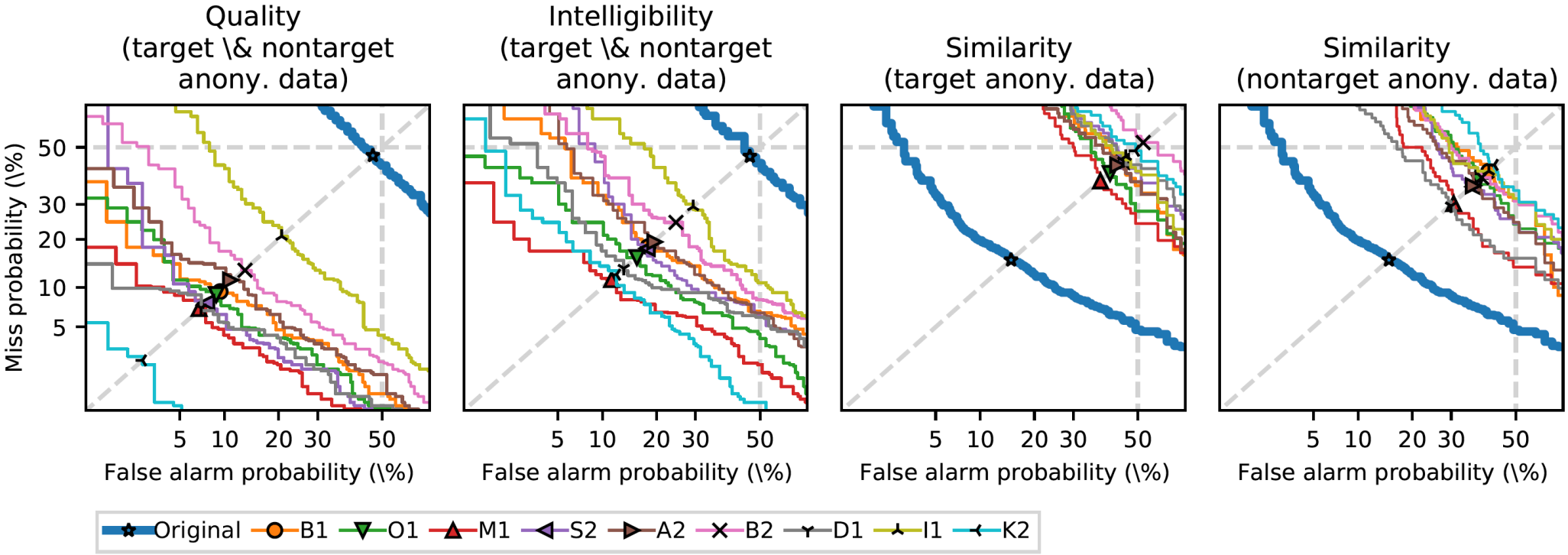
Subjective vs objective evaluation results: EER



- Positive correlation between subjective and objective scores

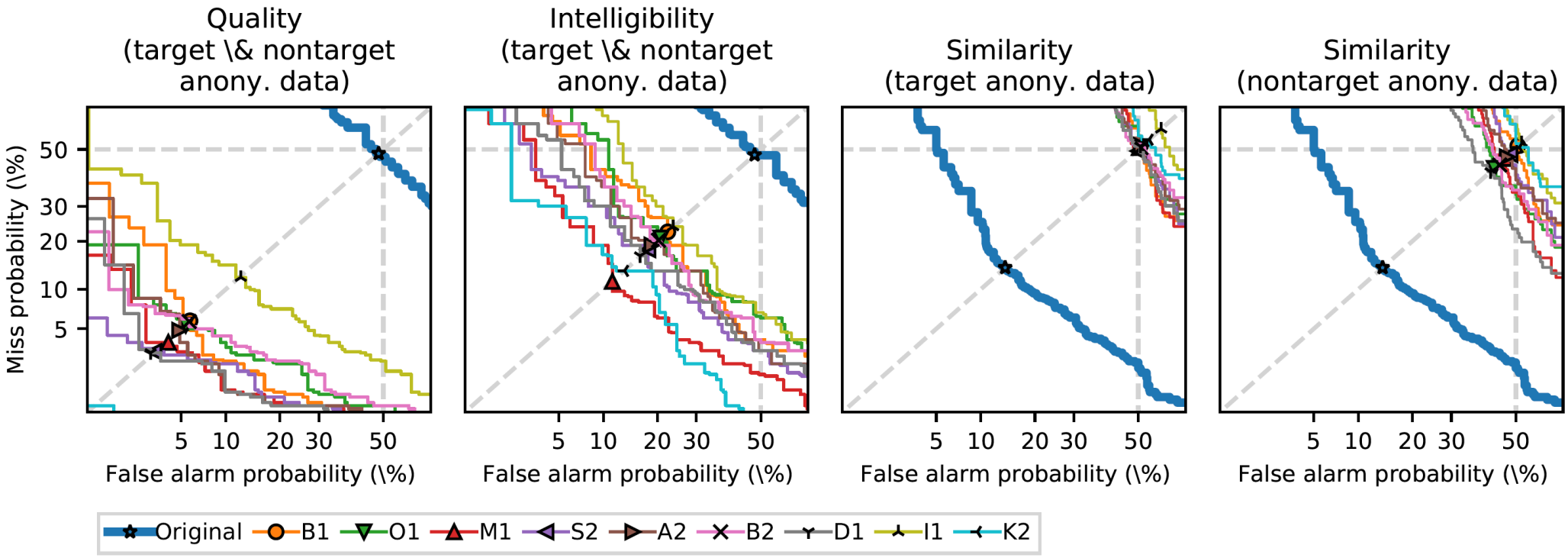
* With ranked per-listener normalization of scores

DET curves*: LibriSpeech



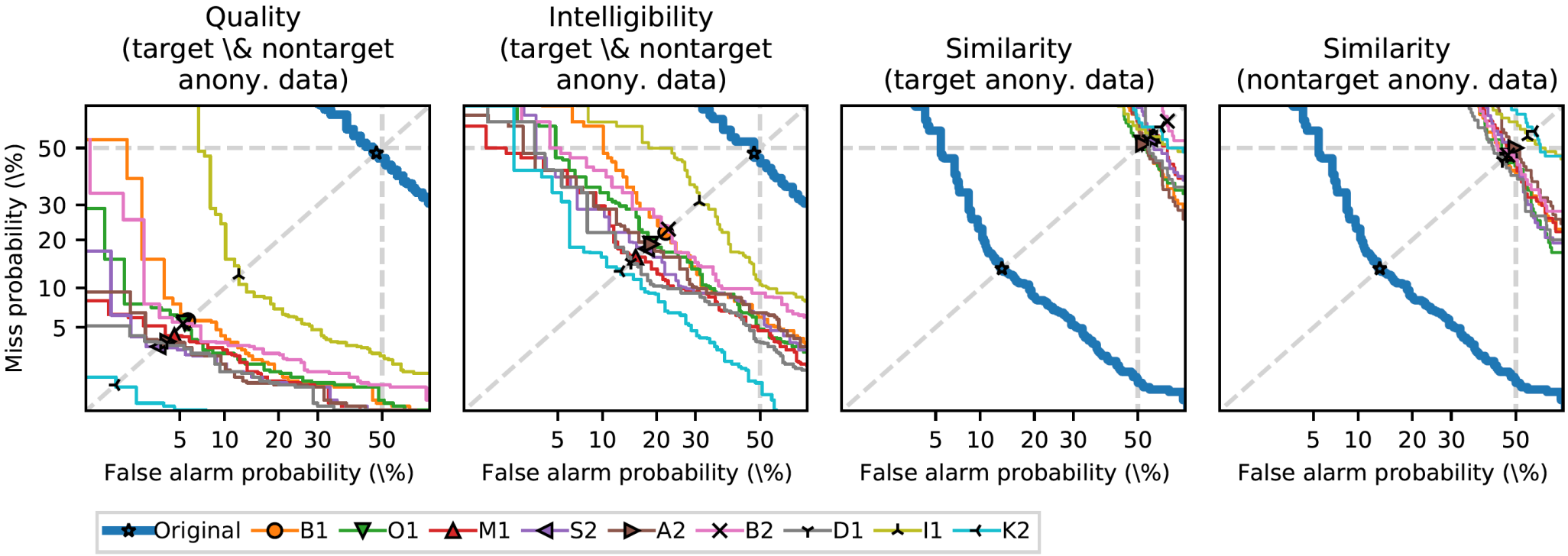
* With ranked normalization

DET curves*: VCTK-common



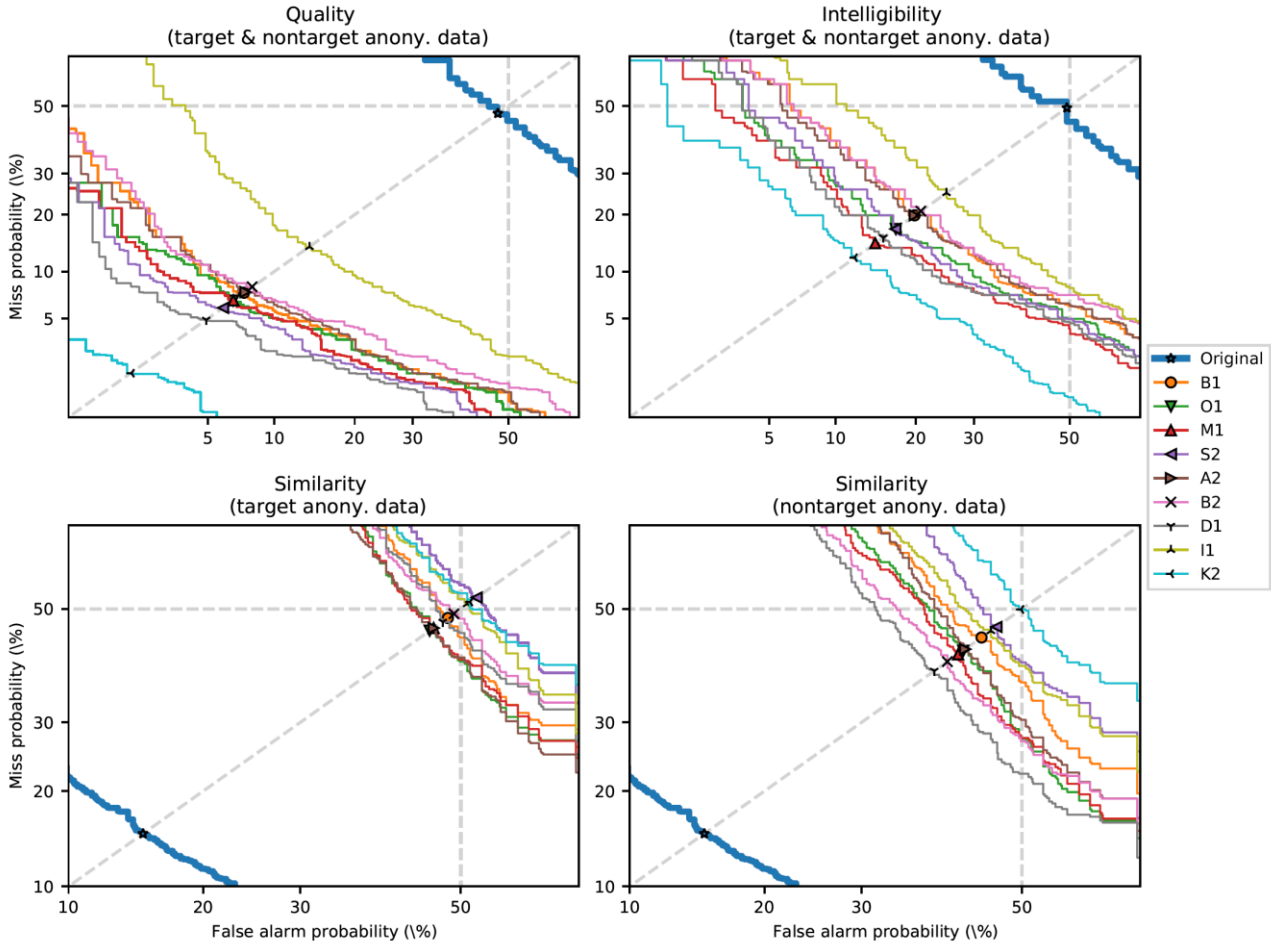
* With ranked normalization

DET curves*: VCTK-different



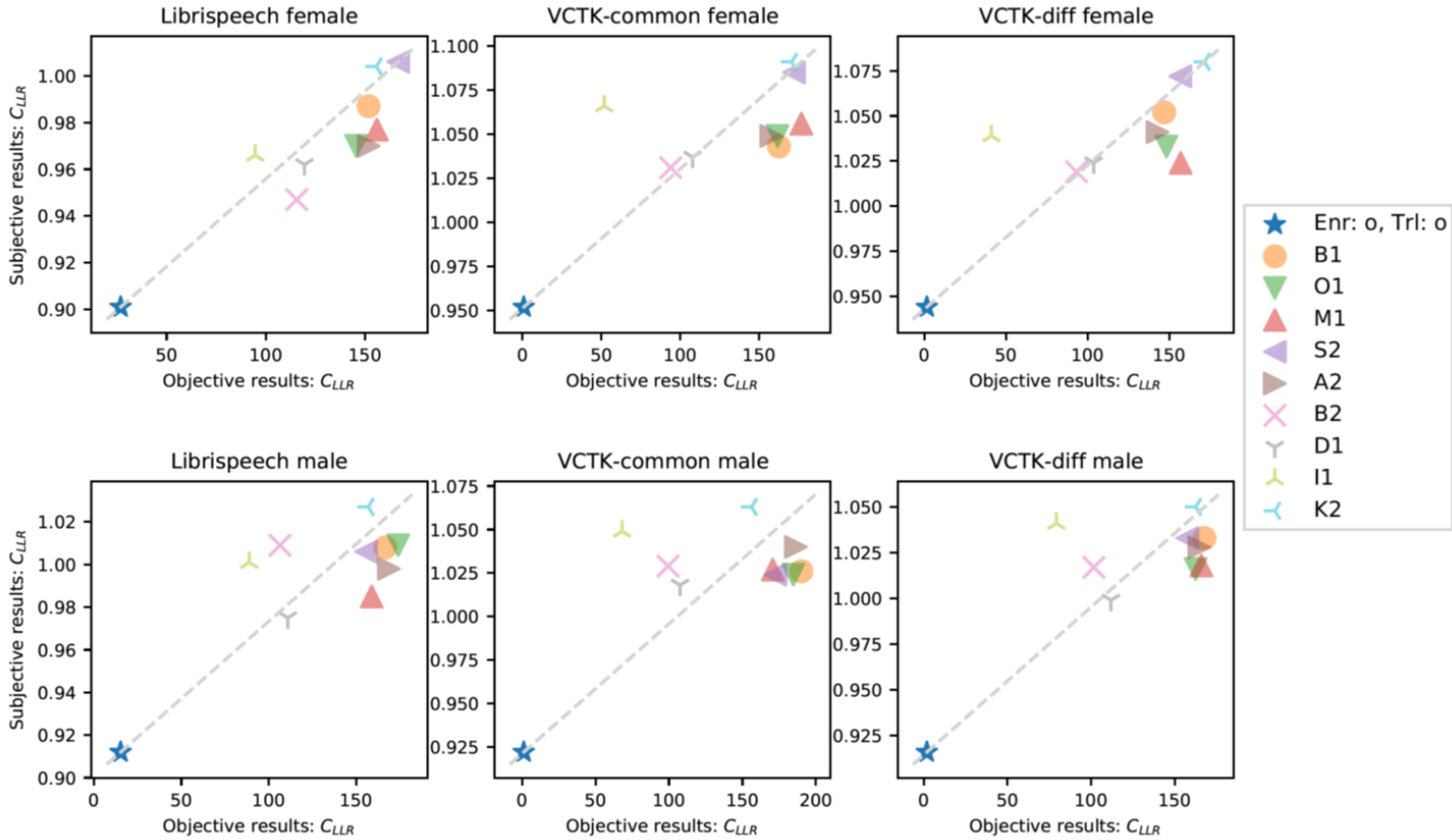
* With ranked normalization

DET curves*: Pooled



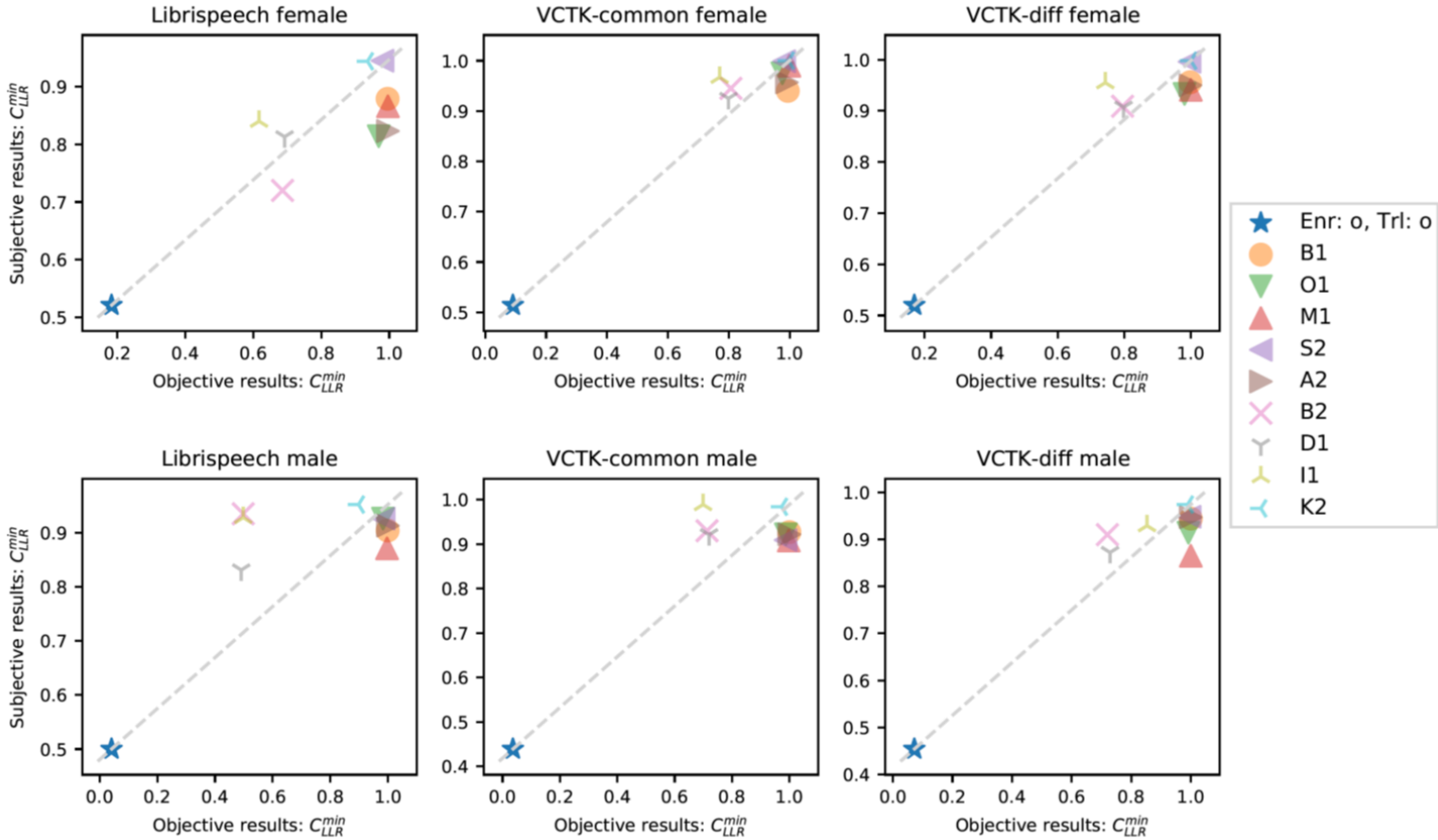
* With ranked normalization

Objective vs subjective: Cllr



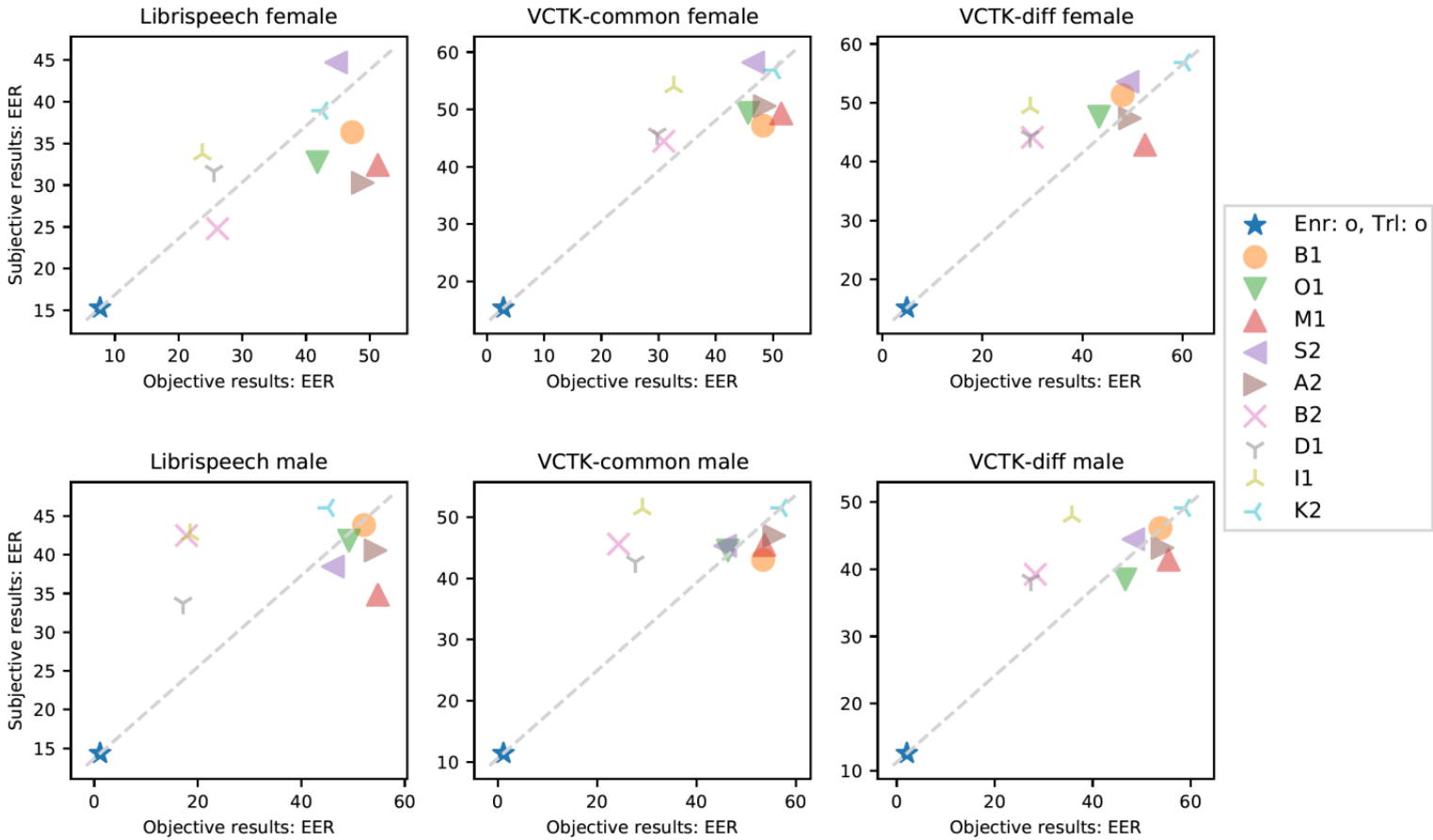
* With ranked normalization

Objective vs subjective: Cllr-min



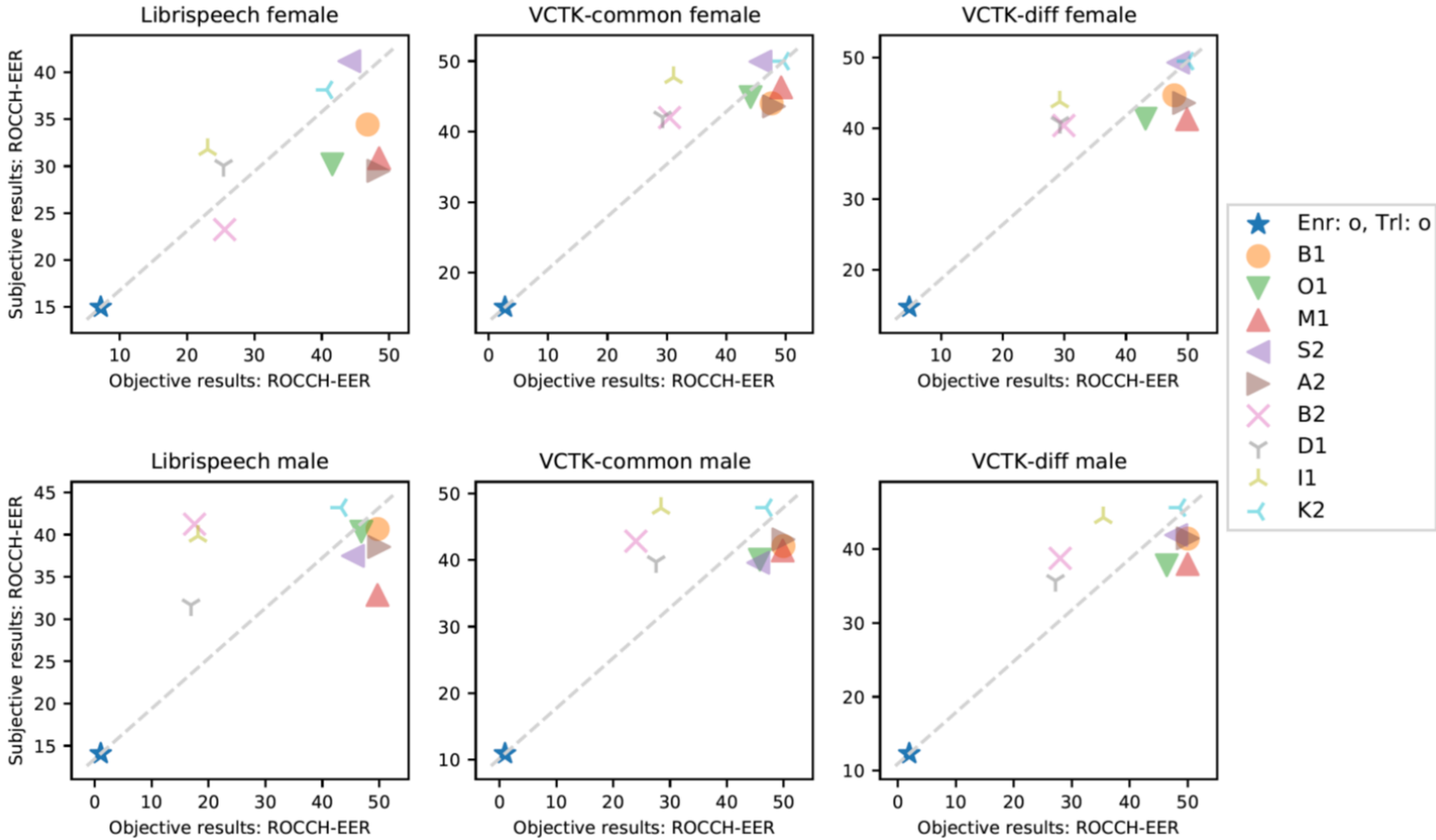
* With ranked normalization

Objective vs subjective: EER



* With ranked normalization

Objective vs subjective: ROCH-EER



* With ranked normalization