Engineering and Physical Sciences Research Council

THE 3RD CLARITY PREDICTION CHALLENGE: A MACHINE LEARNING CHALLENGE FOR HEARING AID INTELLIGIBILITY PREDICTION

Jon Barker[1], *Michael A. Akeroyd*[2], *Trevor J. Cox*[3], *John F. Culling*[4], *Jennifer Firth*[2], *Simone Graetzer*[3], *Graham Naylor*[2]

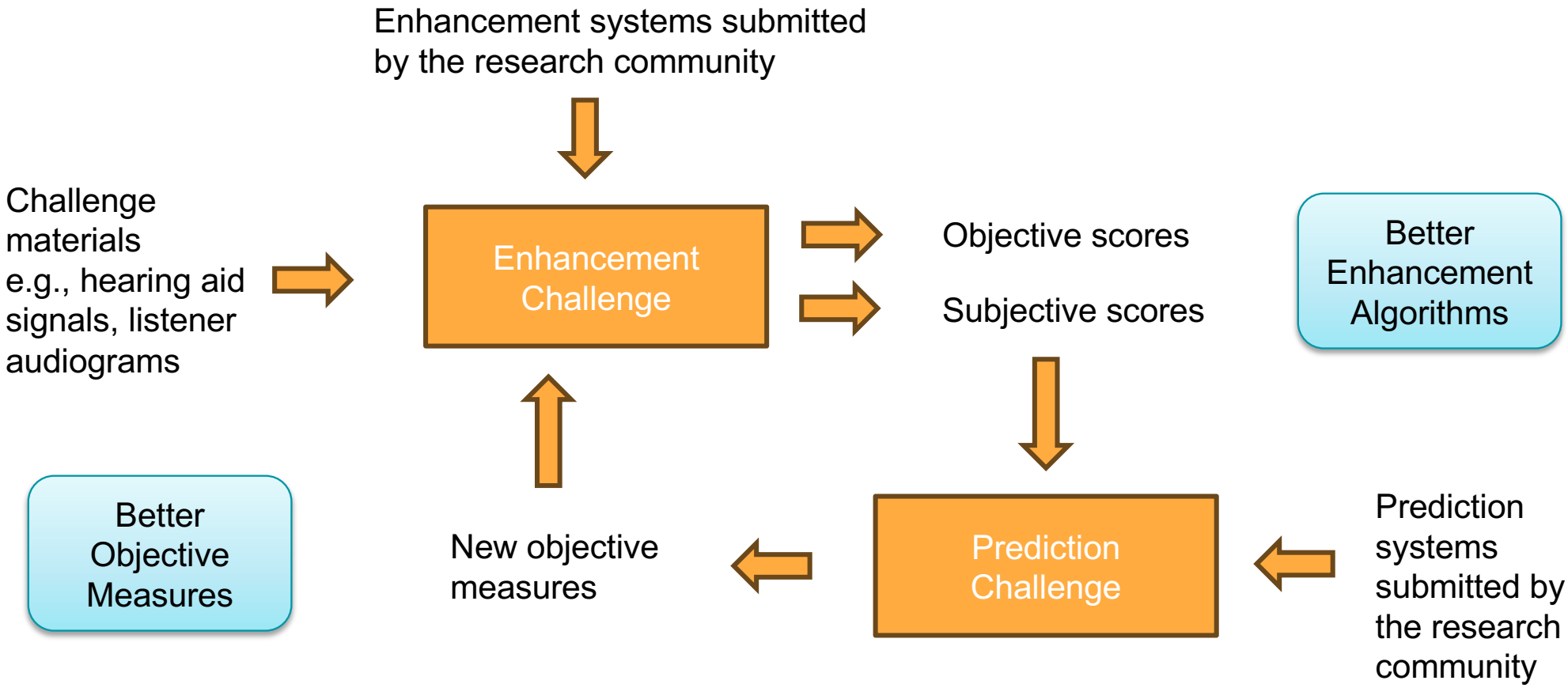
[1] Department of Computer Science, University of Sheffield, UK
[2] School of Medicine, University of Nottingham, UK
[3] Acoustics Research Centre, University of Salford, UK
[4] School of Psychology, Cardiff University, UK

Clarity-2025, 22nd August 2025, Delft, The Netherlands

Motivation

- **Understanding speech in noise is a major challenge for hearing-aid users.**
- New speech processing algorithms are needed.
- Great potential in recent low-latency DNN-based single- and multi-channel speech processing techniques...
- …but application of machine learning approaches is hindered by the lack of sufficiently reliable **objective intelligibility measures**.
- 6-year funding from UK government to run a series of open machine learning challenges for intelligibility enhancement and intelligibility prediction - the Clarity Project.

Challenge methodology



Hearing aid speech enhancement challenges:

- 1st Enhancement Challenge, CEC1, 2021
- 2nd Enhancement Challenge, CEC2, 2022
- ICASSP SP Enhancement Challenge 2022-3
  - Speech intelligibility and quality
- 3rd Enhancement Challenge, CEC3, 2024-5

Speech intelligibility prediction challenges

- 1st Prediction Challenge, CPC1, 2021-2
- 2nd Prediction Challenge, CPC2, 2023
- 3rd Prediction Challenge, CPC3, 2025

Results today

The Clarity Prediction Challenge

Participants are given:

- A **hearing aid output signal** that has arisen from processing **speech in noise**
- The **hearing-impairment severity of the listener** who is using the hearing aid

They must predict:

- The **percentage of words that the listener will correctly recognise**.

Systems are evaluated by computing the RMS prediction error over a large number of signal/listener pairs across a variety of hearing aid algorithms.

Engineering and Physical Sciences Research Council

3rd Clarity Prediction Challenge

The Task and Materials

Clarity-2025, 22nd August 2025, Delft, The Netherlands

The Clarity Challenge Plan

Round 1 (2021)

- Simple stationary scenes.
- Domestic living rooms with speech target and a static domestic noise source.

Round 2 (2022-23)

- Scenes with multiple noise sources
- Listener head movements

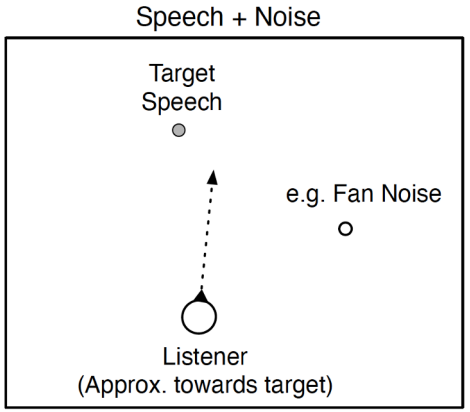
Round 3 (2024-25)

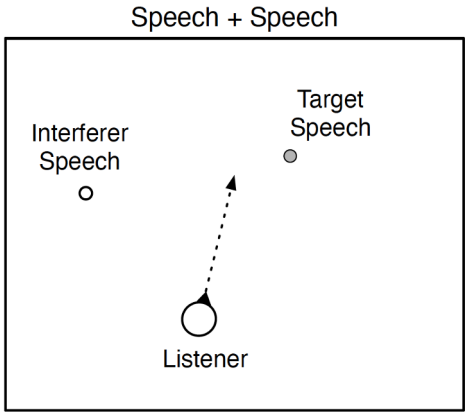
- Fully dynamic scenes.
- Real background; Real hearing aid signals

Round 1

Target speech in presence of a single interferer.

- Target source is within ±30° inclusive in front of listener at >1 m distance and at same height.
  - Human speech directivity and oriented towards the listener.
- Interferer anywhere, except within 1 m of a wall and omnidirectional.
  - Domestic noise source - kettle, washing machine etc
  - Continuous speech stream





Simulated hearing aid inputs

- We use the OlHeaD-HRTF Database (Denk et al., 2018) to simulate input signals for a **3-mic** behind-the-ear (BTE) hearing aid.

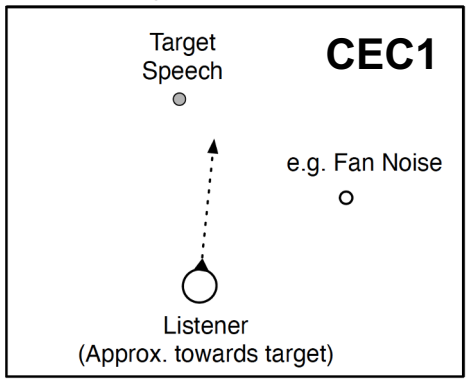
- i.e., the hearing aid algorithms are provided with six channels as input.

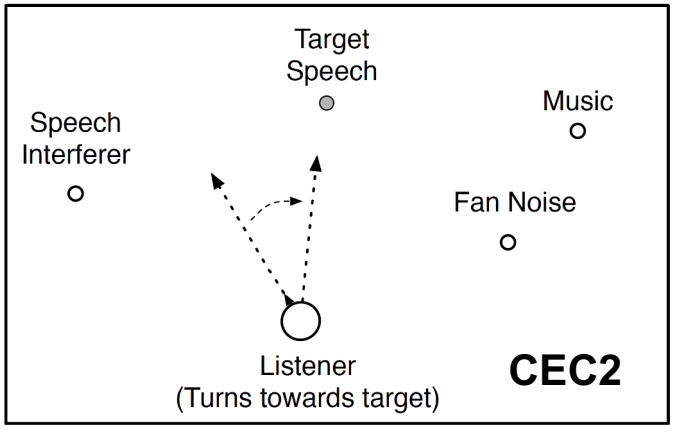
F. Denk, S.M.A. Ernst, S.D. Ewert and B. Kollmeier, (2018): Adapting hearing devices to the individual ear acoustics: Database and target response correction functions for various device styles. Trends in Hearing, vol 22, p. 1-19. DOI:10.1177/2331216518779313

Round 2

Key differences in round 2

- Scenes have **two or three interferers**.
- Interferers are any combination of **speech, noise and music**
- The listener **turns their head** towards the target speaker
- Variability in target speaker onset time
- **Target speaker** is identified by familiarity (4 clean target speaker utterances for learning target id)
- Better Ear SNR ranges from **-12 dB to 6 dB**,
  (cf -6 dB to 6 dB for CEC1)

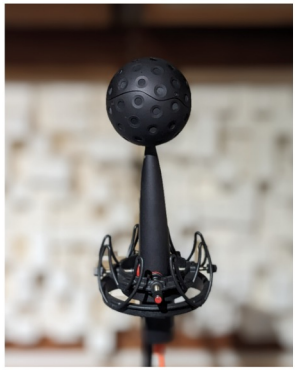




Round 3

Task 1 - real impulse responses

As CEC2 but using measured 6th order ambisonic room impulse responses for development and evaluation data.



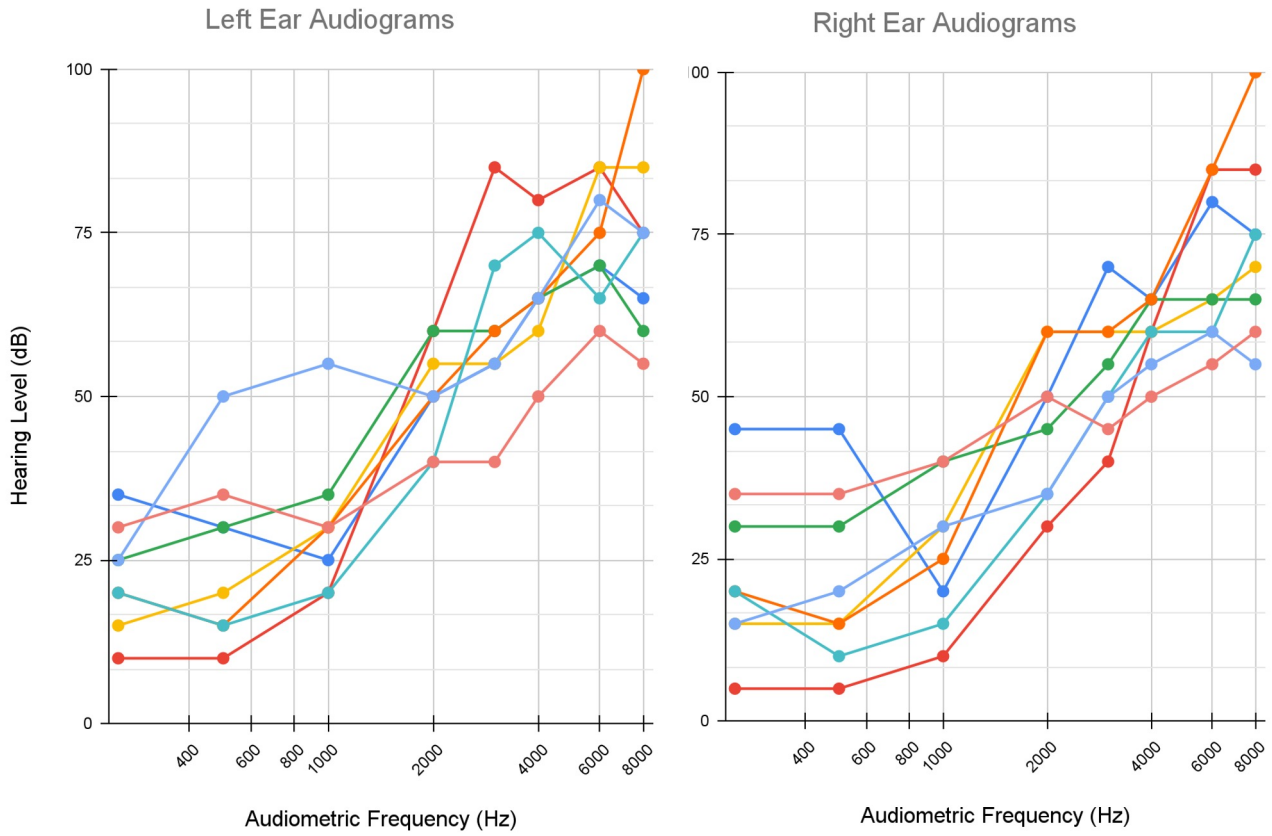
Task 2 - real hearing aid mics

As CEC2 but with scenes played in a real room over loudspeakers and recorded via hearing aid shells.

Task 3 - real noise backgrounds

As CEC2 but using real ambisonic background recordings in place of point source interferers and targets real impulse responses for adding the targets. “Out and about”

Listener Characteristics



Round 1 - 28 listeners.
Round 2 - 17 listeners.
Round 3 - 17 listeners.

Mean left ear = 43 dB
Mean right ear = 40 dB

Mean better ear = 39 dB
Mean worse ear = 45 dB

Mean better-worse difference = 6 dB

CEC1 Enhancement Systems

System	Beamforming	Noise Removal	Hearing Loss Compensation
E002	RLS adaptive	MC Conv-TasNet	Linear, fitting formula
E003		Conv-TasNet	Linear, fitting formula
E004		2D CNN + LSTM, WPE	Baseline system
E005		Binaural Conv-Tasnet	Baseline system
E007	MVDR	Conv-TasNet	Linear, NN-optimised
E009		MC Conv-TasNet	Linear, NN-optimised
E010		U-Net CNN	Linear, fitting formula
E013	MVDR		Linear, fitting formula but AGC
E016	Weighted LCMP		Linear, fitting formula
E018		2D CNN + LSTM, WPE	Dynamic EQ
E019	Weighted LCMP		MBDRC
E021	Weighted LCMP	DNN (Deep MFMBVDR)	MBDRC

CEC2 Enhancement Systems

Team	System	Enhancement	Amplification	Spkr. Extr.	Data+	HR
T01	E009	cf iNeuBe	NALR+DRC+trained	✓	-	-
T02	E031	DRC-NET	NALR	-	-	-
T03	E008	SDD-Net + S-DCCRN	trained	-	✓	-
T03	E008	ibid.	trained	-	✓	✓
T03	E008	ibid.	trained	-	-	✓
T03	E008	ibid.	trained	-	-	-
T04	E037	EaBNet + mod. MTFAA	POGO II + trained	-	-	-
T04	E022	ibid.	POGO II	-	-	-
T05	E024	SuDoRM-RF	PCS	-	-	✓
T05	E024	ibid.	PCS	-	-	-
T06	E036	TCN-conformer	NALR	✓	-	-
T06	E038	TCN	NALR	✓	-	-
T07	E032	Extr-DenseUNet	trained	✓	-	-
-	Baseline	-	NALR	-	-	-
-	None	-	-	-	-	-

Spkr. Extr. = Used speaker extraction;
Data+ = Augmented training data; HR = used head-rotation signal

Hearing Aid output samples

	Good	Fair	Poor
S08502 / L0106			

“And it is the most incredible thing”

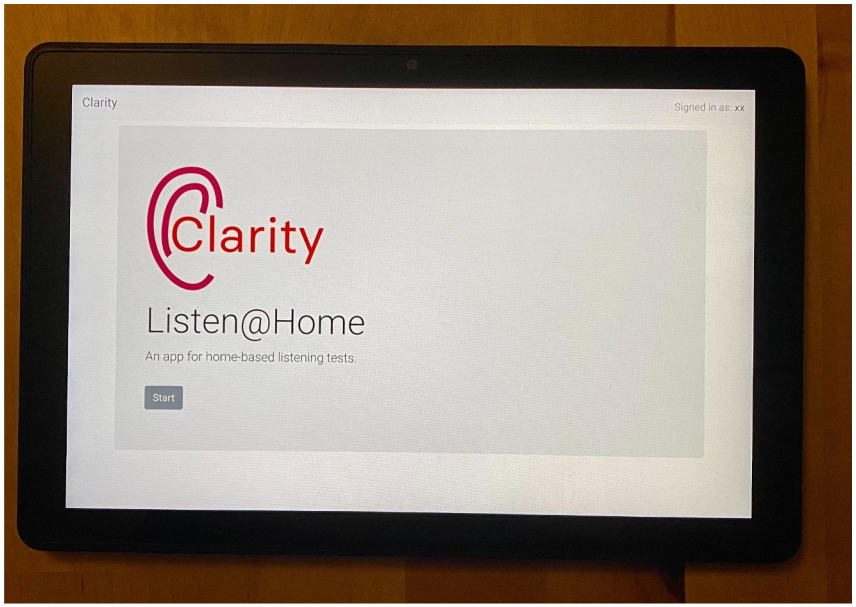
	Good	Fair	Poor
S08501 / L0104			

“Roll over and repeat on the other side”

Listen@Home

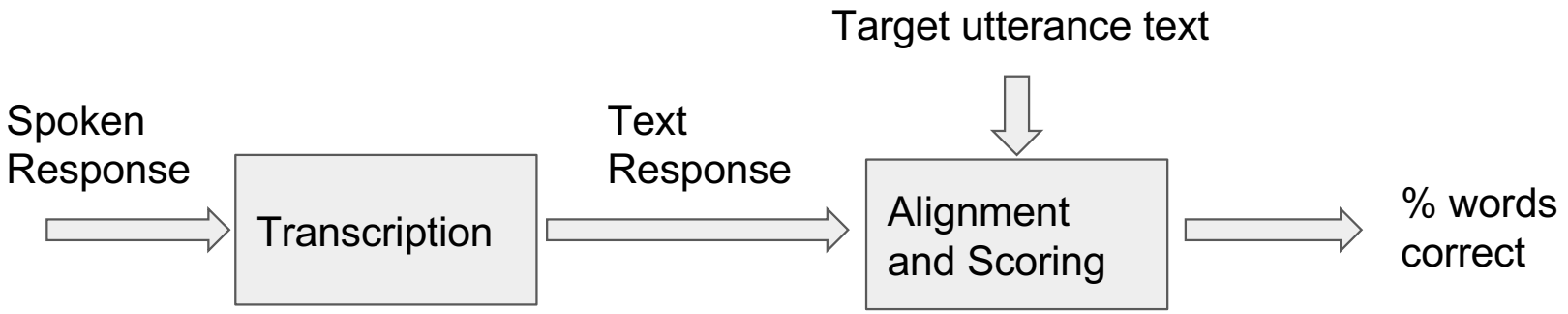
Lenovo 10e chromebook tablet and Sennheiser PC-8 headphone+mic headset. Posted to every participant's home.

Participants listen to processed speech-in-noise and then respeak the sentence that they've heard.



Intelligibility Scoring

- The target signals are short sentences, 7-10 words long spoken by British English speakers (Graetzer, et al., 2022)
- Per sentence intelligibility is measured as the percentage of words heard correctly.



- e.g., Target: She **did not return to** **land** again.

  Response: He **did not return to** the **land**.

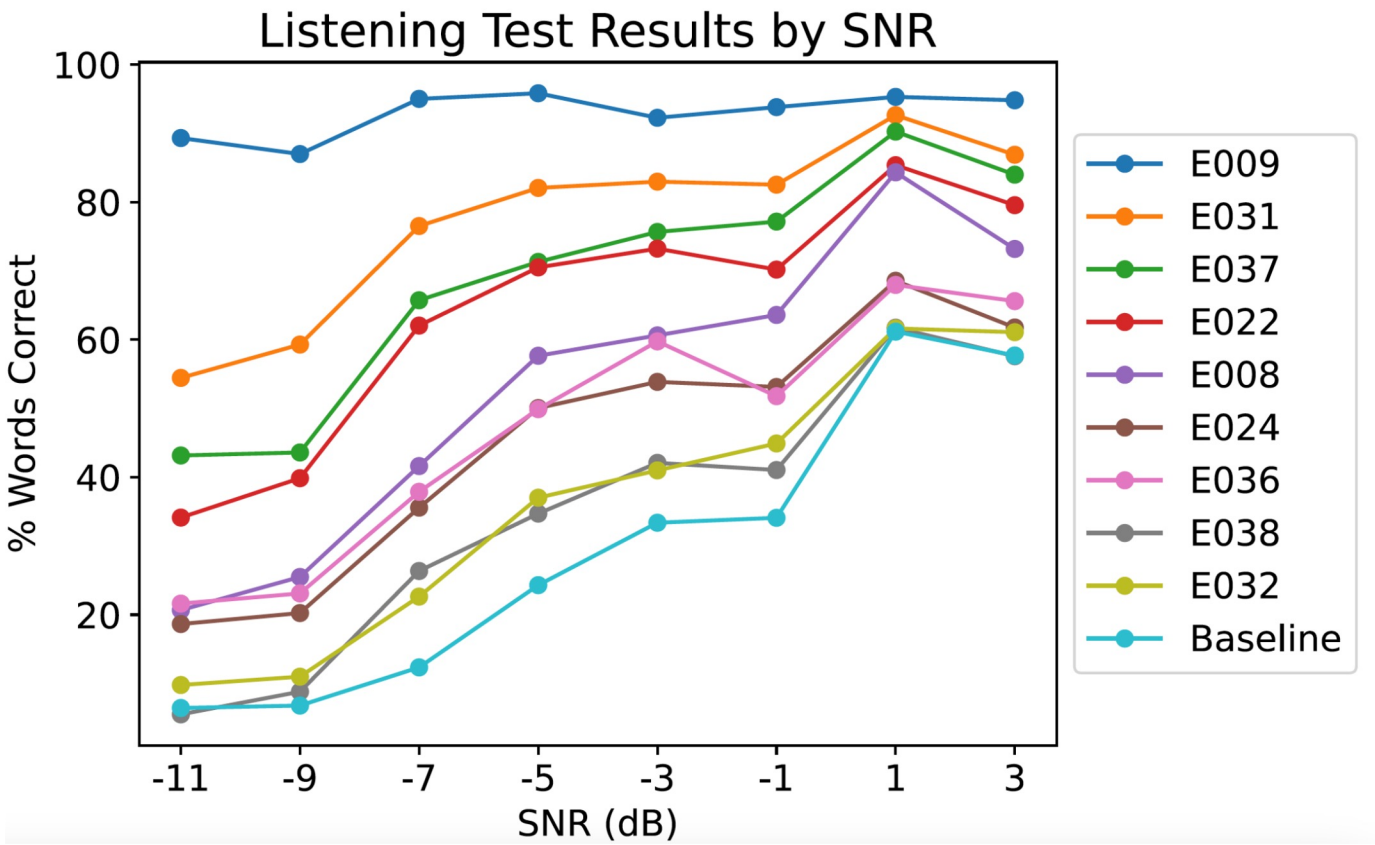
  Would score 5 out of 7 correct. (71%)

CEC2 Listening test scores

Team	System	Enhancement	Amplification	Spkr. Extr.	Data+	HR	HASPI	Listener
T01	E009	cf iNeuBe	NALR+DRC+trained	✓	-	-	**0.966**	93.2
T02	E031	DRC-NET	NALR	-	-	-	**0.801**	76.5
T03	E008	SDD-Net + S-DCCRN	trained	-	✓	-	**0.800**	-
T03	E008	ibid.	trained	-	✓	✓	0.794	-
T03	E008	ibid.	trained	-	-	✓	0.784	52.6
T03	E008	ibid.	trained	-	-	-	0.777	-
T04	E037	EaBNet + mod. MTFAA	POGO II + trained	-	-	-	**0.775**	68.4
T04	E022	ibid.	POGO II	-	-	-	0.721	65.5
T05	E024	SuDoRM-RF	PCS	-	-	✓	**0.630**	44.8
T05	E024	ibid.	PCS	-	-	-	0.617	-
T06	E036	TCN-conformer	NALR	✓	-	-	**0.599**	45.6
T06	E038	TCN	NALR	✓	-	-	0.554	34.1
T07	E032	Extr-DenseUNet	trained	✓	-	-	**0.549**	35.3
-	Baseline	-	NALR	-	-	-	0.258	27.0
-	None	-	-	-	-	-	0.172	-

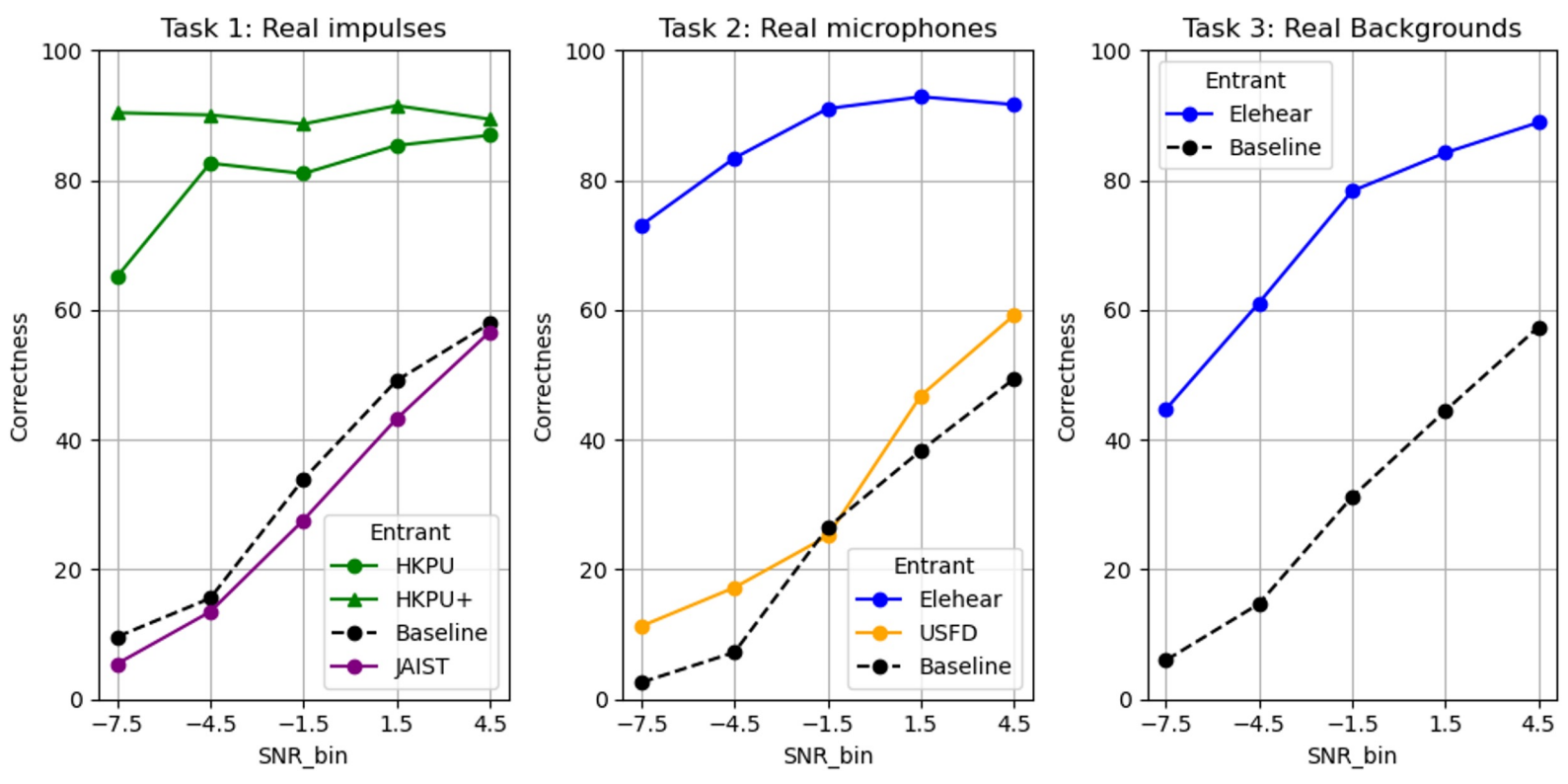
Spkr. Extr. = Used speaker extraction;
Data+ = Augmented training data; HR = used head-rotation signal

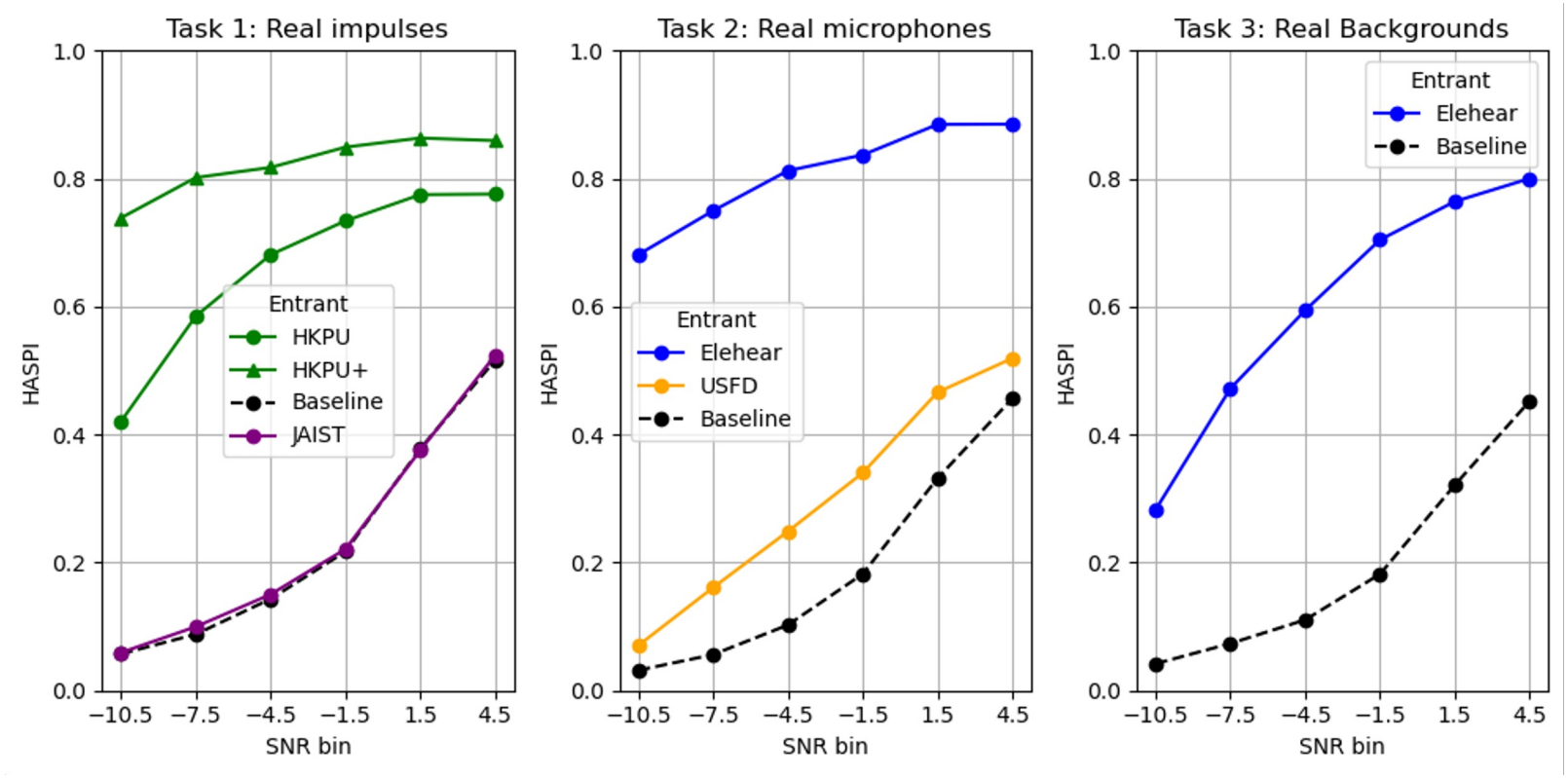
Performance vs SNR



CEC3 Enhancement Systems

System	Team	Task	Listener	HASPI
E001	T001	Task 1	39.3	0.241
E002	T002	Task 1	35.7	0.246
E003	T003	Task 1	82.5	0.665
E004	T003	Task 1	90.1	0.823
E005	T001	Task 2	30.0	0.193
E006	T004	Task 2	36.8	0.298
E007	T005	Task 2	88.7	0.806
E008	T001	Task 3	36.3	0.198
E009	T005	Task 3	76.7	0.806


Task 1: Real impulses
Correctness
SNR_bin
Entrant
HKPU
HKPU+
Baseline
JAIST
Task 2: Real microphones
Correctness
SNR_bin
Entrant
Elehear
USFD
Baseline
Task 3: Real Backgrounds
Correctness
SNR_bin
Entrant
Elehear
Baseline


Task 1: Real impulses
Task 2: Real microphones
Task 3: Real Backgrounds
HASPI
SNR bin
1.0
0.8
0.6
0.4
0.2
0.0
−10.5
−7.5
−4.5
−1.5
1.5
4.5
Entrant
HKPU
HKPU+
Baseline
JAIST
Elehear
USFD

Engineering and Physical Sciences Research Council

Clarity Prediction Challenge

Challenge Datasets and Rules

Clarity-2025, 22nd August 2025, Delft, The Netherlands

Training Data

- All the signal-listener pairs from CEC1 and CEC2
- 20,256 single-response pairs in total
- 20 different hearing-aid systems
- 34 listeners
- Ground truth listener scores made available for training.

Dev Data

- A subset of the CEC3 listening data
- 8 listeners
- 4 systems
- 926 single-response pairs in total
- Ground truth scores hidden, but remote evaluation via submission to ‘leaderboard’

Eval Data

- Remainder of CEC3 data
- 7674 single-response pairs in total
- 16 listeners (dev listeners + 8 more)
- 9 systems (dev systems + 5 more)
- Ground truth hidden and only one submission allowed

The Clarity Prediction Challenge

Participants are given:

- A **hearing aid output signal** that has arisen from processing **speech in noise**
- The **hearing-impairment severity of the listener** who is using the hearing aid
  - i.e. only know whether the impairment is mild, moderate or moderate-severe

They must predict:

- The **percentage of words that the listener will correctly recognise**.

Systems are evaluated by computing the RMS prediction error over a large number of signal/listener pairs across a variety of hearing aid algorithms.

Engineering and Physical Sciences Research Council

Clarity Prediction Challenge

Entries and Results

The Entrants

- We had **21 system submissions** arising from **14 separate teams**.
- Teams submitted technical papers which were reviewed to check compliance with the rules.
- Systems were classified as either **Intrusive or Non-intrusive** (i.e. whether they used the undistorted reference speech signal or not)

- Systems were scored by
  - computing the **RMS error** between the true and estimated sentence intelligibility
  - computing the **correlation** between the true and estimated sentence intelligibility.
  - RMS error is the main metric used for system ranking.

CPC2 Results

Paired t-test showed E011 significantly better than E002

Better-ear HASPI v2, Kates + Arehart, 2021

Always output the training set average

Team	System	Intr.	Non-Intr.	RMSE ↓	Corr ↑
T01	E011		X	**25.1 ± 0.8**	0.78
T02	E002		X	25.3 ± 0.8	0.77
T03	E009	X		25.4 ± 0.8	0.78
T04	E022	X		25.7 ± 0.9	0.77
T05	E023		X	26.4 ± 0.9	0.76
T05	E016		X	26.8 ± 0.9	0.75
T04	E025		X	27.9 ± 0.9	0.72
Base.	beHASPI	X		28.7 ± 1.0	0.70
T06	E003		X	31.1 ± 1.0	0.64
T06	E024		X	31.7 ± 1.0	0.62
T07	E015		X	35.0 ± 1.1	0.60
T08	E019		X	– ± –	–
T09	E020		X	39.8 ± 1.3	0.33
Base.	Prior		X	40.0 ± 1.3	–

CPC3 Results

Paired t-test showed E025 significantly better than E019

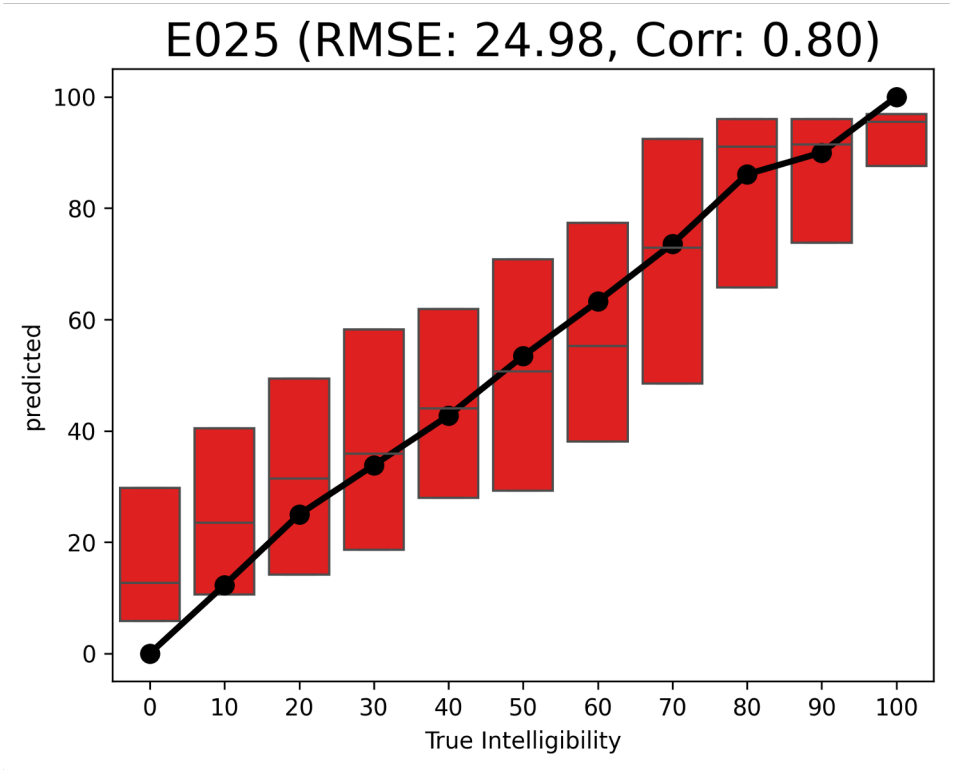
Better-ear HASPI v2, Kates + Arehart, 2021

Always output the training set average

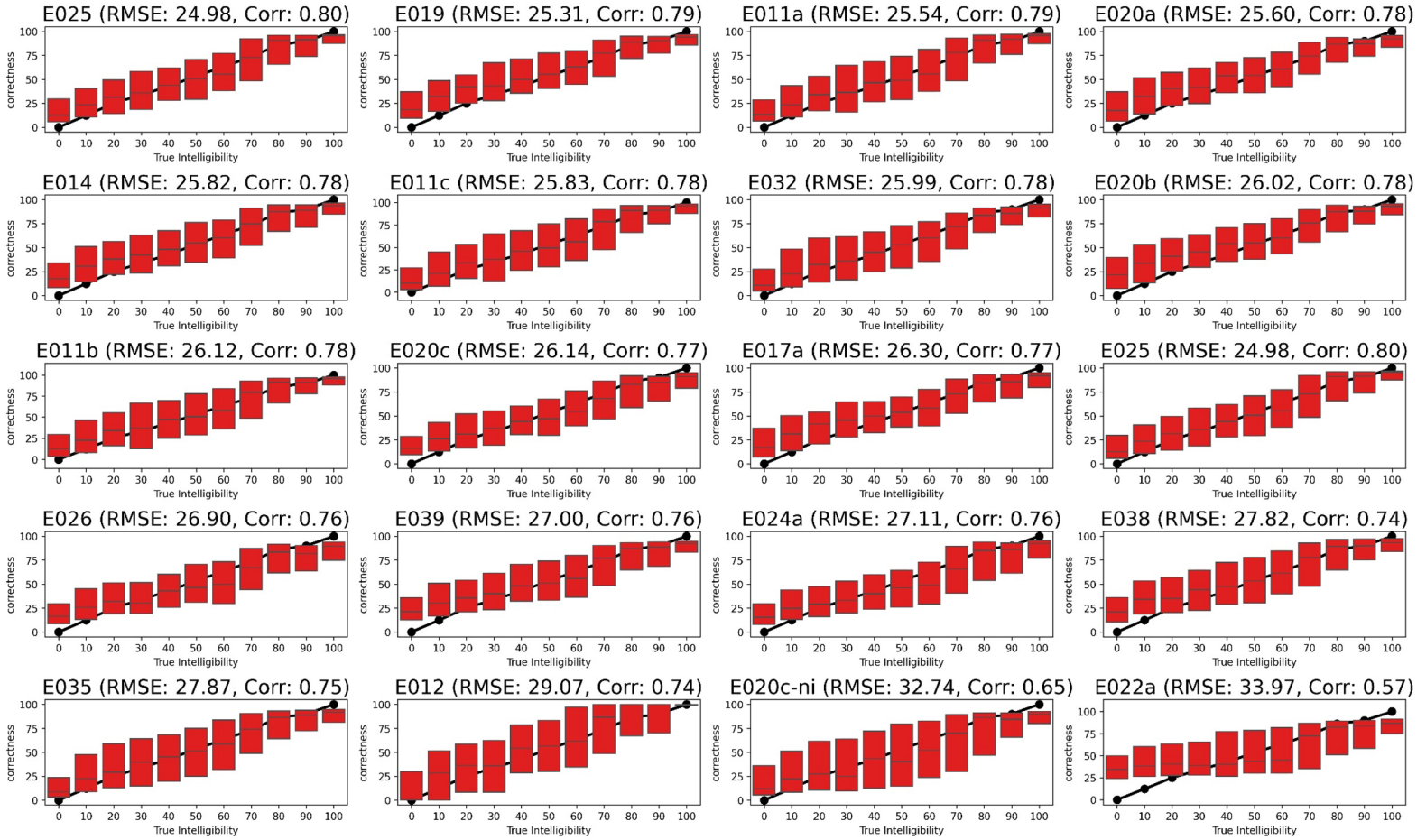
Team	System	Intr.	Non-Intr.	Dev RMSE ↓	Eval RMSE ↓	Corr ↑
T001	E025	X		22.36	**24.98 ± 0.29**	**0.80**
T002	E019		X	**21.87**	25.31 ± 0.29	0.79
T003	E011a		X	22.80	25.54 ± 0.29	0.79
T004	E020a	X		23.15	25.60 ± 0.29	0.78
T005	E014		X	22.95	25.82 ± 0.29	0.78
T003	E011c		X	22.89	25.83 ± 0.29	0.79
T006	E032	X (Text)		23.60	25.99 ± 0.30	0.78
T004	E020b	X		23.47	26.02 ± 0.30	0.78
T003	E011b		X	22.89	26.12 ± 0.30	0.78
T004	E020c	X		24.81	26.14 ± 0.30	0.77
T007	E017	X		24.05	26.30 ± 0.30	0.77
T008	E024b		X	24.74	26.58 ± 0.30	0.77
T009	E026	X		24.64	26.90 ± 0.31	0.76
T010	E039	X		25.61	27.00 ± 0.31	0.76
T008	E024a		X	24.18	27.11 ± 0.31	0.76
T011	E038	X		24.88	27.82 ± 0.32	0.74
T012	E035	X			27.87 ± 0.32	0.75
T013	E012		X	26.28	29.07 ± 0.33	0.75
Base.	**HASPI**	X		**28.00**	**29.47 ± 0.34**	**0.70**
T004	E020c-ni		X	30.16	32.74 ± 0.37	0.65
T014	E022a		X	31.11	33.97 ± 0.39	0.57
T014	E022b		X	33.10	35.48 ± 0.40	0.56
Base.	Prior		X	40.20	41.33 ± 0.47	—

Analysis

Predicted vs observed intelligibility for winning system

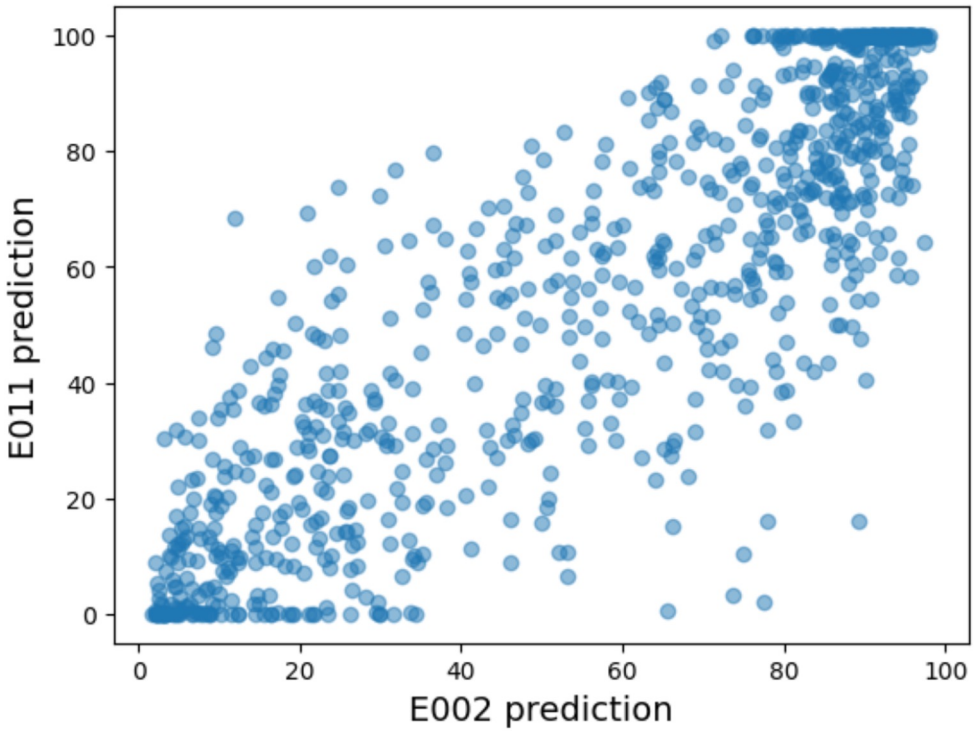


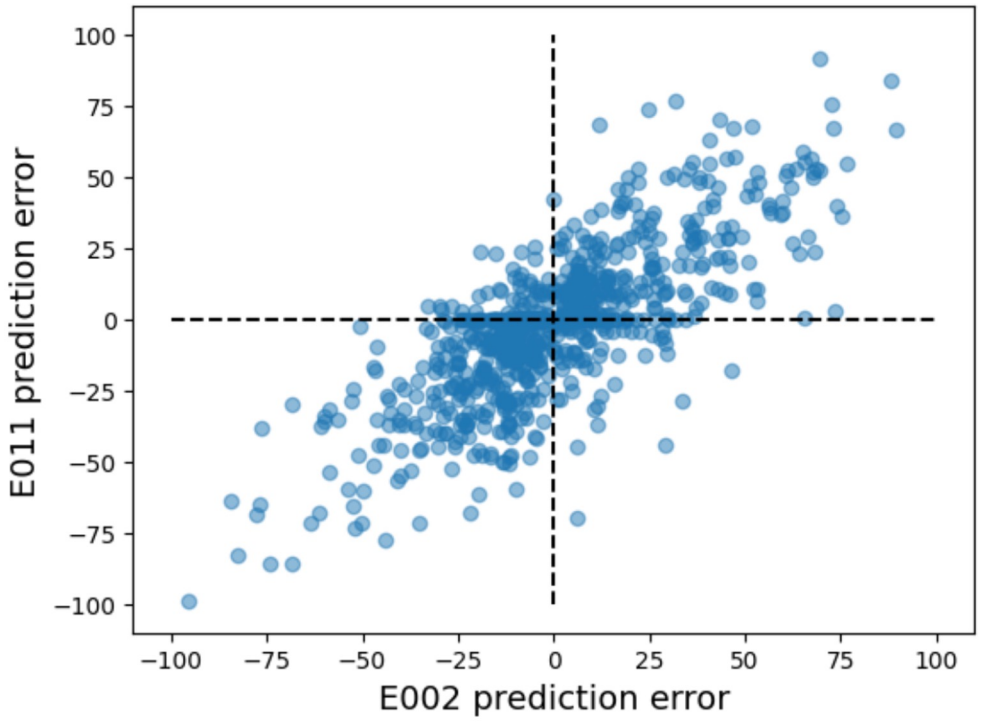
Analysis


E025 (RMSE: 24.98, Corr: 0.80)
E019 (RMSE: 25.31, Corr: 0.79)
E011a (RMSE: 25.54, Corr: 0.79)
E020a (RMSE: 25.60, Corr: 0.78)
E014 (RMSE: 25.82, Corr: 0.78)
E011c (RMSE: 25.83, Corr: 0.78)
E032 (RMSE: 25.99, Corr: 0.78)
E020b (RMSE: 26.02, Corr: 0.78)
E011b (RMSE: 26.12, Corr: 0.78)
E020c (RMSE: 26.14, Corr: 0.77)
E017a (RMSE: 26.30, Corr: 0.77)
E025 (RMSE: 24.98, Corr: 0.80)
E026 (RMSE: 26.90, Corr: 0.76)
E039 (RMSE: 27.00, Corr: 0.76)
E024a (RMSE: 27.11, Corr: 0.76)
E038 (RMSE: 27.82, Corr: 0.74)
E035 (RMSE: 27.87, Corr: 0.75)
E012 (RMSE: 29.07, Corr: 0.74)
E020c-ni (RMSE: 32.74, Corr: 0.65)
E022a (RMSE: 33.97, Corr: 0.57)
correctness
True Intelligibility

Analysis

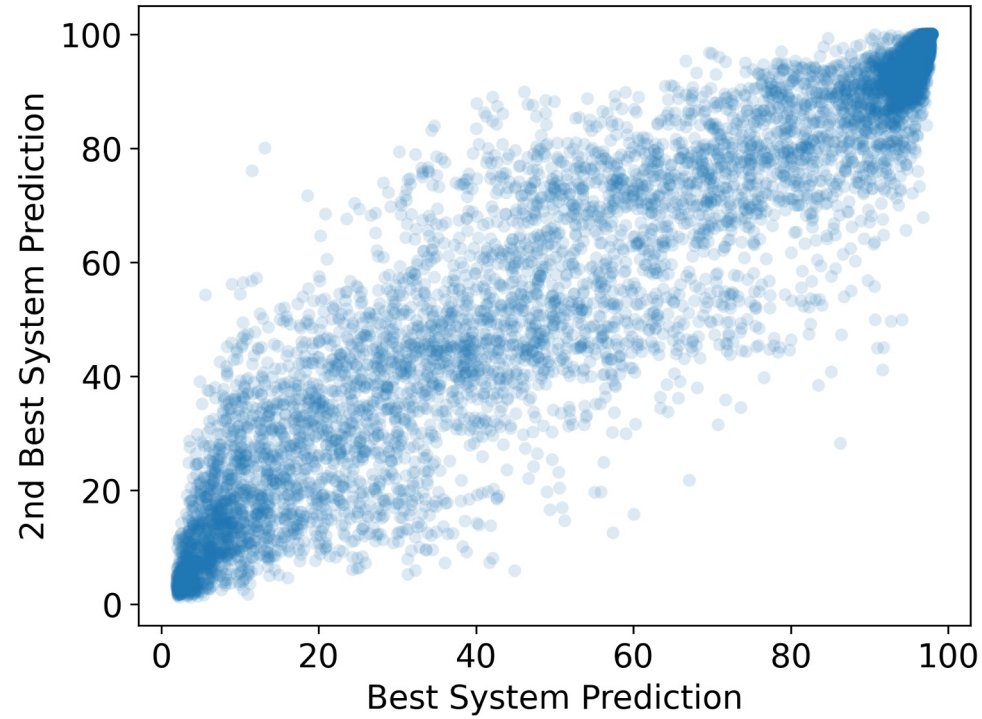
In CPC2, we observed complementarity among top systems

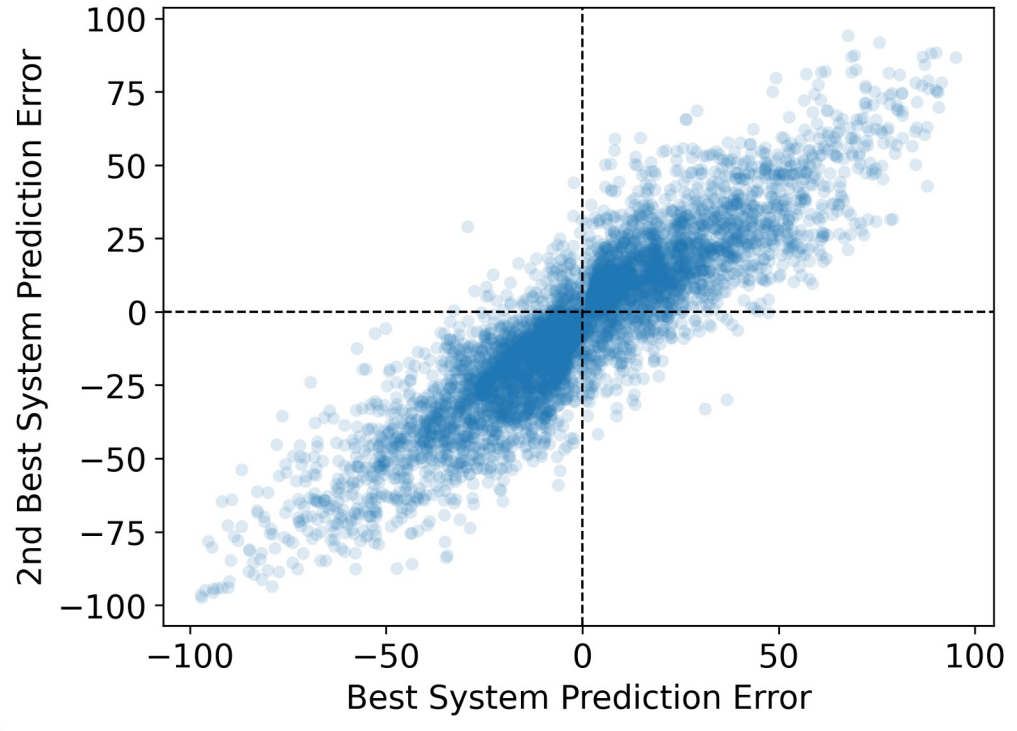




Analysis

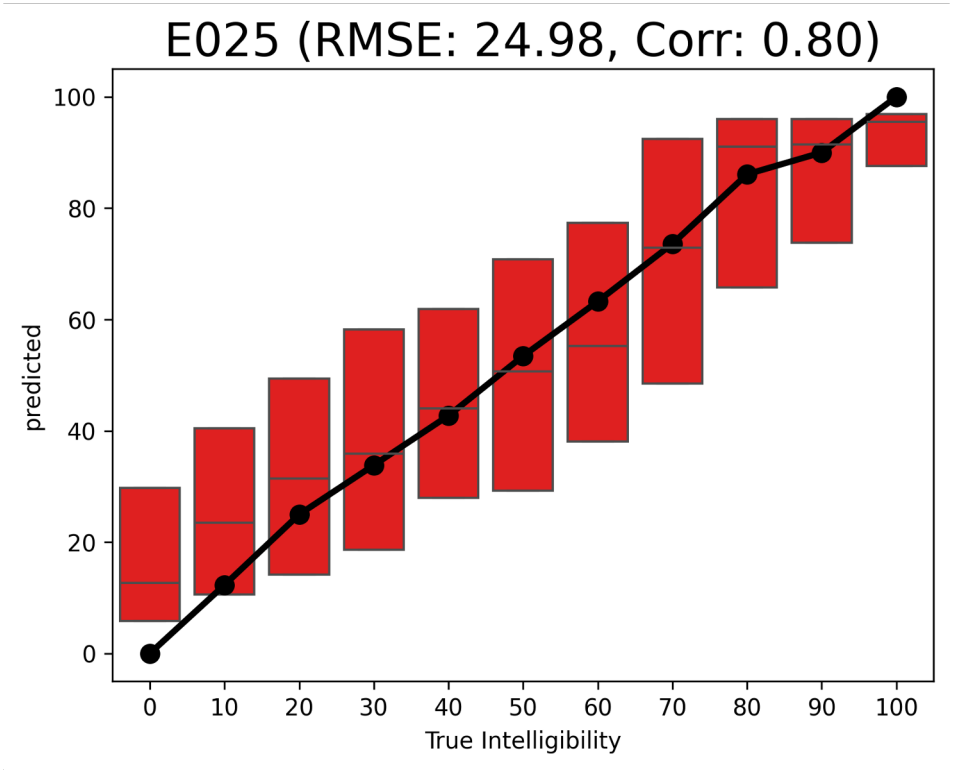
… similar pattern for top systems in CPC3





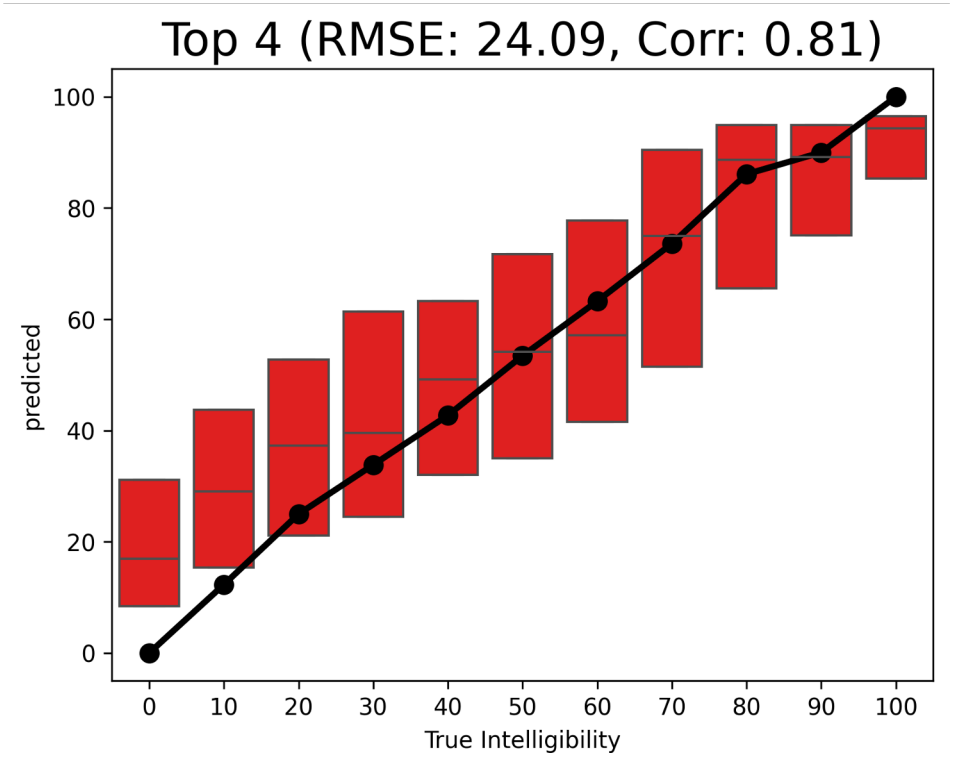
Analysis

Predicted vs observed intelligibility for winning system



Analysis

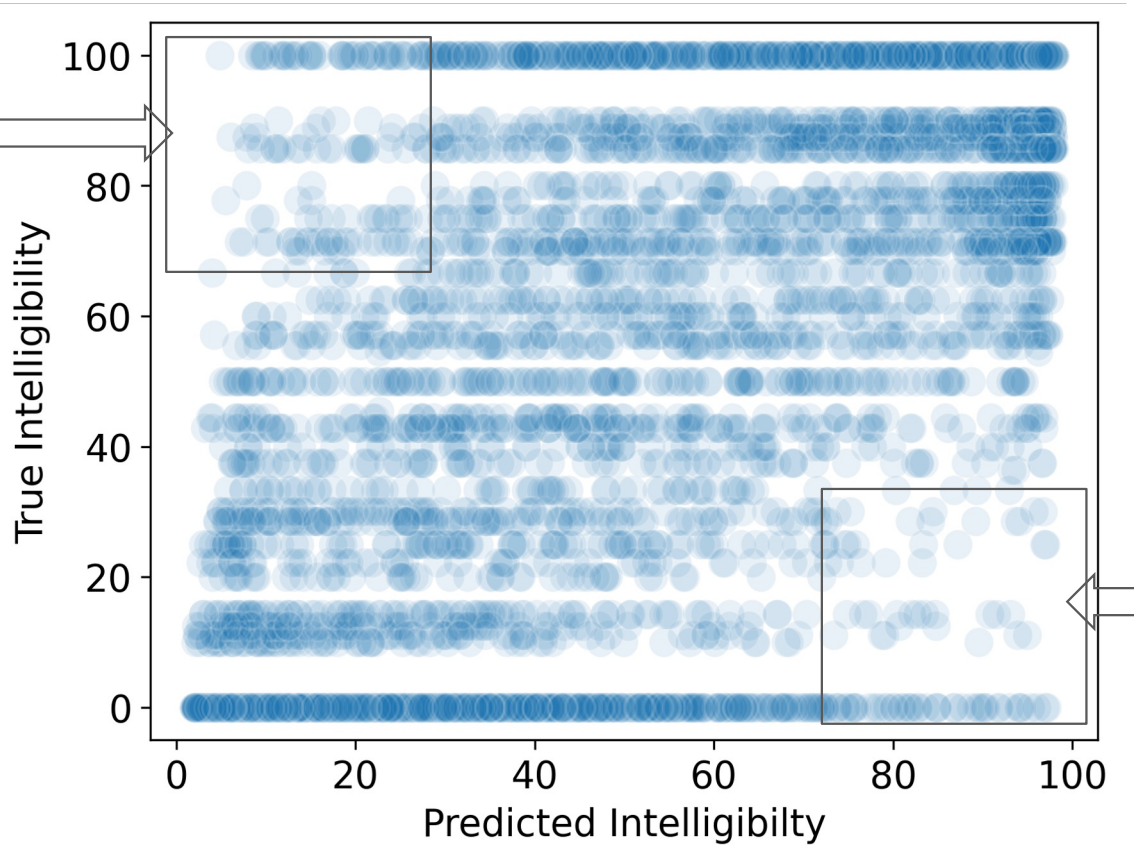
Predicted vs observed intelligibility when averaging top four systems



Observations

Predicted to be poorly intelligible but listener scored well.

Interesting.

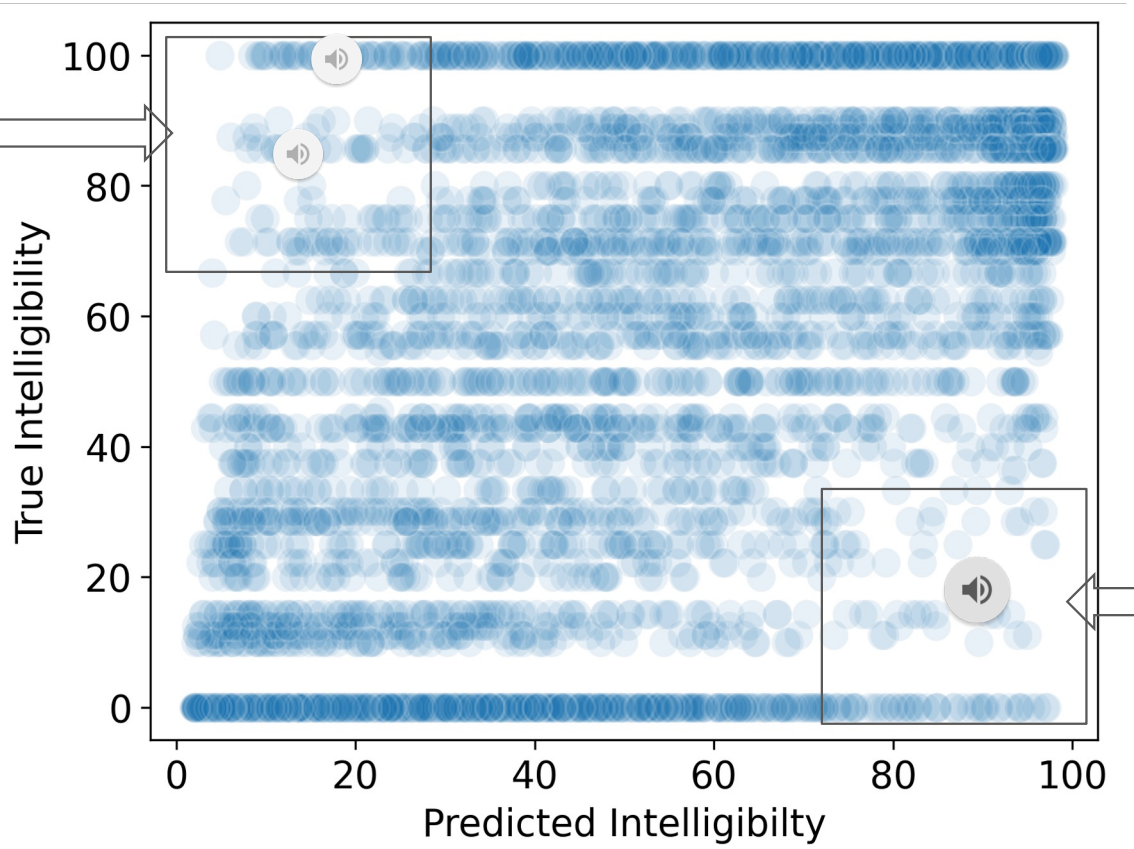


Predicted to be highly intelligible but listener scored poorly.

Many possible reasons.

Considerations

Predicted to be poorly intelligible but listener scored well.



Target:
“Cutting their pay will do nothing to induce a recovery”

Response
“having induced nothing for recovery”

Only 20% correct.

Predicted to be highly intelligible but listener scored poorly.

Many possible reasons.

Some Preliminary Conclusions

- Most of the submitted systems are performing better than the HASPI baseline.
- Many strong non-intrusive approaches are using pre-trained speech models (eg. Whisper).
- Best system was intrusive but it scored only marginally better than the best non-intrusive approach.
- Evidence of real progress in system performance since CPC1, CPC2
  - Non-intrusive systems outperforming intrusive systems
  - Best systems beating HASPI baseline by similar margin to CPC2 despite harder conditions
- Seems very hard to get the RMSE scores down lower than 20%. Many factors simply not predictable from the signal and HL severity alone.

Thank you for listening.

Questions?