The 2nd Clarity Prediction Challenge: A machine learning challenge for hearing aid intelligibility prediction

Jon Barker[1], *Michael A. Akeroyd*[2], *Will Bailey*[1], *Trevor J. Cox*[3], *John F. Culling*[4], *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

claritychallengecontact@gmail.com

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**.
- 5-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

Systems submitted by
the research community

Challenge materials
i.e., signals,
listeners

Enhancement Challenge

Objective scores

Subjective scores

Better Enhancement Algorithms

Better Objective Measures

New objective measures

Prediction Challenge

Systems submitted by the research community

Enhancement of hearing aids

- 1st Enhancement Challenge, **CEC1**, 2021
- 2nd Enhancement Challenge, **CEC2**, 2022
    - ICASSP SP Enhancement Challenge 2022-3
- 3rd Enhancement Challenge, **CEC3**, 2024-5

Coming soon

Prediction of speech intelligibility

- 1st Prediction Challenge, **CPC1**, 2021-2
- 2nd Prediction Challenge, CPC2, 2023

Results today!

The Clarity Prediction Challenge

Participants are given:

- A **hearing aid output signal** that has arisen from processing **speech in noise**
- The **audiogram 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 algorithm.

Clarity Prediction Challenge

The Task and Materials

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)

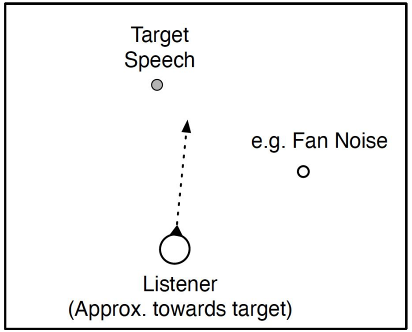
- Fully dynamic scenes.
- Yet to be fully defined.

Round 1

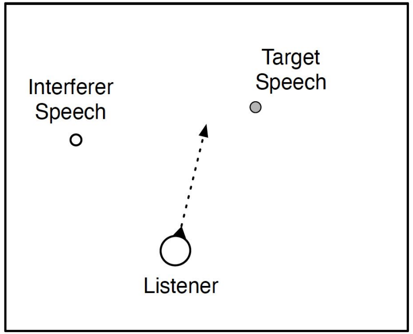
Target speech in presence of a single interferer.

- 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

Speech + Noise



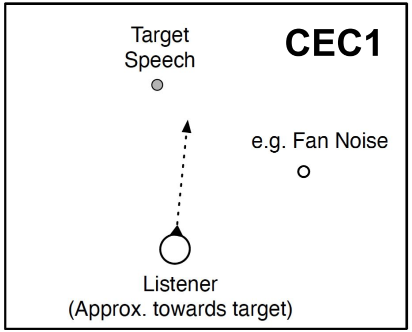
Speech + Speech

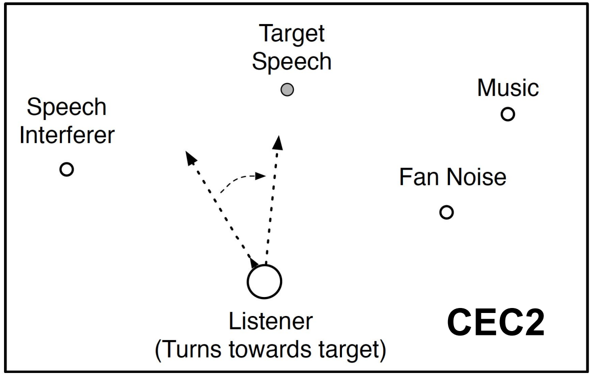


Round 2

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)



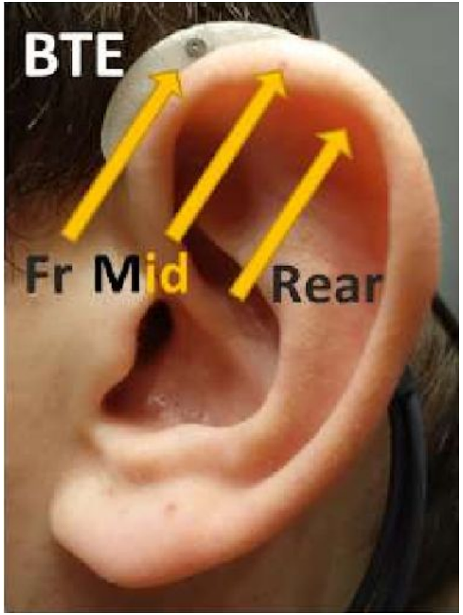


Simulated hearing aid inputs

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

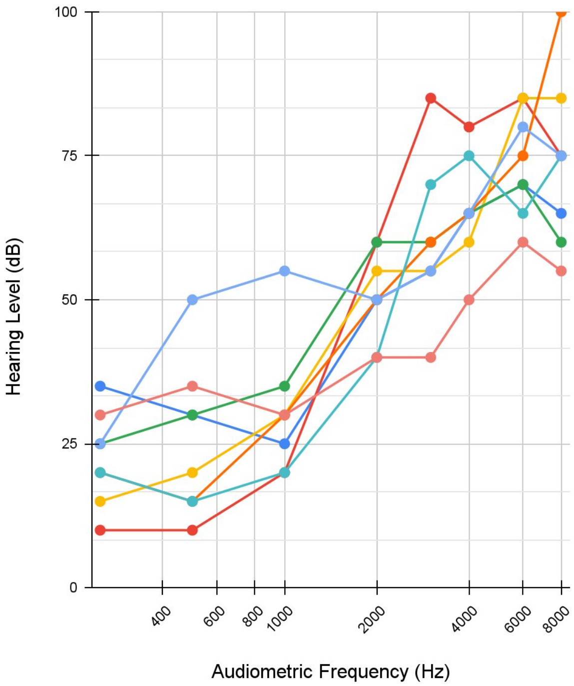
- i.e., the hearing aid algorithms are provided with six channels as input.

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

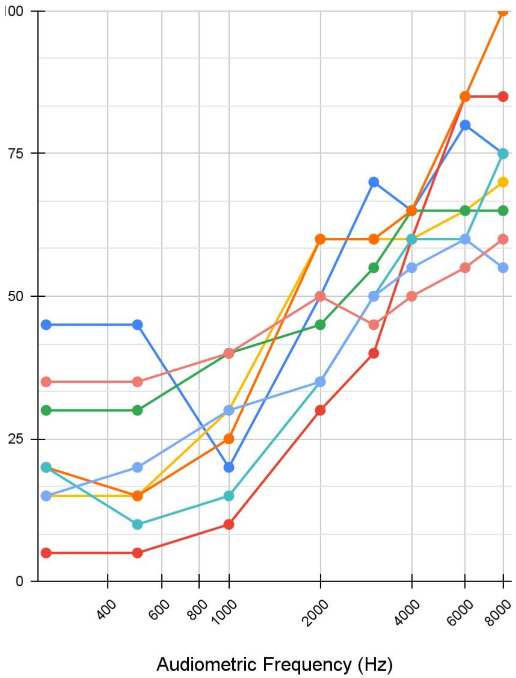


Listener Characteristics

Left Ear Audiograms



Right Ear Audiograms



Round 1 - 28 listeners.

Round 2 - 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

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			

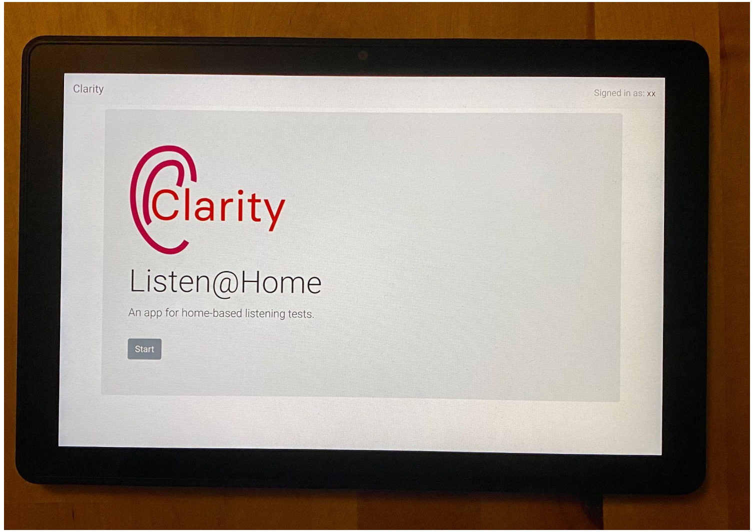
“And it is the most incredible thing”

	Good	Fair	Poor
S08501 / L0104			

“Roll over and repeat on the other side”

Listen@Home

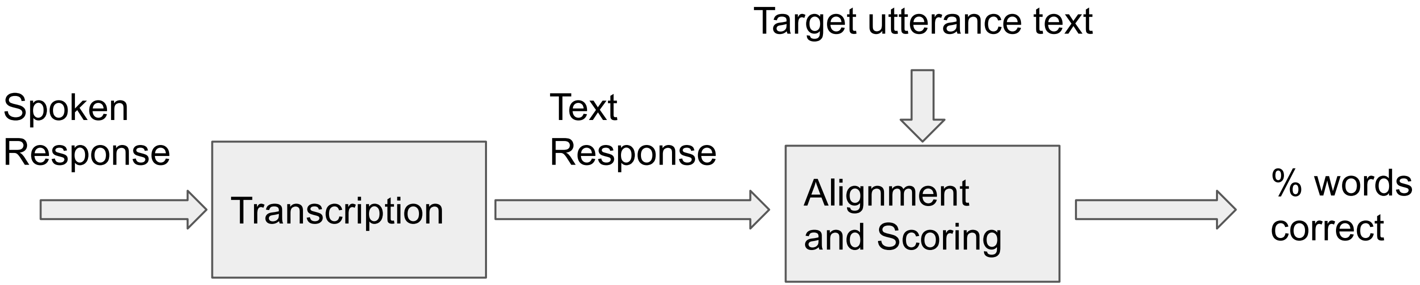
Participants listen to processed speech-in-noise and then respeak the sentence that they've heard.



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

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**.

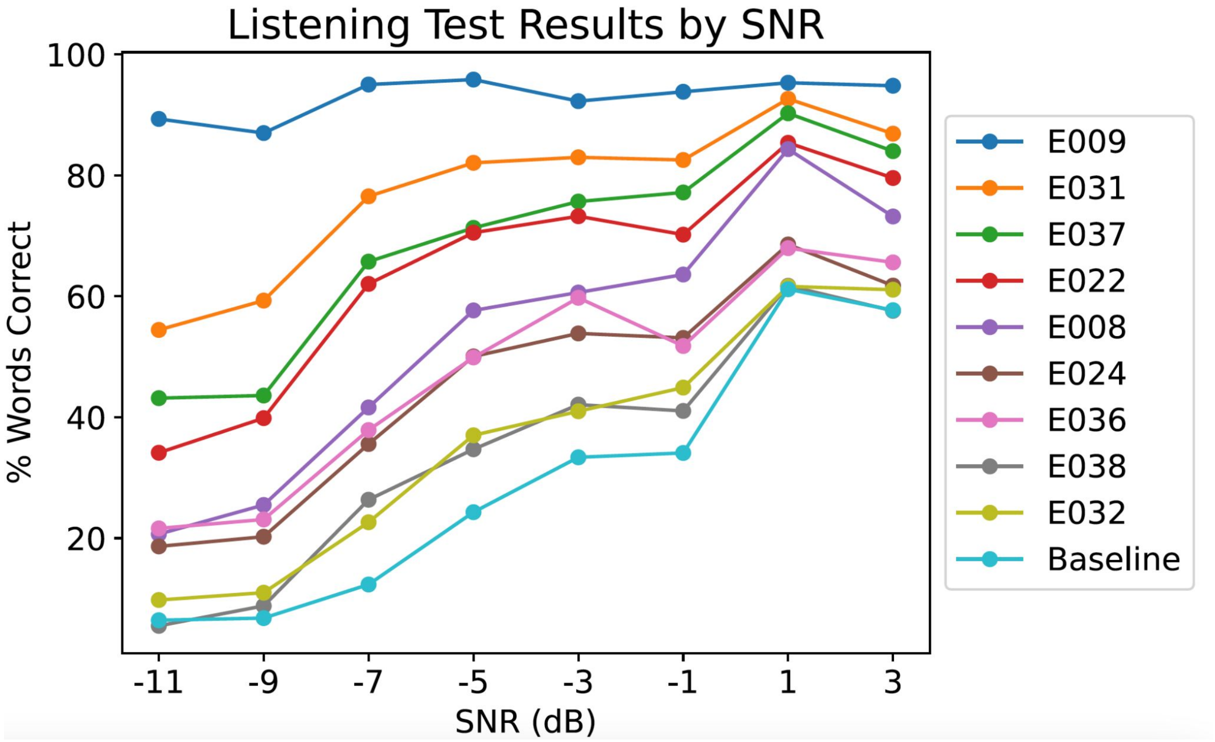
  Would score 5 out of 7 correct. (71%)

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

Clarity

Performance vs SNR



Clarity Prediction Challenge

Challenge Datasets and Rules

Train and Test set construction

10 systems and 15 listeners used for the challenge data.

	Training Data: CEC2 Data	Training Data: CEC1 Data	Test Data: CEC2 Data	
Partition 1	7 Systems 10 Listeners 2780 responses	10 Systems 22 Listeners 5124 responses	3 Systems 5 Listeners 305 responses	Test Set 1
Partition 2	7 Systems 10 Listeners 2780 responses	10 Systems 23 Listeners 5539 responses	3 Systems 5 Listeners 294 responses	Test Set 2
Partition 3	7 Systems 10 Listeners 2780 responses	10 Systems 24 Listeners 5820 responses	3 Systems 5 Listeners 298 responses	Test Set 3

Data organised into 3 partitions to allow all systems and listeners to appear in the test sets while keep the training and test sets disjoint. Data from the simpler CEC1 scenes also provided to increase size of training sets.

Clarity Prediction Challenge

Entries and Results

The Entrants

- We had **12 system submissions** arising from **9 separate teams**.
- Teams submitted technical papers which were reviewed to check compliance with the rules.
- Systems were classified as either **Intrusive or Non-intrusive**

- Systems were scored by
  - computing the **RMS error** between the true and estimated sentence intelligibilities
  - computing the **correlation** between the true and estimated sentence intelligibilities.
  - 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 ↑
Base.	beHASPI	X		28.7 ± 1.0	0.70
Base.	Prior		X	40.0 ± 1.3	–

CPC1 Results

RMSE = root mean squared intelligibility prediction error

Corr = Correlation between predicted and actual scores

Better-ear HASPI v2, Kates + Arehart, 2021

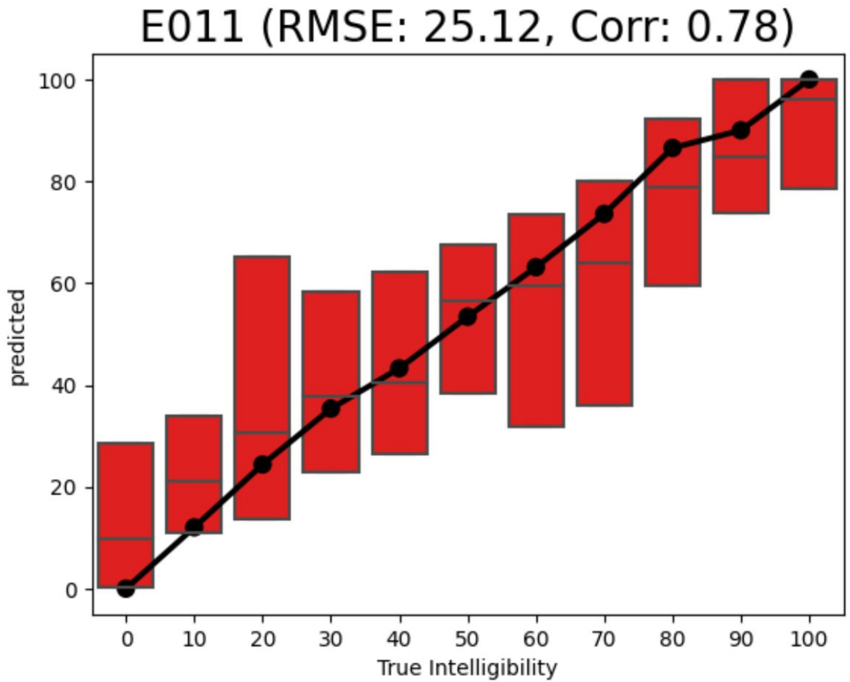
MSBG + MBSTOI

Always output training set average

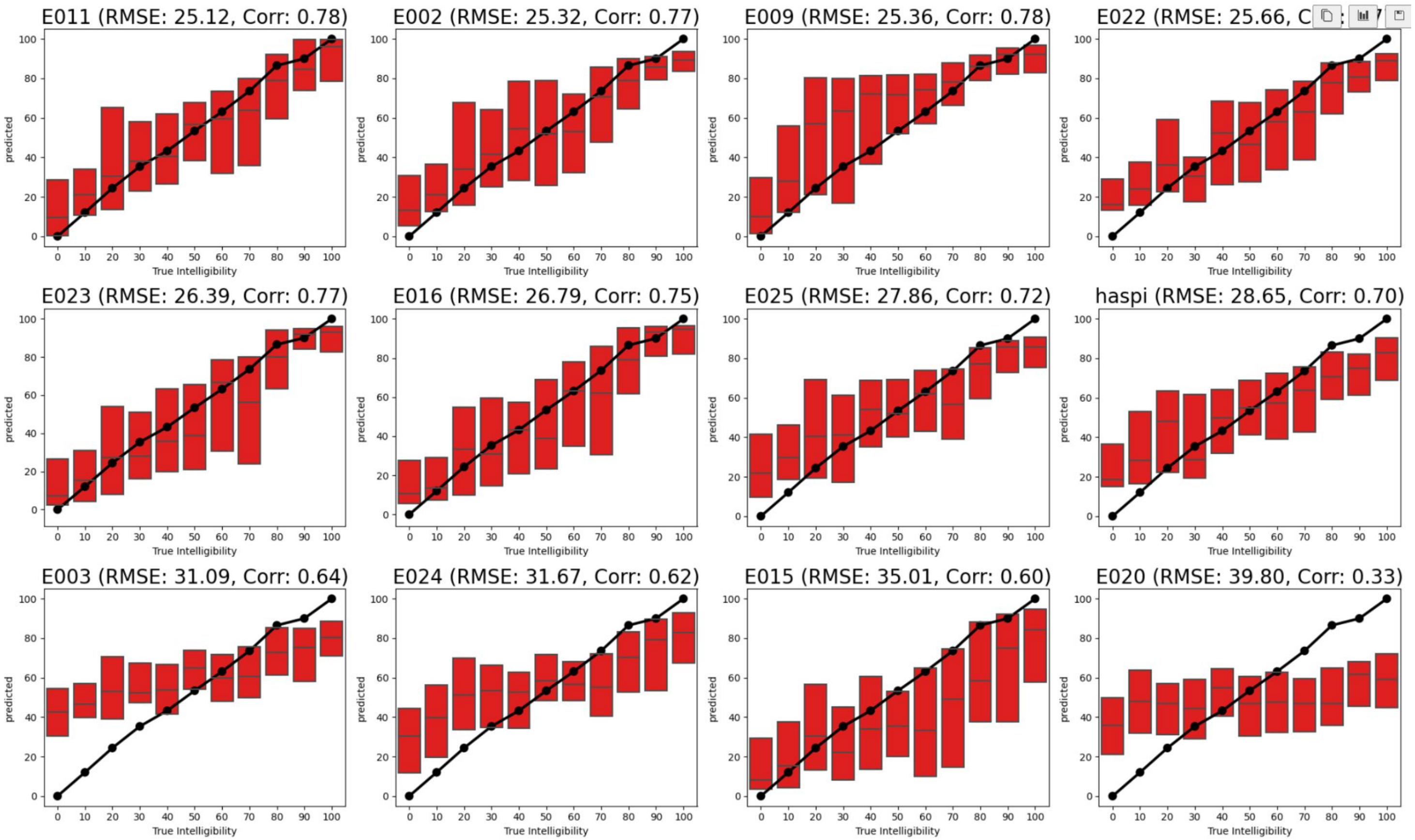
Entrant	Intr.	Track 1 (closed) RMSE ↓	Track 1 (closed) Corr ↑	Track 2 (open) RMSE ↓	Track 2 (open) Corr ↑
E30 [22]	Yes	**22.5 ± 0.5**	0.79	–	–
E32 [23]	Yes	23.1 ± 0.5	0.77	**23.5 ± 0.9**	0.76
E29 [24]	No	23.3 ± 0.5	0.77	24.6 ± 1.0	0.73
E36 [25]	Yes	24.0 ± 0.5	0.76	29.2 ± 1.2	0.60
E33 [26]	No	24.1 ± 0.5	0.75	28.9 ± 1.1	0.65
E16 [26]	No	24.7 ± 0.5	0.74	30.7 ± 1.2	0.59
E22 [27]	No	25.9 ± 0.5	0.70	32.1 ± 1.2	0.54
beHASPI	Yes	26.1 ± 0.5	0.70	27.3 ± 1.1	0.66
E19 [28]	Yes	27.5 ± 0.6	0.66	28.1 ± 1.1	0.63
Base. [1]	Yes	28.5 ± 0.6	0.62	36.5 ± 1.4	0.53
E06 [29]	No	32.0 ± 0.7	0.50	–	–
E34 [29]	No	33.4 ± 0.7	0.43	–	–
E35 [30]	No	35.4 ± 0.7	0.25	35.7 ± 1.4	0.22
Prior	No	36.4 ± 0.7	–	36.2 ± 1.4	-
E31 [31]	Yes	37.2 ± 0.7	0.41	28.3 ± 1.1	0.67
E23 [32]	No	41.5 ± 0.7	0.07	43.7 ± 1.5	0.05
E02 [33]	Yes	–	–	35.2 ± 1.4	0.38
E38 [33]	Yes	–	–	49.7 ± 1.5	0.30

Analysis

Predicted vs observed intelligibility for winning system

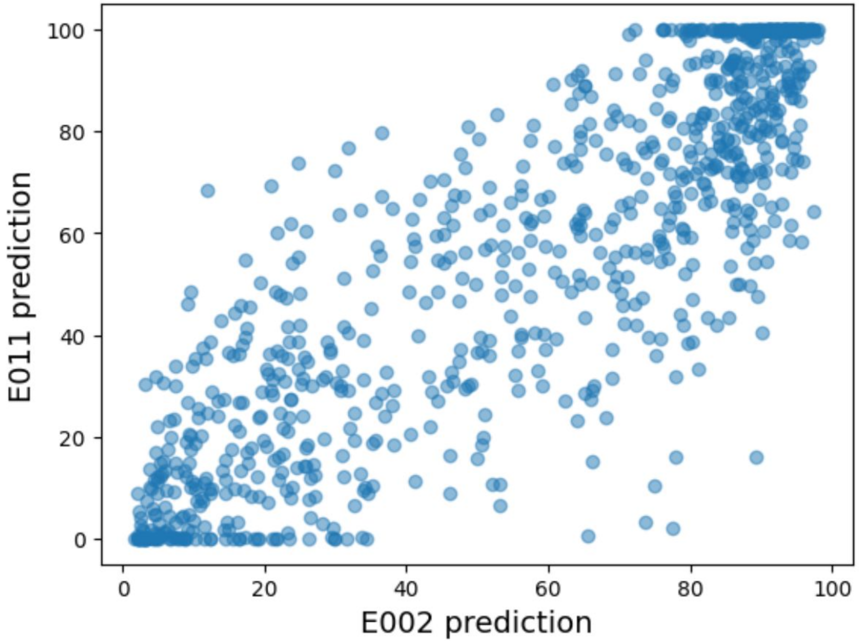


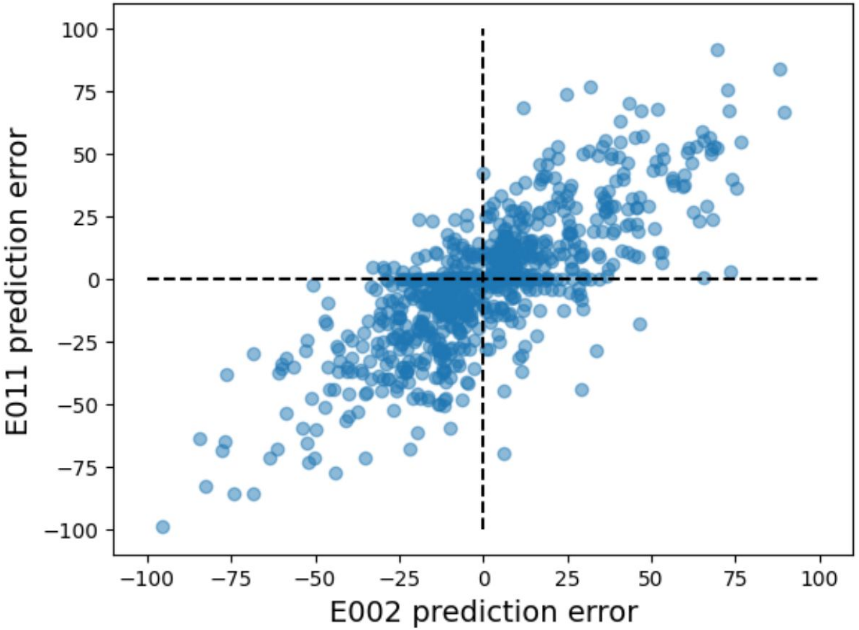
Analysis


E011 (RMSE: 25.12, Corr: 0.78)
E002 (RMSE: 25.32, Corr: 0.77)
E009 (RMSE: 25.36, Corr: 0.78)
E022 (RMSE: 25.66, C
E023 (RMSE: 26.39, Corr: 0.77)
E016 (RMSE: 26.79, Corr: 0.75)
E025 (RMSE: 27.86, Corr: 0.72)
haspi (RMSE: 28.65, Corr: 0.70)
E003 (RMSE: 31.09, Corr: 0.64)
E024 (RMSE: 31.67, Corr: 0.62)
E015 (RMSE: 35.01, Corr: 0.60)
E020 (RMSE: 39.80, Corr: 0.33)
predicted
True Intelligibility

Analysis

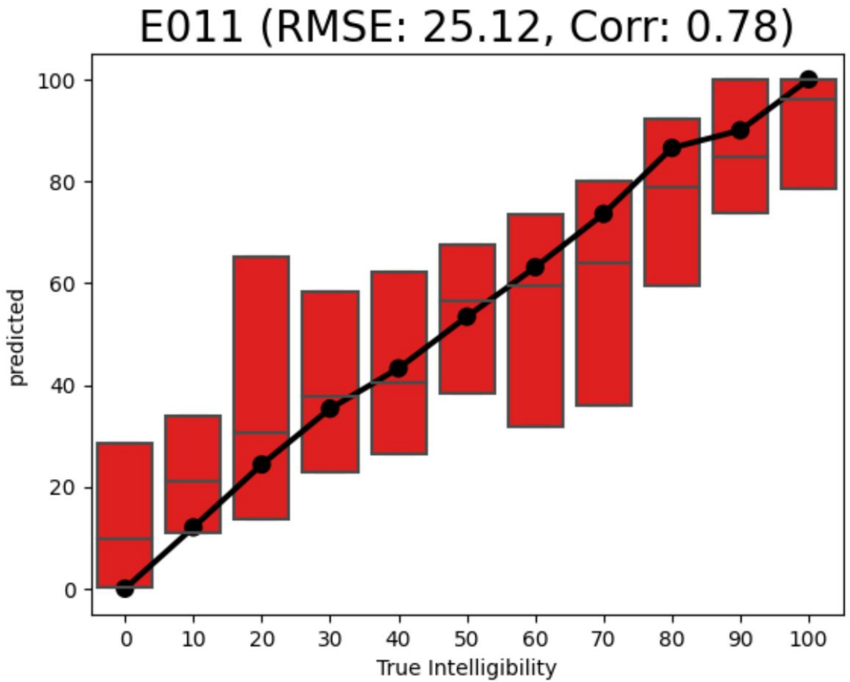
Complementarity of top 2 systems





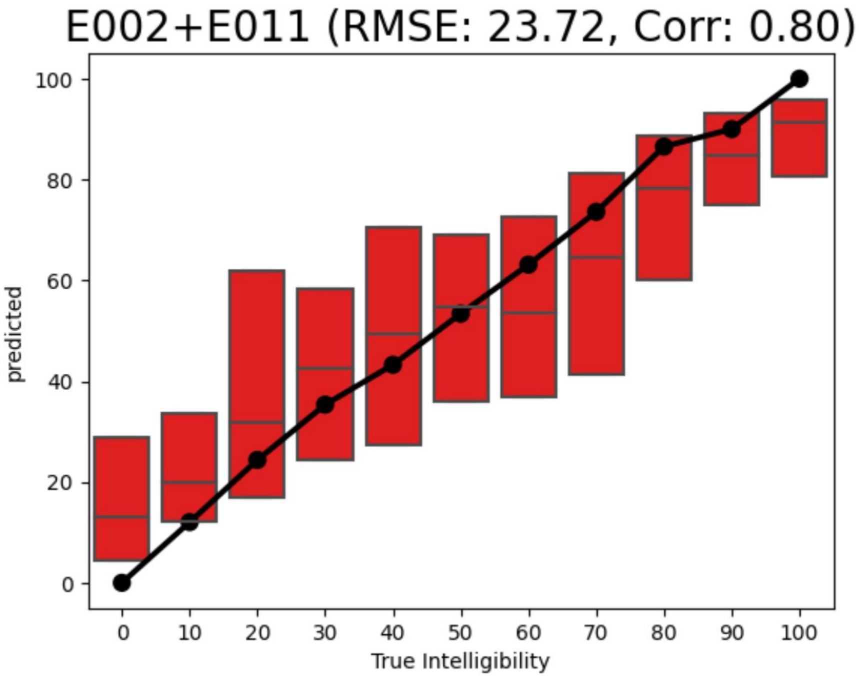
Analysis

Predicted vs observed intelligibility for baseline winning system



Analysis

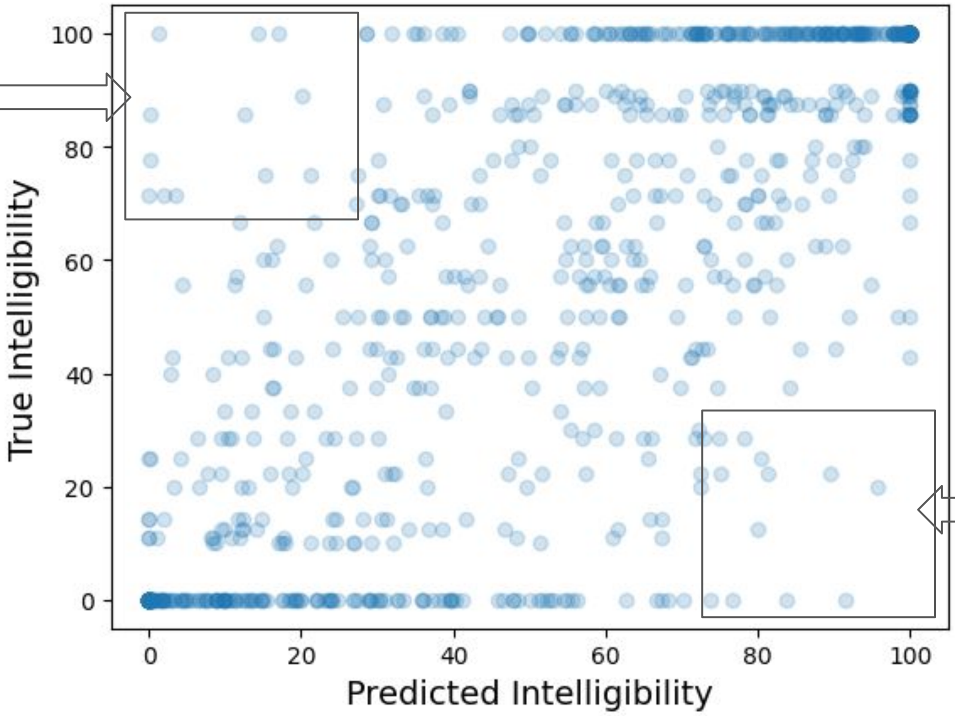
Predicted vs observed intelligibility for baseline winning system



Observations

Predicted to be poorly intelligible but listener scored well.

Interesting.

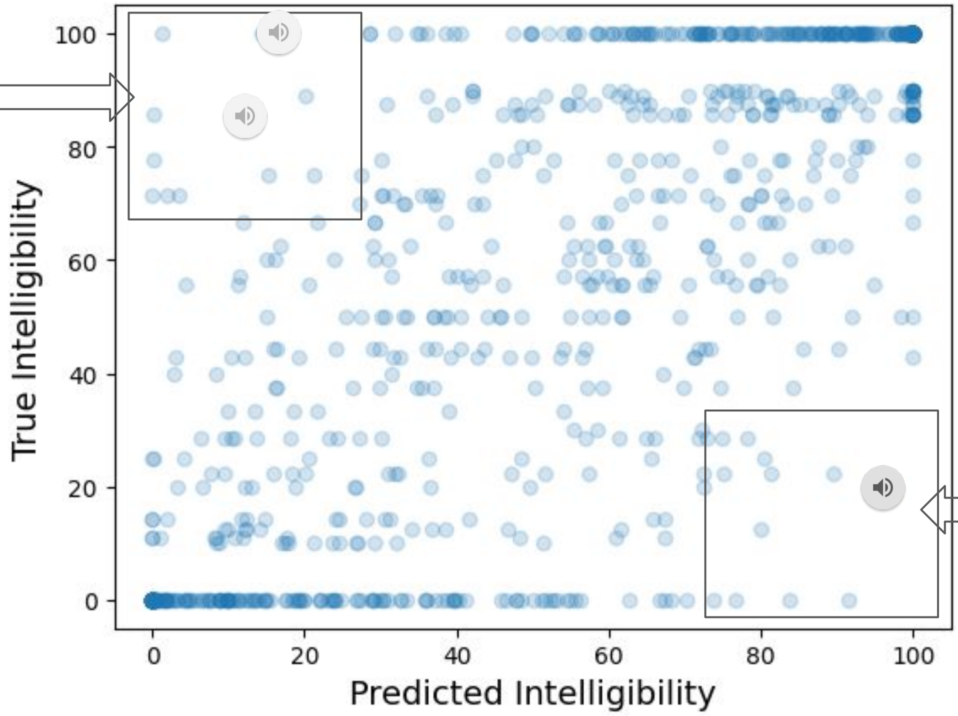


Predicted to be highly intelligible but listener scored poorly.

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 were non-intrusive
- Non-intrusive approaches are using DNN-based acoustic models that leverage developments in automatic speech recognition.
- 5 team produce non-intrusive systems that outperformed the intrusive HASPI baseline
- Evidence of real progress in system performance since CPC1
    - Non-intrusive systems outperforming intrusive systems
    - Increase in non-intrusive RMSE scores despite the task being harder
- More work needed to measure how well these systems generalise.

Thank you.

Audio Examples

Scene	SNR	Interferers	Mixed	Reference
S06033	4 dB	Music, speech, microwave		
S06001	2 dB	Speech, washing machine		
S06019	-1 dB	Speech, dishwasher, music		
S06032	-8 dB	Music, vacuum cleaner		
S06039	-11 dB	Speech, washing machine, vacuum		

HASPI vs Listener Scores