Engineering and Physical Sciences Research Council

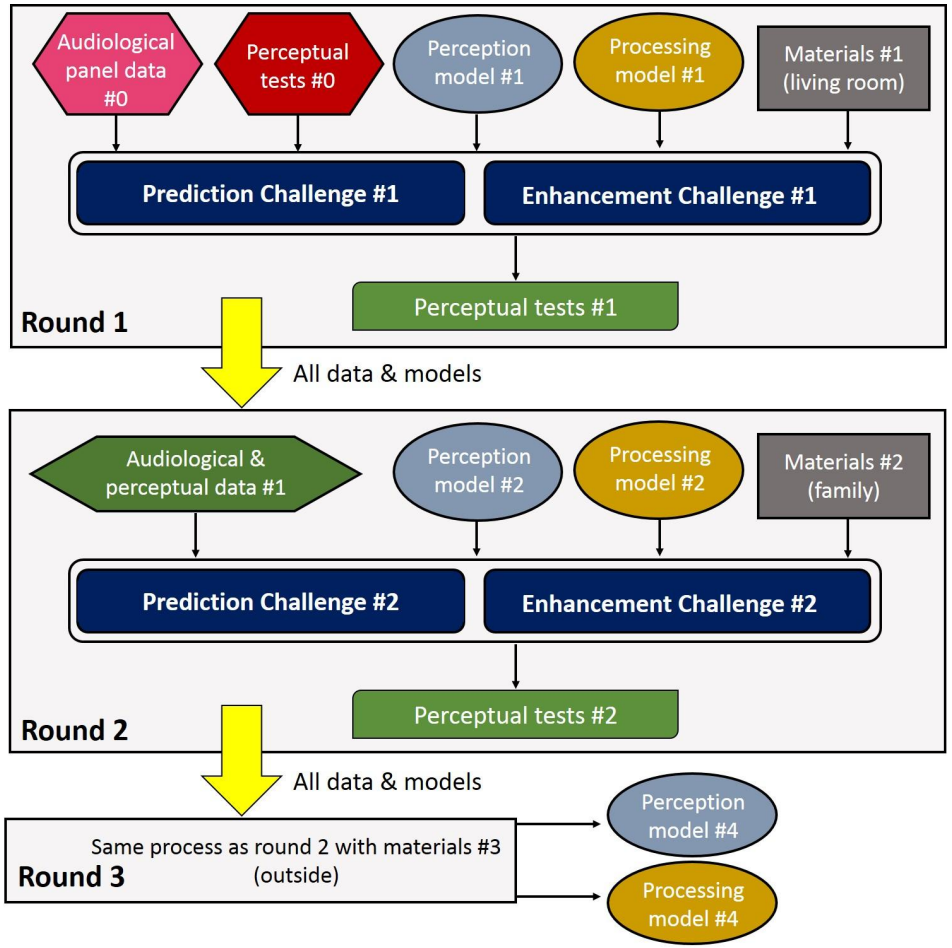
The 1st Clarity Prediction Challenge

Jon Barker, Michael A. Akeroyd, Trevor J. Cox, John F. Culling, Jennifer Firth, Simone Graetzer, Holly Griffiths, Lara Harris, Rhoddy Viveros-Munoz, Graham Naylor, Zuzanna Podwinska, Eszter Porter

- Two parallel challenges

  Enhancement challenge
  - Hearing aid signal processing

  Prediction challenge
  - Signal intelligibility prediction
- Three rounds over 5 years
  - Increasingly challenging listening scenarios
  - Each round will build on previous one, *i.e.*, data, tools, baseline
- First round launched Jan. 2021



The Clarity Challenge Plan

Round 1 (2021)

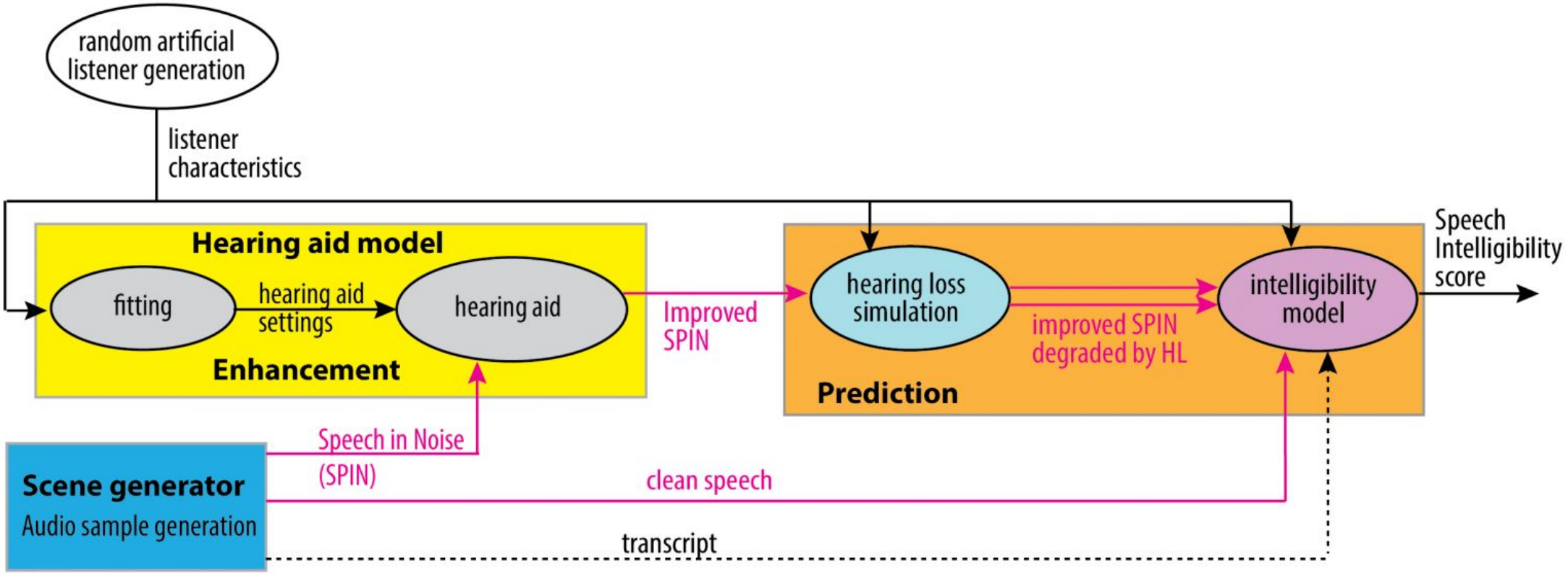
- Simple stationary scenes.
- Domestic living rooms with speech target and either i) a competing static speech source, or ii) a static domestic noise source.

Round 2 (2022)

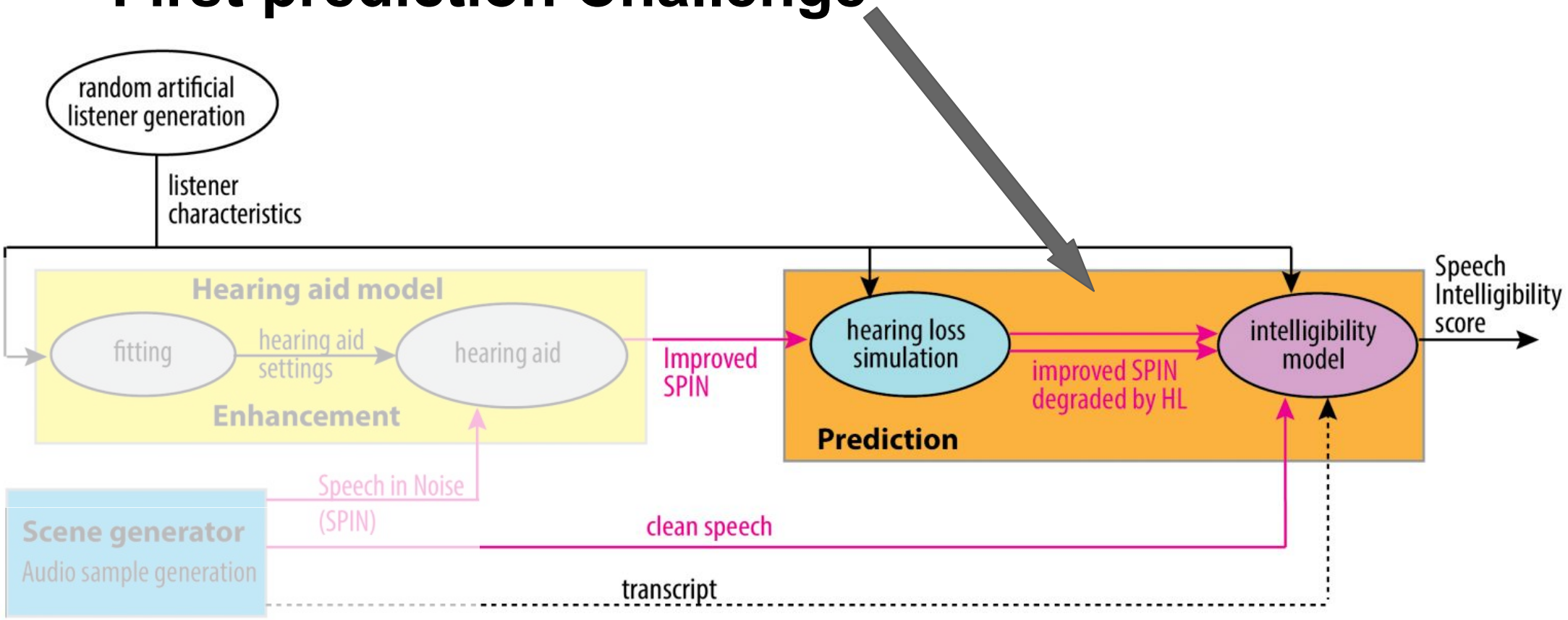
- Scenes with multiple noise sources
- Listener head movements

Round 3 (2023)

- Fully dynamic scenes.
- Yet to be defined.


random artificial listener generation
listener characteristics
Hearing aid model
fitting
hearing aid settings
hearing aid
Enhancement
Improved SPIN
hearing loss simulation
improved SPIN degraded by HL
intelligibility model
Prediction
Speech Intelligibility score
Speech in Noise (SPIN)
clean speech
Scene generator
Audio sample generation
transcript

First prediction Challenge



Engineering and Physical Sciences Research Council

Clarity Prediction Challenge

The Challenge Task

Task: to predict a **hearing-impaired listener**'s judgement of the intelligibility of a **speech-in-noise signal** that has been processed by a **hearing-aid algorithm**.

Competitors are given

<processed signal> and <listener id>

And must predict

<intelligibility score>

Intelligibility scores:

- The signals are short sentences, 7-10 words long
- The per-sentence intelligibility is reported as the number of words in the sentence recognised correctly, expressed as a percentage.

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%)

Engineering and Physical Sciences Research Council

Clarity Prediction Challenge

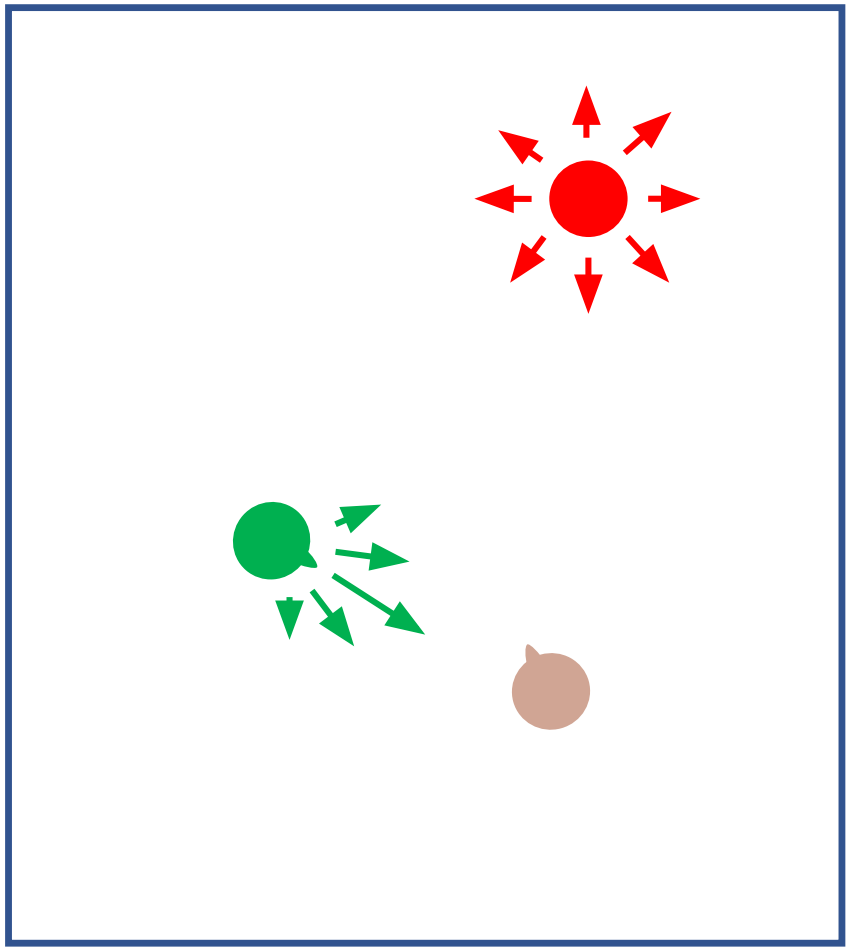
The Speech in Noise signals

Spatial configuration

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. It has human speech directivity and is oriented towards the listener.

Interferer anywhere, except within 1 m of a wall and omnidirectional. Domestic noise source - kettle, washing machine etc



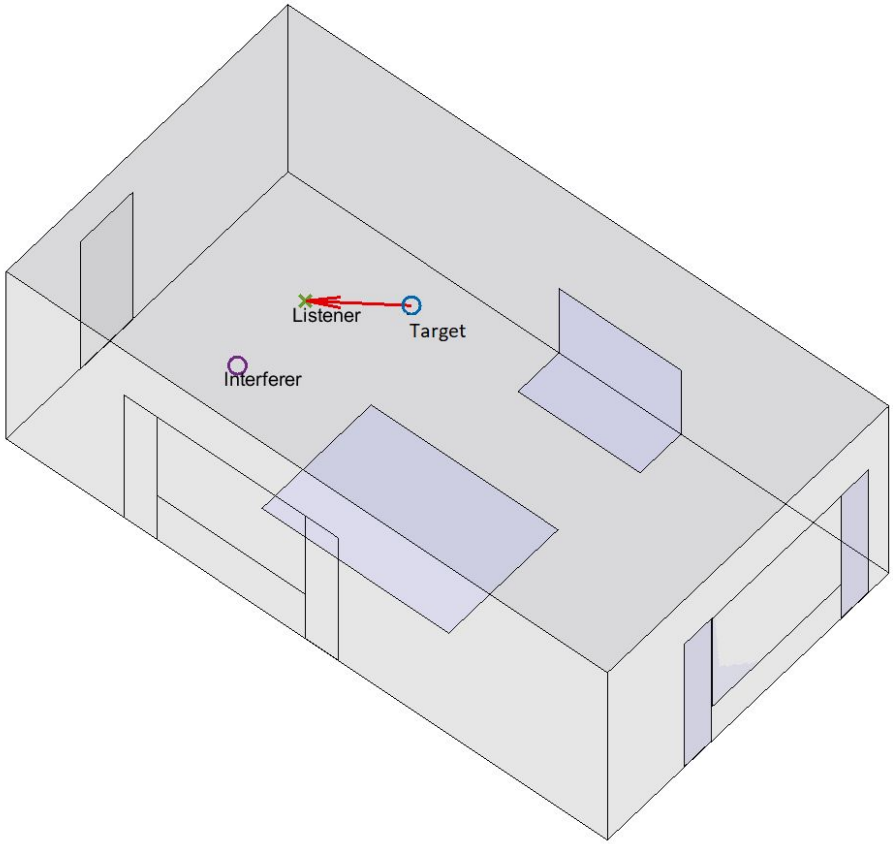
Speech Targets

- 10,000 different sentences selected from the British National Corpus (www.natcorp.ox.ac.uk) of (mainly) written text materials (novels, pamphlets etc., but excluding poetry).
- Screened to contain 7-10 words, all with a word frequency of at least one in the Kucera and Francis database, and hand checked for acceptable grammar and vocabulary by the Clarity project team.
- Recorded (at home, due to Covid-19) by 40 voice actors from a radio production company, reading 250 sentences each.

Graetzer, S., et al. (2022). Dataset of British English speech recordings for psychoacoustics and speech processing research: The clarity speech corpus. Data in Brief, 41(107951), 2711.

Environment

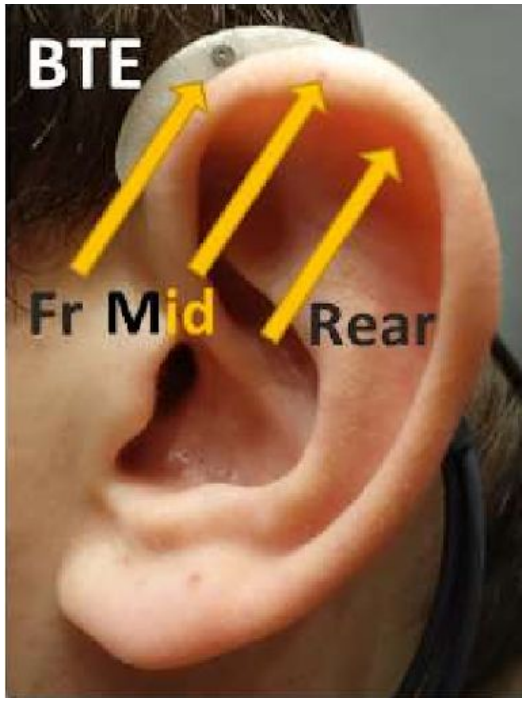
- Room impulse responses from each source to six hearing-aid mics in 10,000 spatial configurations generated by RAVEN.
- The rooms were based on the statistics of British living rooms – dimensions and reverberation times (Burgess & Utley, 1985).
- Rooms are all rectangular, but feature variations in surface absorption to represent doors, window, curtains rugs, furniture etc., combined with scattering coefficient of 0.1.



Simulated hearing aid inputs

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

- i.e. The hearing aid algorithms have 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

Engineering and Physical Sciences Research Council

Clarity Prediction Challenge

The hearing aid algorithms

Overview of approaches

Hearing aid algorithms were the entrants of the Clarity Enhancement Challenge (CEC1)

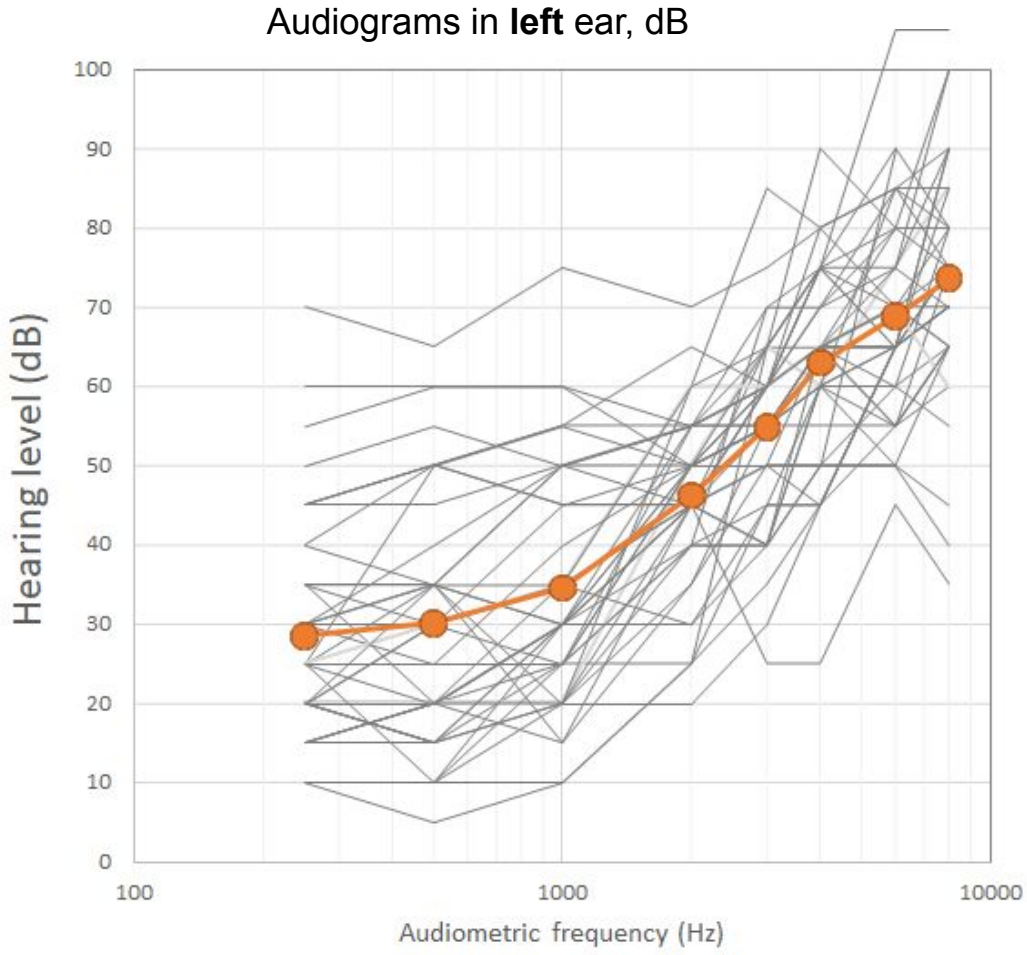
Entrant	Beamforming	DNN Noise Removal	Hearing Loss Compensation
E001			Baseline
E003	RLS	Conv-TasNet	Linear, fitting formula
E005		Binaural Conv-Tasnet	
E007	MVDR	Conv-TasNet	Linear, NN-optimised
E009		MC Conv-TasNet	Linear, NN-optimised
E010		U-Net CNN	Linear, fitting formula
E013	MVDR		Linear, fitting formula but AGC
E018		2D CNN + LSTM, WPE	Dynamic EQ
E019	Weighted LCMP		MBDRC
E021	Weighted LCMP	DNN (Deep MFMBVDR)	MBDRC

Engineering and Physical Sciences Research Council

Clarity Prediction Challenge

The listening tests

Listener Characteristics

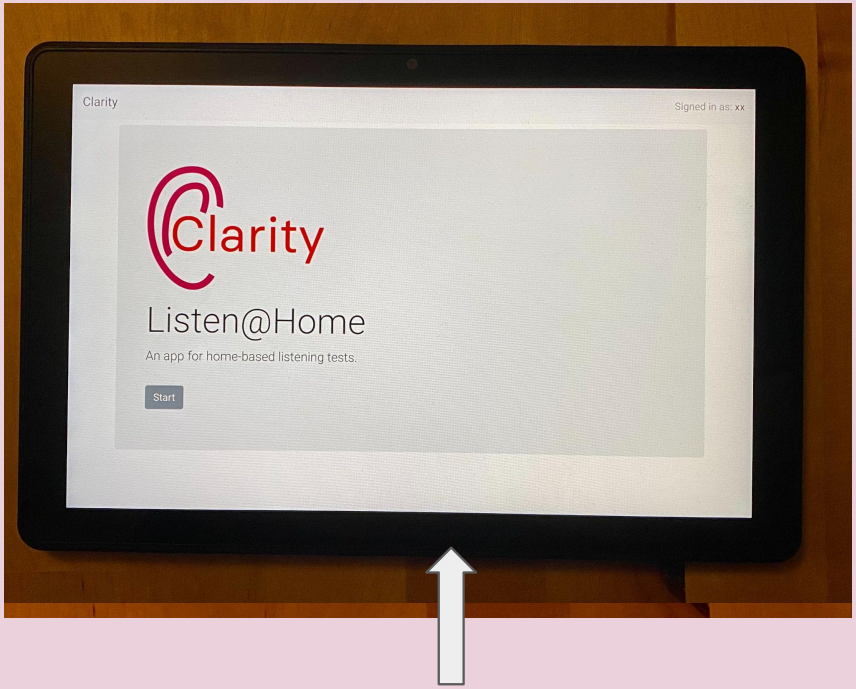


Hearing Loss

Mean better ear	= 40 dB
Mean worse ear	= 47 dB
Mean better-worse difference	= 7 dB
Mean left ear	= 43 dB
Mean right ear	= 43 dB

“Listen@Home”

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



Headphone measurements - PC8

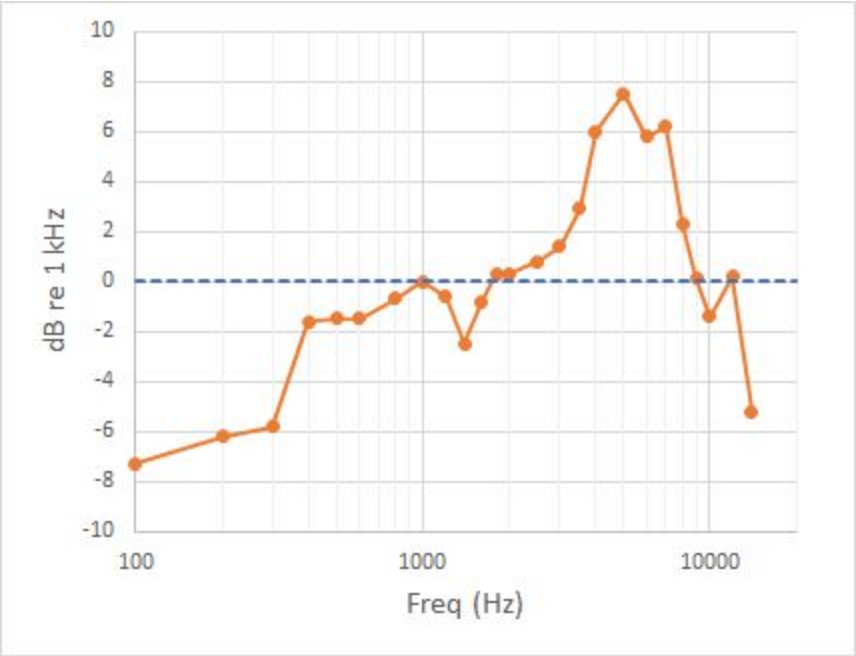
Levels measured as dB SPL produced by a +/- fullscale sinusoid @ 1 kHz and so is the maximum volume from the headset. (B&K 4192 ½" mic on a 4153 artificial ear to a 2260 SLM)

"Reference" set gave 99 dB.

Actual sets (43 of them):

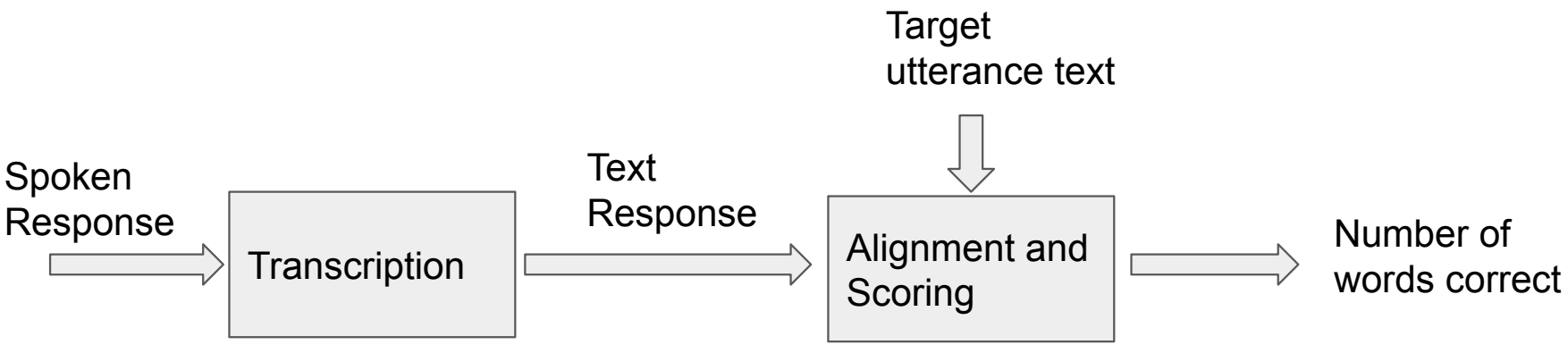
1 @ 94 dB
8 @ 96 dB
16 @ 97 dB
12 @ 98 dB
4 @ 99 dB
2 @ 100 dB

… so some variation across our sample.



Intelligibility Scoring

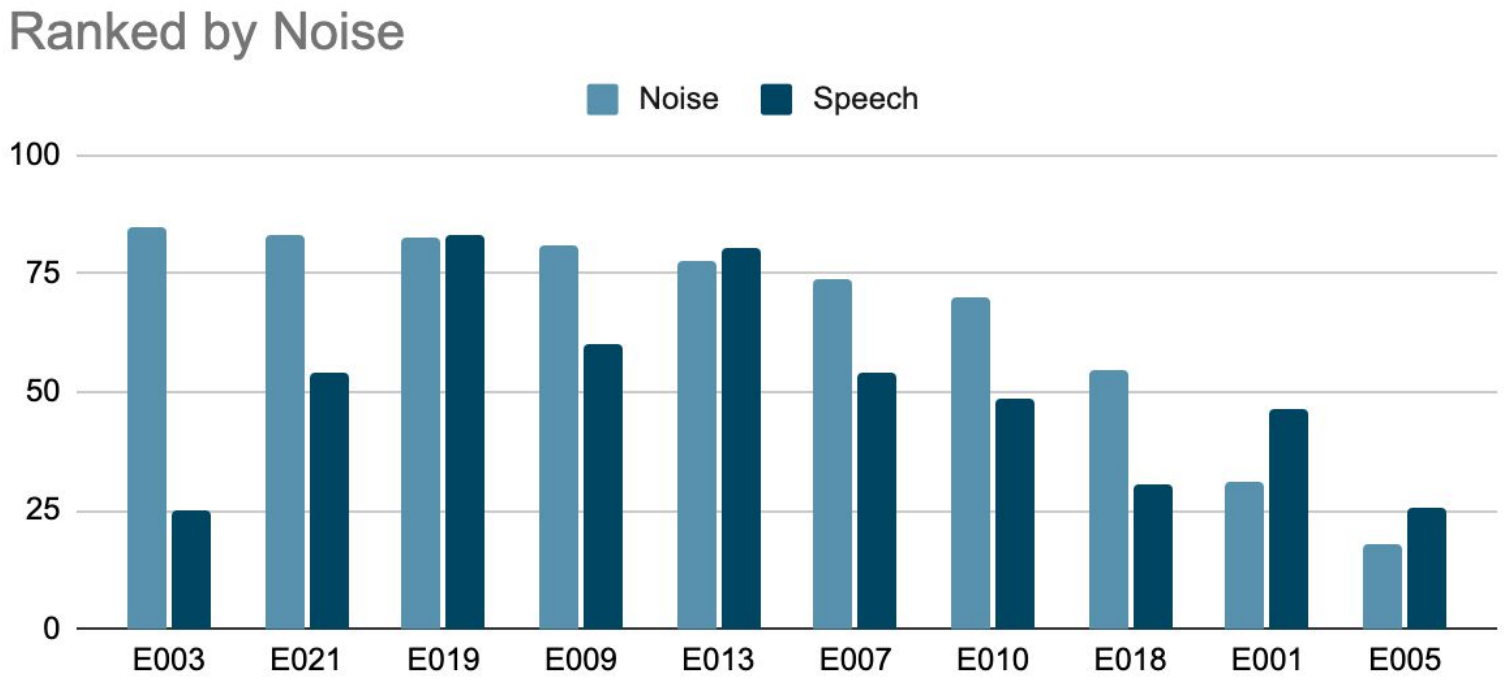
Tests are scored as percentage of words recognised/identified correctly.



Target: She **did not return to** **land** again.

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

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


Ranked by Noise
Noise
Speech
100
75
50
25
0
E003
E021
E019
E009
E013
E007
E010
E018
E001
E005

Engineering and Physical Sciences Research Council

Clarity Prediction Challenge

Challenge Datasets and Rules

Open vs Closed set evaluation

Total of 7233 responses from 27 listeners using 10 systems.
Data partitioned in two ways

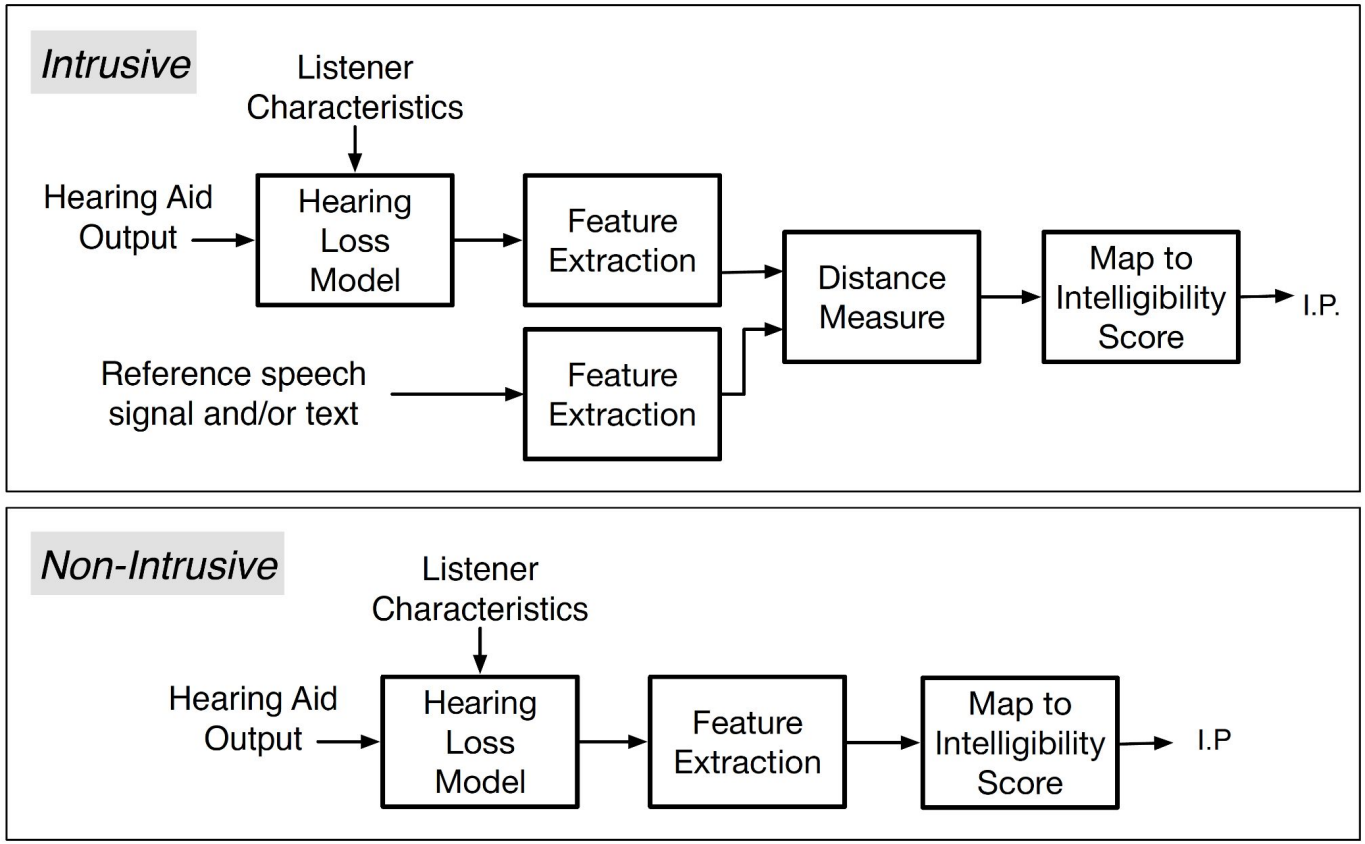
Track 1 (closed set).

- Same listeners and HA systems in the training set (4812 responses) and test (2421 responses).

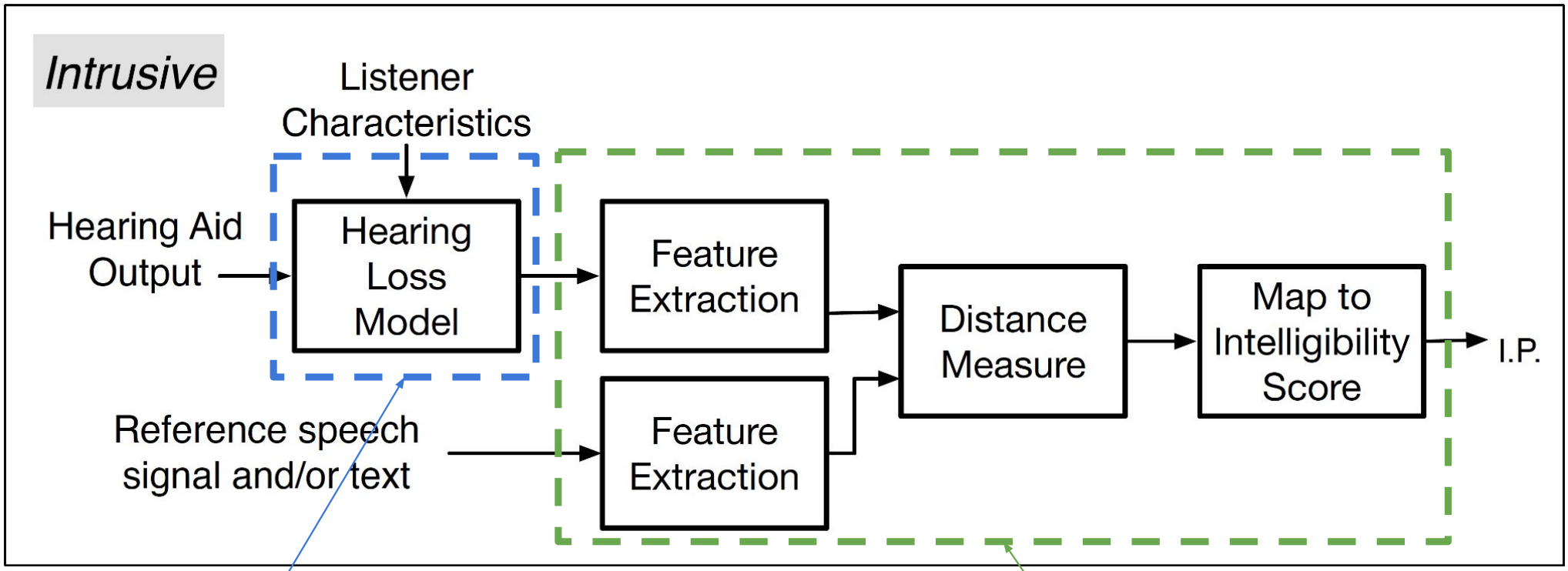
Track 2 (open set, i.e. unseen listener or unseen system).

- **Train set:** 22 listeners and 9 systems (3545 responses),
- **Test set:**
  - unseen listeners (5 listeners, 432 responses)
  - unseen system (1 system, 249 responses)

Intrusive vs non-intrusive system

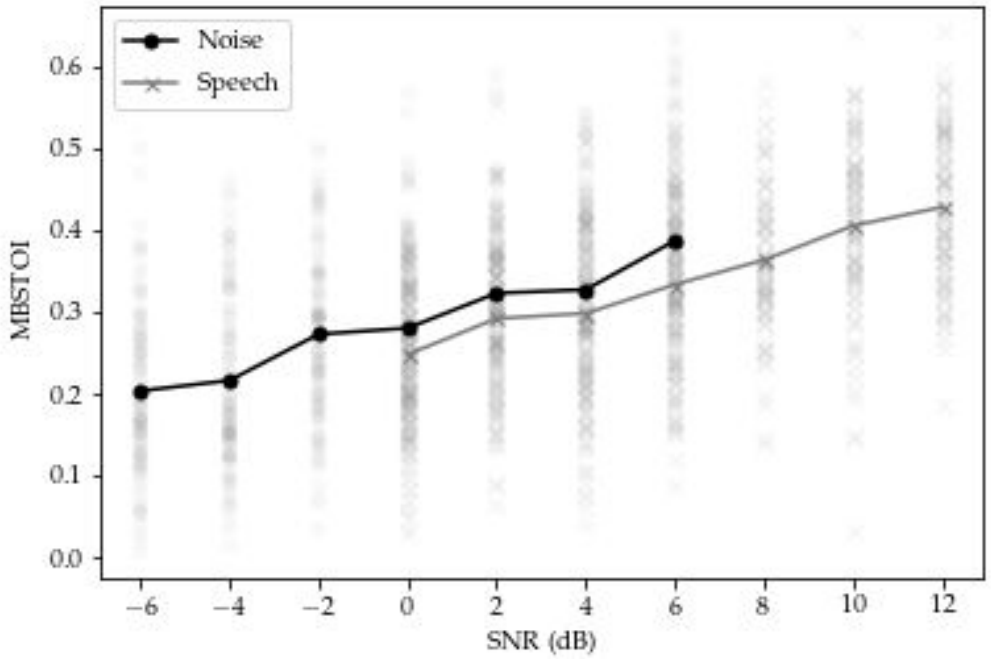


Baseline prediction system



MBSTOI by listening condition

For the baseline system



MBSTOI behaving sensibly

- Increases with SNR
- Decreases as the distance between the target and listener increases
- Decreases as average frequency hearing loss increases

Engineering and Physical Sciences Research Council

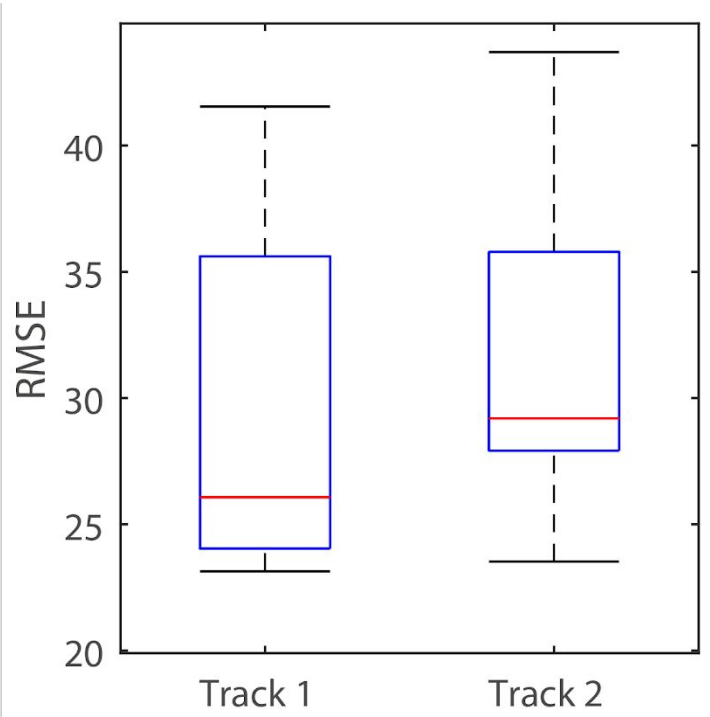
Clarity Prediction Challenge

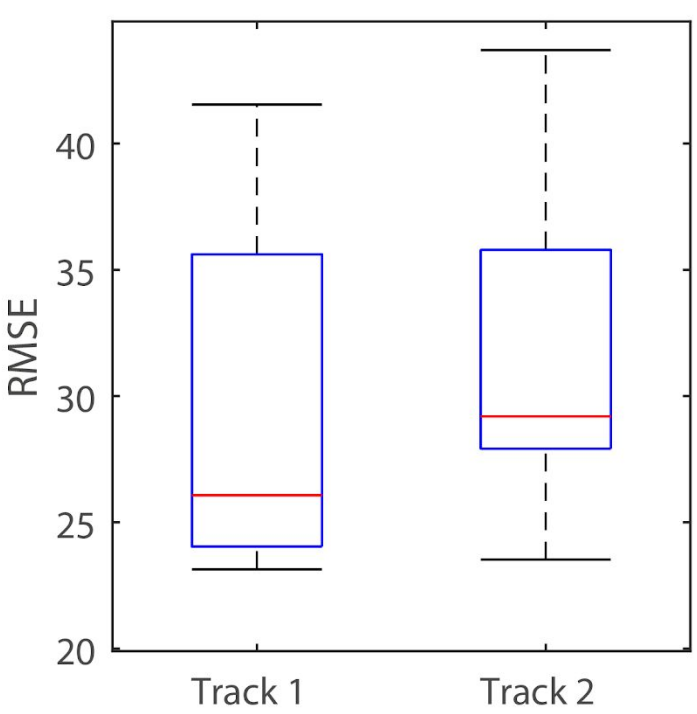
Entries and Results

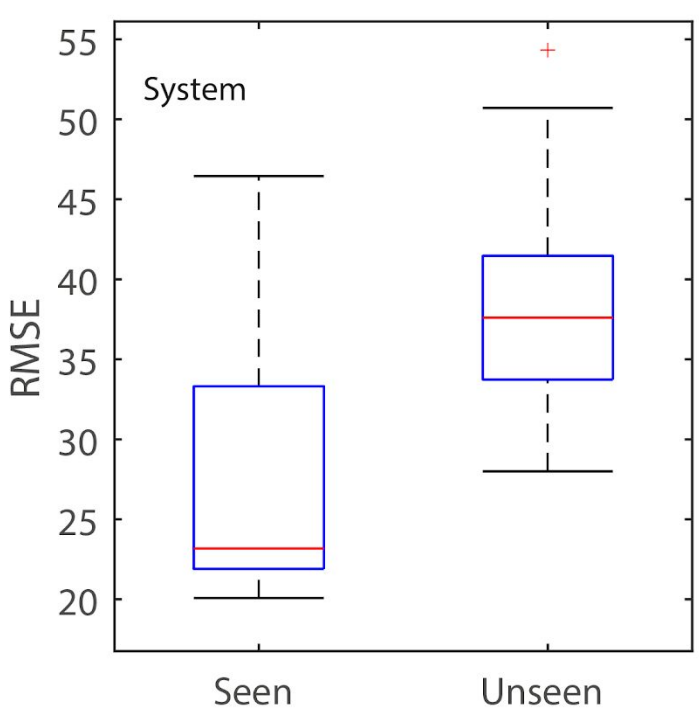
The Entrants

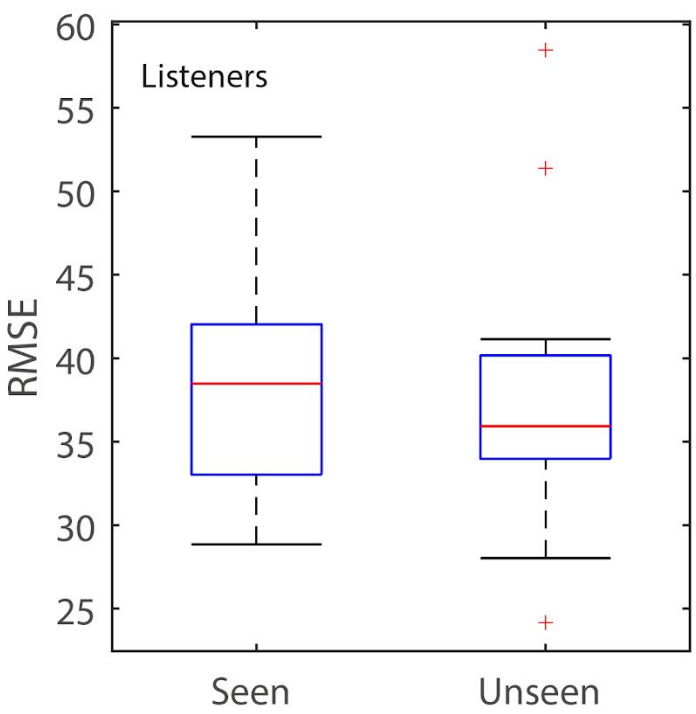
- We had **15 system submissions** arising from **9 separate teams**.
- Teams submitted technical papers which were reviewed to check compliance. **All submissions complied with the rules**.
- Systems were classified as either **Intrusive or Non-intrusive**
- Also included in analysis:
  - Predictions using HASPI
  - A simple algorithm (‘prior’) that just guessed the mean of the training set intelligibility for every example.

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
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

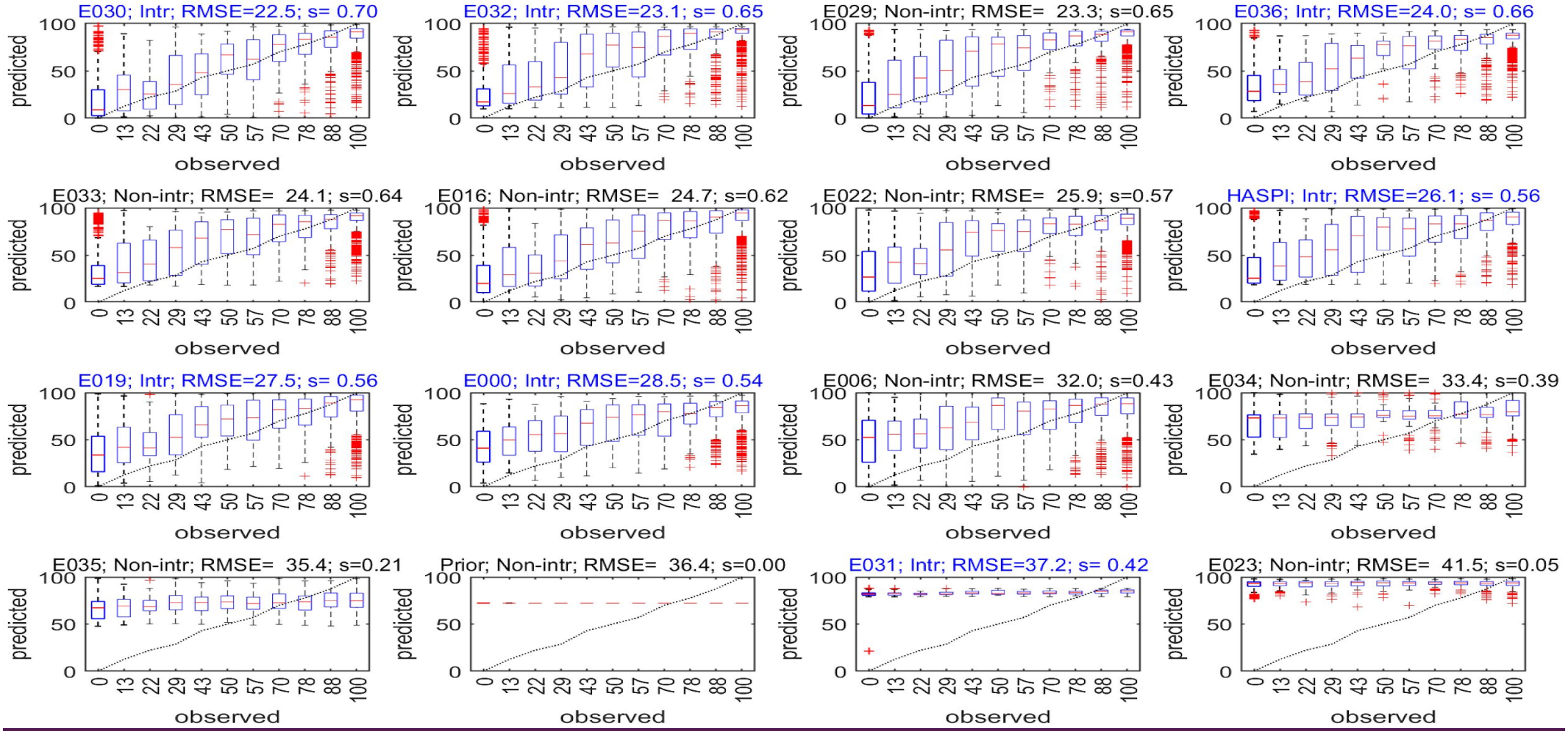



RMSE
40
35
30
25
20
Track 1
Track 2


System
RMSE
55
50
45
40
35
30
25
20
Seen
Unseen


Listeners
RMSE
60
55
50
45
40
35
30
25
Seen
Unseen

Track 1


E030; Intr; RMSE=22.5; s= 0.70
E032; Intr; RMSE=23.1; s= 0.65
E029; Non-intr; RMSE= 23.3; s=0.65
E036; Intr; RMSE=24.0; s= 0.66
E033; Non-intr; RMSE= 24.1; s=0.64
E016; Non-intr; RMSE= 24.7; s=0.62
E022; Non-intr; RMSE= 25.9; s=0.57
HASPI; Intr; RMSE=26.1; s= 0.56
E019; Intr; RMSE=27.5; s= 0.56
E000; Intr; RMSE=28.5; s= 0.54
E006; Non-intr; RMSE= 32.0; s=0.43
E034; Non-intr; RMSE= 33.4; s=0.39
E035; Non-intr; RMSE= 35.4; s=0.21
Prior; Non-intr; RMSE= 36.4; s=0.00
E031; Intr; RMSE=37.2; s= 0.42
E023; Non-intr; RMSE= 41.5; s=0.05
predicted
observed
0
50
100
0 13 22 29 43 50 57 70 78 88 100

Observations

- Lots of approaches.
- The best entrant systems had improved performance when compared to:
  - Baseline system
  - Current state-of-the-art metric (HASPI).
- Intrusive (double-ended) and non-intrusive (blind, single-ended) had similar performance.
- Listener characteristics were less useful than expected.
- Even for the best systems, the prediction errors were quite large, equivalent to getting 2 words wrong in a 9 word sentence.
- Look out for special session at Interspeech, September 2022