Background Adaptation for Improved Listening Experience in Broadcasting

- ***Yan Tang* [a], *Qingju Liu* [b], *Trevor Cox* [a], *Bruno Fazenda* [a], *Weuwu Wang* [b]**
- [a] Acoustics Research Centre, University of Salford, UK
- [b] Centre for Vision, Speech and Signal Processing, University of Surrey, UK

Intelligibility issue in broadcast

S³A

- Factors causing low speech intelligibility [1]:
  - Background sound effects
  - Intrinsically unintelligible speech
  - Unfamiliar accents
  - Loud ambient noise

- Intelligibility enhancement [2-4] :
  - Reduced perceived quality of the modified speech [2]
  - Escalated annoyance when listening to modified speech.

[1] Armstrong et al, 2015 [2] Tang & Cooke, 2018; [3] Zoril et al, 2017; [4] Jokinen et al, 2016

How about adapting the background sound(s)?

S3A

- Assumption
  - Both speech and background sound(s) are separately accessible (OBA).
  - Adapting the background sound may be less intrusive to listeners.

- Applying modification to the background signal
  - Can maintain the background level for design or artistic purposes

- Spectral weighting [1]
  - Similar to post-filtering: computationally cheap
  - Learning optimal weightings is time-consuming
  - Need a fast implementation for online processing

[1] Tang & Cooke, 2018

Spectral weighting for background

- Adaptation: to reallocate the energy of the background, *s*, across 34 frequencies on the ERB scale.

$$s'(t) = k \cdot \sum_{f=1}^{F=34} s_f(t) \cdot 10^{W_f/20},$$

s’: adapted s
k: scalar for renormalising the broadband signal energy
W_f *:* spectral weighting

- Problem: to seek for a set of optimal *W*

Factors affecting overall listening experienceca

- Perceptual guides:
  - Speech intelligibility: High-Energy Glimpse proportion (HEGP [1-3])
  - Overall audio quality: Perceptual Evaluation of Audio Quality (PEAQ [4])

- A linear combination of HEGP and PEAQ

  - k_{si}: weight for HEGP
  - k_{aq}: weight for PEAQ

$$OM = k_{si} \cdot \mathrm{HEGP} + k_{aq} \cdot \mathrm{PEAQ}, \text{ w.r.t } k_{si} + k_{aq} = 1$$

[1] Tang & Cooke, 2016 [2] Tang et al, 2018a [3] Tang et al, 2018b [4] ITU-R BS.1387

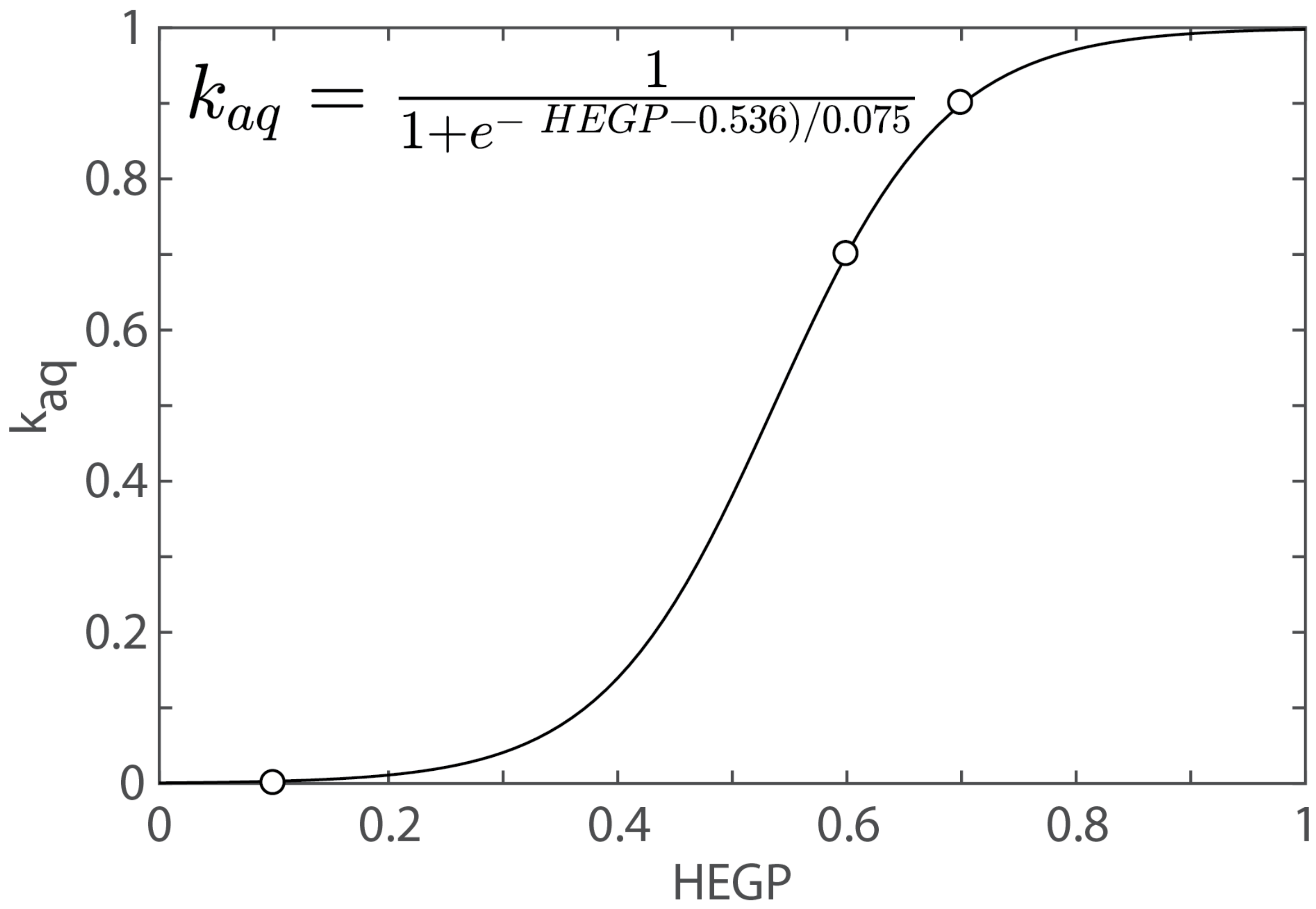
Relationship between intelligibility and audio quality

S3A

- When HEGP < 0.1, i.e. no intelligibility
  - Prioritising increasing intelligibility
  - k_{aq} = 0

  - When HEGP ≈ 0.6, i.e. threshold of full intelligibility
  - Both intelligibility and quality affect listening experience
  - k_{aq} = 0.7

- When HEGP ≈ 0.7, i.e. more favourable SNR
  - Overall quality is dominant
  - k_{aq} = 0.9



Closed-loop optimisation for spectral weightings *W*

- Task: to learn a set of optimal W_f (in dB) for each speech-background pair at a specified SNR.

- Optimisation procedure [1]
  - Algorithm: Pattern Search with MATLAB implementation
  - Variables: a vector of 34 elements, representing W_f
  - Objective function: the linear combination of HEGP and PEAQ, *OM*

- But Closed-loop optimisation is slow; not applicable for real time processing

[1] Tang et al, 2018a

Neural network implementation

- A two-hidden-layer recurrent NN with backpropagation

- Input features:
  - 34 mean log-compressed speech spectra E^s_f and 34 noise spectra E^n_f
  - 34 mean band SNRs, i.e. E^s_f - E^n_f
  - A vector of 102 elements

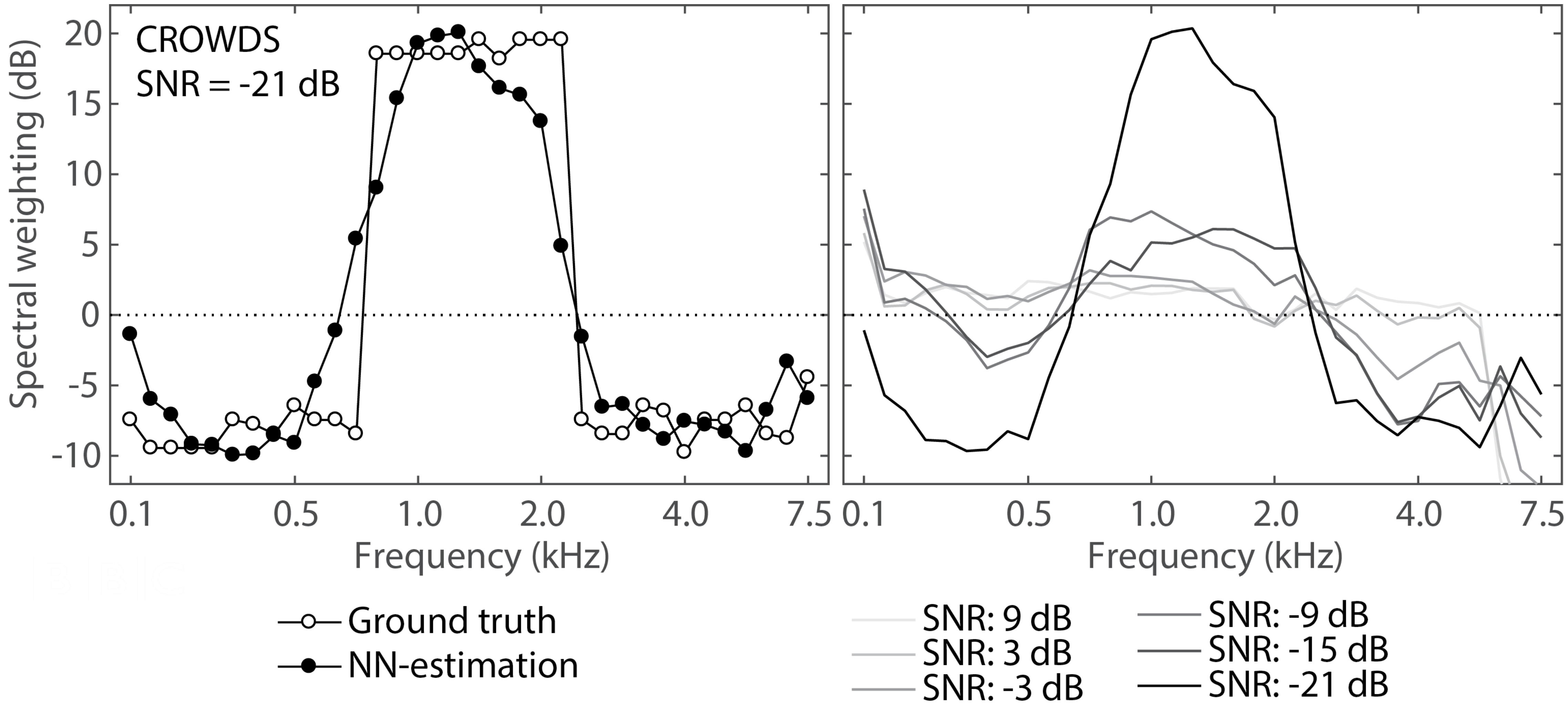
- Grand-truth: 34 optimal weightings, W_f
  - Learnt from maximising the linear combination of HEGP and PEAQ, *OM*

Experiments

- NN Training data
  - 120 Harvard sentences sampled at 16 kHz; male talker
  - 6 background sounds:
    - café noise (CAFE)
    - female competing speech (CS)
    - stadium crowd noise (CROWDS)
    - a pop song (SONG)
    - the same song with vocal being removed (SONG-VR)
    - classic music (CLASSICAL)
  - SBRs: from -21 to 9 dB with steps of 3 dB
  - 7920 samples

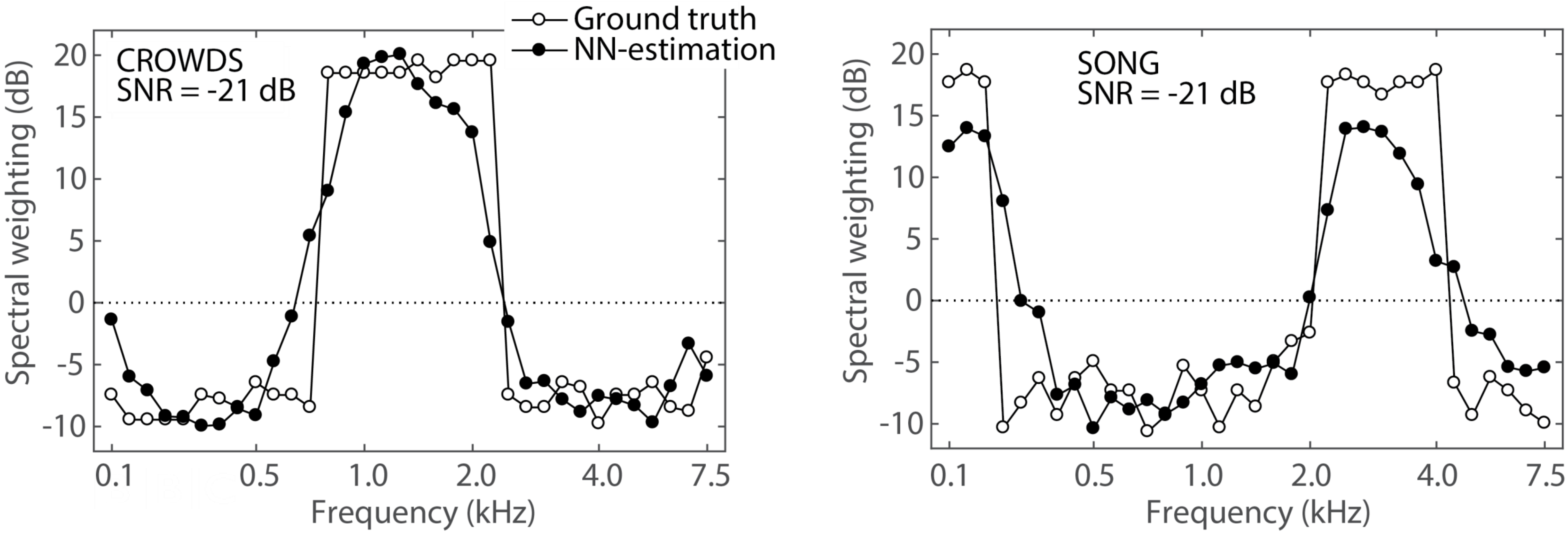
- Test data
  - 300 sentences not appearing in training
  - SBRs: from -19.5 to 10.5 dB with steps of 3 dB

Results I

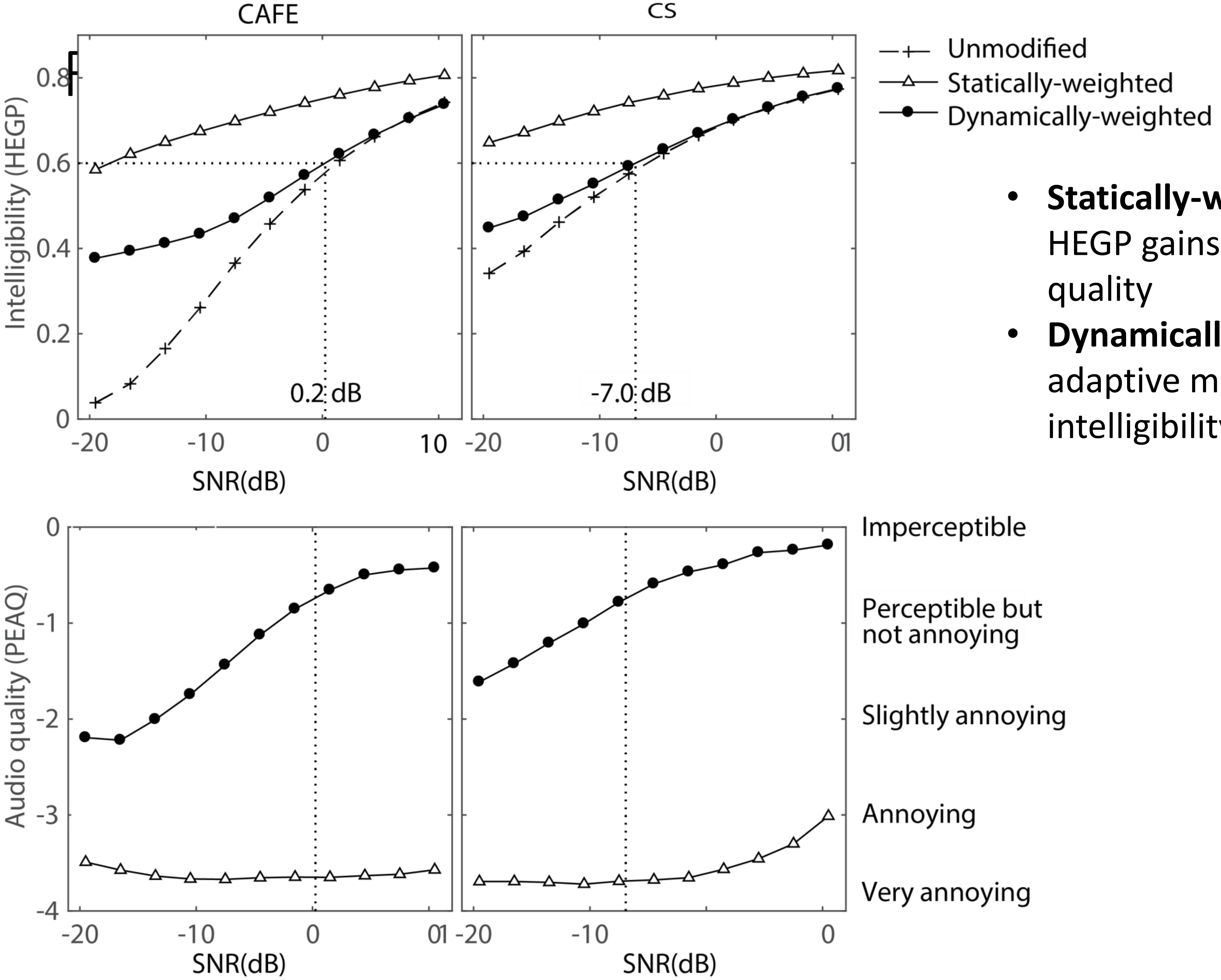


CROWDS
SNR = -21 dB
Spectral weighting (dB)
20
15
10
5
0
-5
-10
0.1
0.5
1.0
2.0
4.0
7.5
Frequency (kHz)
Ground truth
NN-estimation
SNR: 9 dB
SNR: 3 dB
SNR: -3 dB
SNR: -9 dB
SNR: -15 dB
SNR: -21 dB

Results II



Ground truth
NN-estimation
CROWDS
SNR = -21 dB
SONG
SNR = -21 dB
Spectral weighting (dB)
20
15
10
5
0
-5
-10
0.1
0.5
1.0
2.0
4.0
7.5
Frequency (kHz)



- **Statically-weighted** leads to substantial HEGP gains at cost to the overall audio quality
- **Dynamically-weighted** shows more adaptive manner in preserving both intelligibility and audio quality

Conclusions

- Spectral weighting inspired by near-end intelligibility enhancement is applied to the background signal, in order to enhance speech intelligibility while preserving the overall audio quality.

- With an adaptive function which models the relationship between intelligibility and audio quality, the optimised spectral weightings balance the two factors while modifying the background signal.

- A pre-trained NN is able to estimate the optimal spectral weightings from easy-to-compute acoustic features.

- Perceptual listening experiments are needed for further validating the method.

Thank you!

This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1) and the BBC as part of the BBC Audio Research Partnership.