



## A Method for Conversational Signal-to-Noise Ratio Estimation in Real-World Sound Scenarios

**Mansour, Naim; Marschall, Marton; May, Tobias; Westermann, Adam; Dau, Torsten**

*Published in:*  
Acoustical Society of America. Journal

*Link to article, DOI:*  
[10.1121/1.5101769](https://doi.org/10.1121/1.5101769)

*Publication date:*  
2019

*Document Version*  
Publisher's PDF, also known as Version of record

[Link back to DTU Orbit](#)

*Citation (APA):*  
Mansour, N., Marschall, M., May, T., Westermann, A., & Dau, T. (2019). A Method for Conversational Signal-to-Noise Ratio Estimation in Real-World Sound Scenarios. *Acoustical Society of America. Journal*, 145, 1873.  
<https://doi.org/10.1121/1.5101769>

---

### General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

## 1. Introduction

- Analysis of conversational signal-to-noise ratios (SNRs) measured in real-world scenarios can provide insights into communicative strategies and difficulties, and guide development of hearing devices [1].
- Measuring SNRs accurately and realistically is challenging in typical recording conditions, where only a mixture of sound sources is captured. Typical single-channel methods [2] rely on subtracting estimates of noise in a frame  $\mathbf{N}_r(\mathbf{f})$  of the recording from the mixture of speech and noise  $(\mathbf{S}(\mathbf{f})+\mathbf{N}(\mathbf{f}))_r$  (1).

$$\text{SNR}(\mathbf{f}) = \frac{(\mathbf{S}(\mathbf{f})+\mathbf{N}(\mathbf{f}))_r - \mathbf{N}_r(\mathbf{f})}{\mathbf{N}_r(\mathbf{f})} \quad (1) \quad \text{SNR}(\mathbf{f}) = \frac{\mathbf{S}_r(\mathbf{f})}{\mathbf{N}_r(\mathbf{f})} \quad (2)$$

- A novel in-situ estimation method is proposed, where the speech signal of a person in natural conversation is captured by a cheek-mounted microphone, free-field adjusted, and then convolved with a measured impulse response to estimate the clean speech receiver component  $\mathbf{S}_r(\mathbf{f})$  (2).
- The method is analyzed using in-situ recordings of a real-world workspace meeting and compared to the single-channel technique in terms of its resulting SNR distribution.

## 2. Method

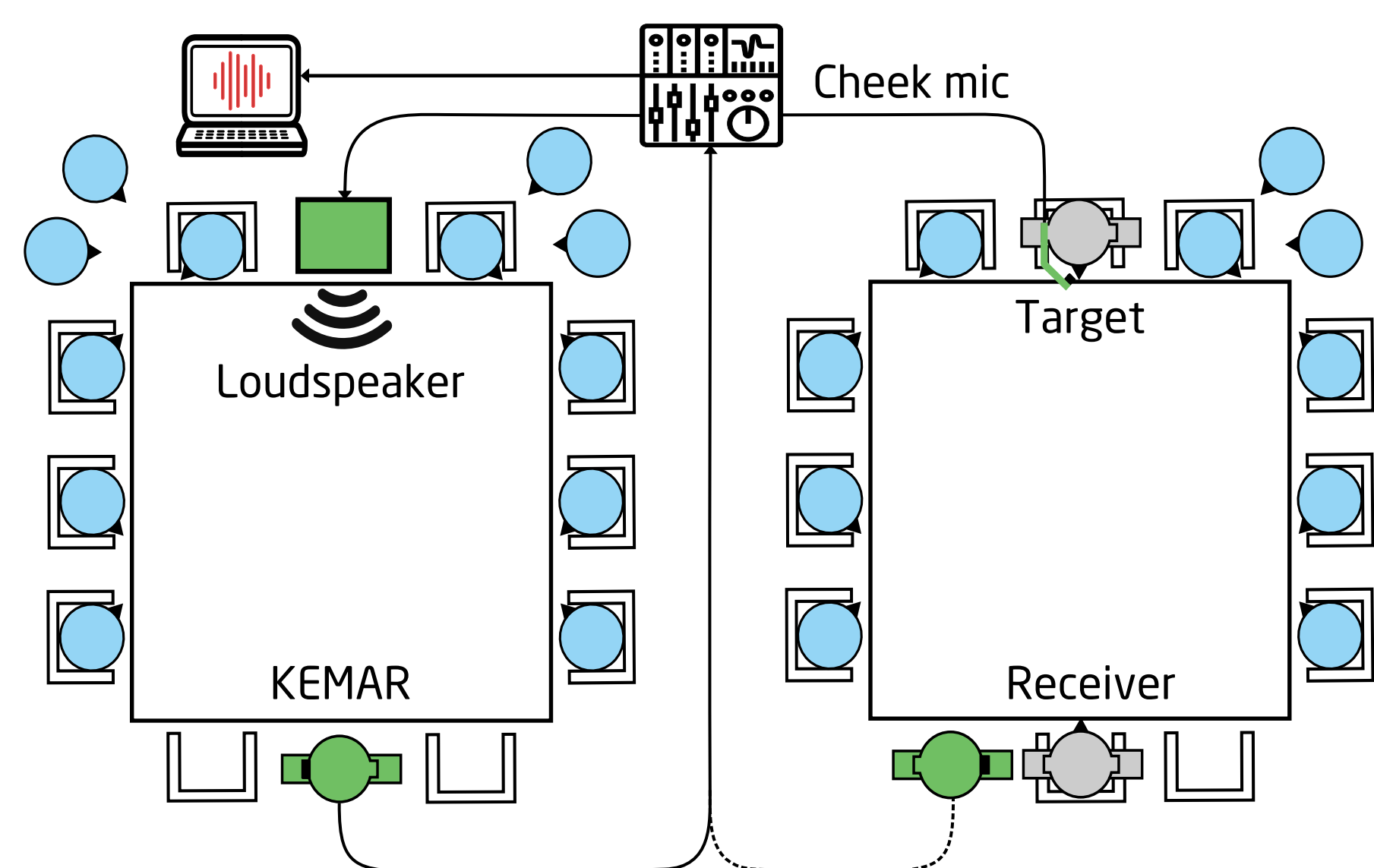
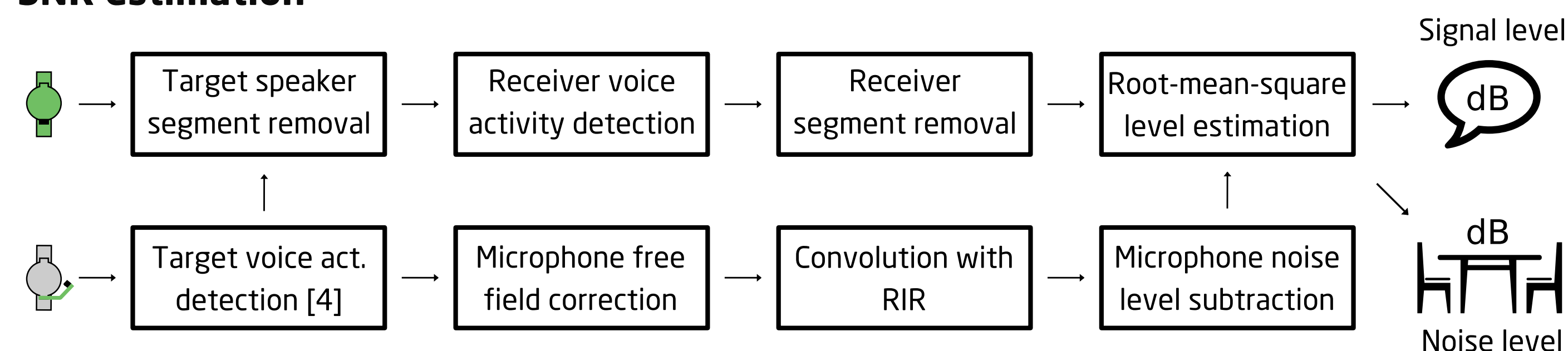


Figure 1. Left: Room impulse response recording - Right: Conversation recording (top-down view)

### Controlled recording of workspace meeting

- Realistic enactment of a typical workspace meeting in a corporate office meeting room
- Conversation between two normal-hearing (NH) people seated across a square conference table, in a background (BG) of 10 NH talkers conversing in pairs about work-related topics
- Room impulse response (RIR) between target and receiver position recorded by KEMAR [3] manikin (left ear), while BG talkers were quiet
- Cheek microphone (CM) captures source speech, KEMAR (right ear) noise at receiver

### SNR estimation



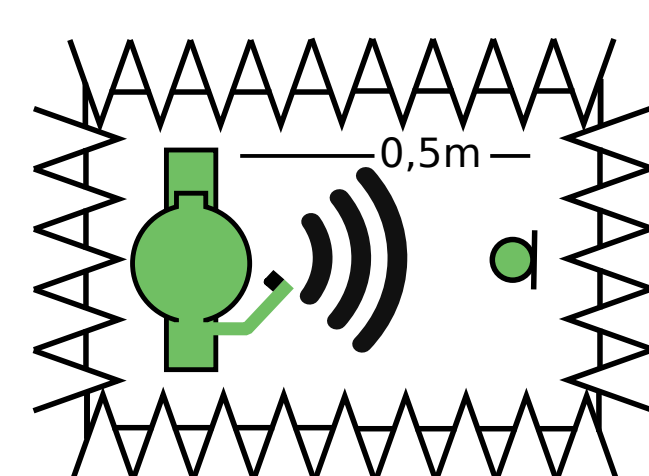
- Frame-based processing, with 10s frame length, and 90% overlap between frames
- Broad-band SNR based on root-mean-square (RMS) power within each frame

### Voice activity detection (VAD) & segment removal

- Energetic VAD [4] to separate speech from background in CM & KEMAR
- Target speaker segment removal in KEMAR recording through binary mask from CM VAD

### Microphone free-field correction

- Transfer function between CM and reference microphone to obtain free-field acoustical conditions in the target speech signal
- Recorded in anechoic chamber with KEMAR producing white noise at a level of 90dB, and reference microphone at a distance of 0.5m



### Convolution with room impulse response

- CM signal convolved with appropriately scaled RIR recorded between target & receiver

### Microphone noise level subtraction

- VAD-driven CM RMS level reduction, subtracting BG noise power in frame  $\mathbf{n}_{\text{RMS}}(\mathbf{f})$  from signal  $\mathbf{s}_{\text{RMS}}(\mathbf{f})$

$$\mathbf{s}_c(\mathbf{f}) = \mathbf{s}(\mathbf{f}) \frac{\mathbf{s}_{\text{RMS}}(\mathbf{f}) - \mathbf{n}_{\text{RMS}}(\mathbf{f})}{\mathbf{s}_{\text{RMS}}(\mathbf{f})}$$

## 3. Results

### Free-field correction & room impulse response

- IR computed from 15s exponential frequency sweep, FFC low-pass filtered at 10kHz

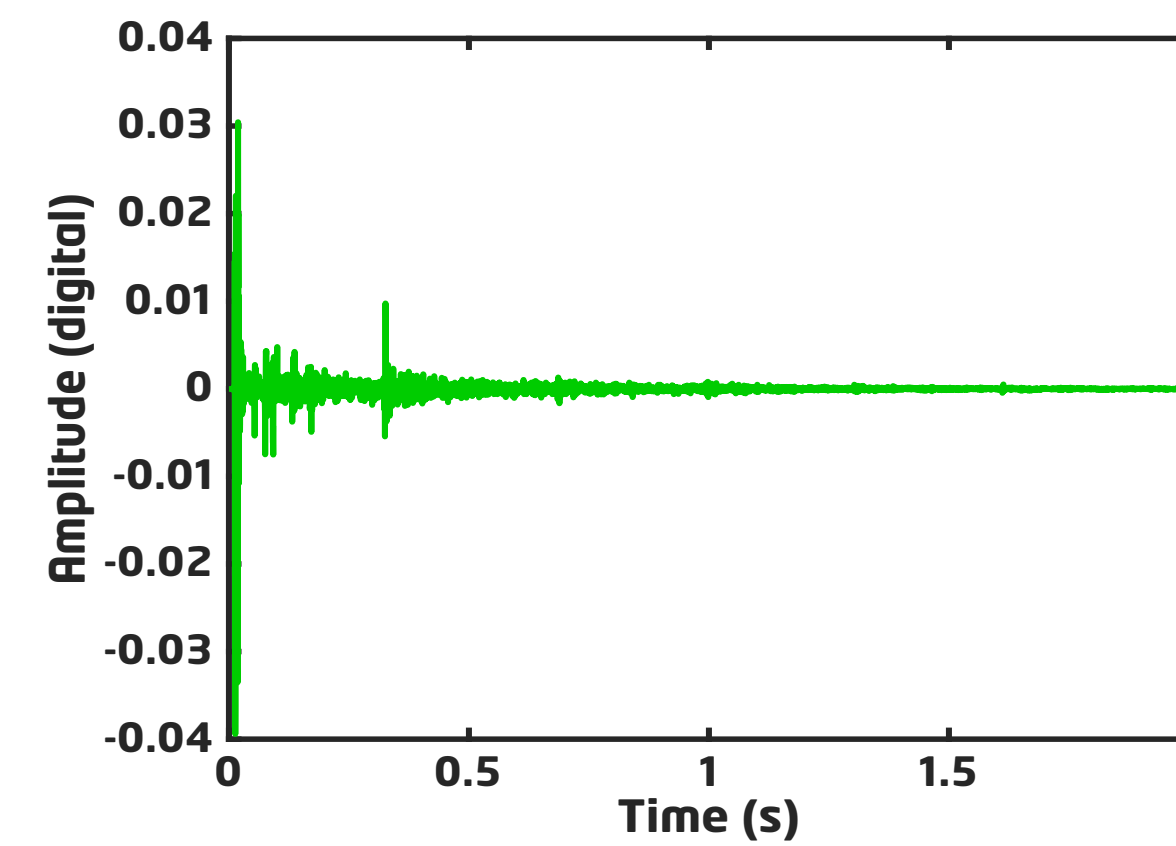


Figure 3. Impulse response

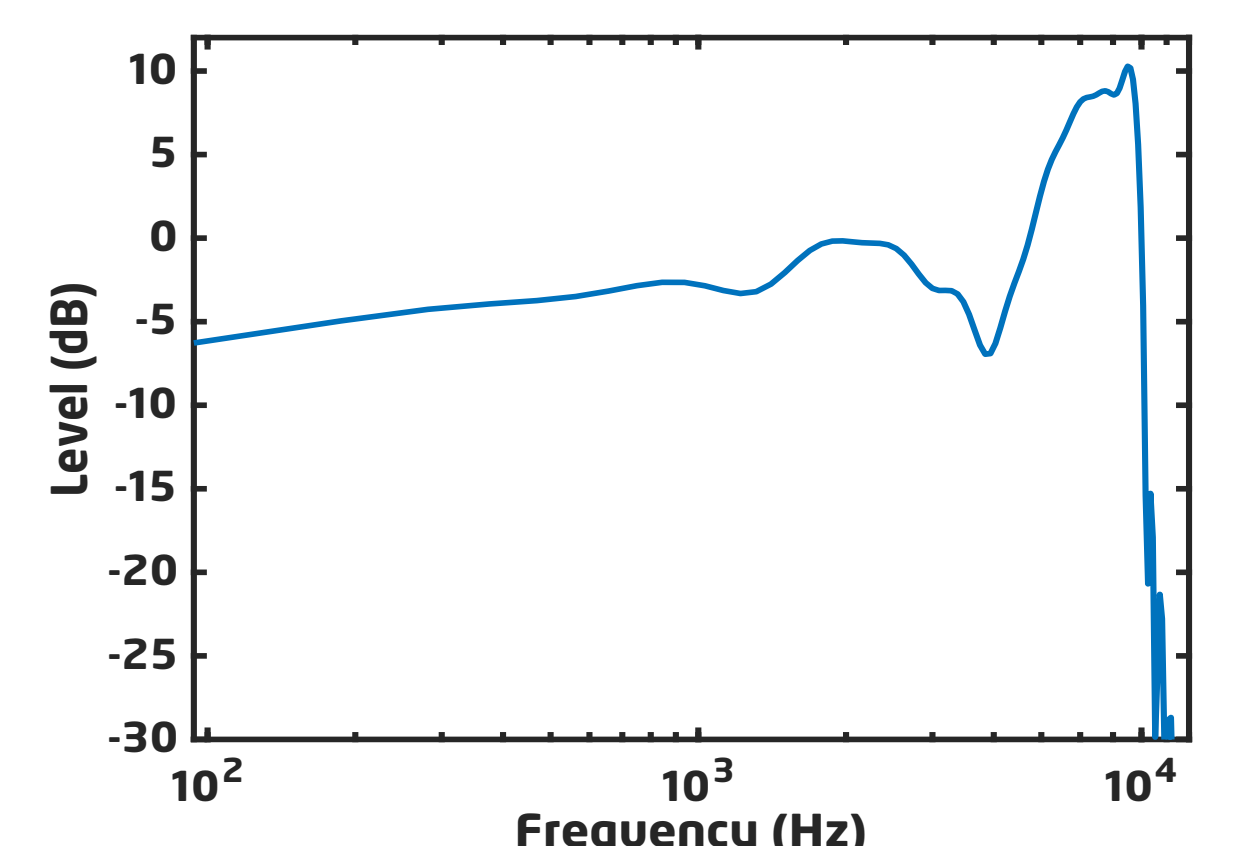


Figure 4. Free-field correction

### Speech and background levels

- Derived from 6-minute recording, temporal progression and level distributions are shown

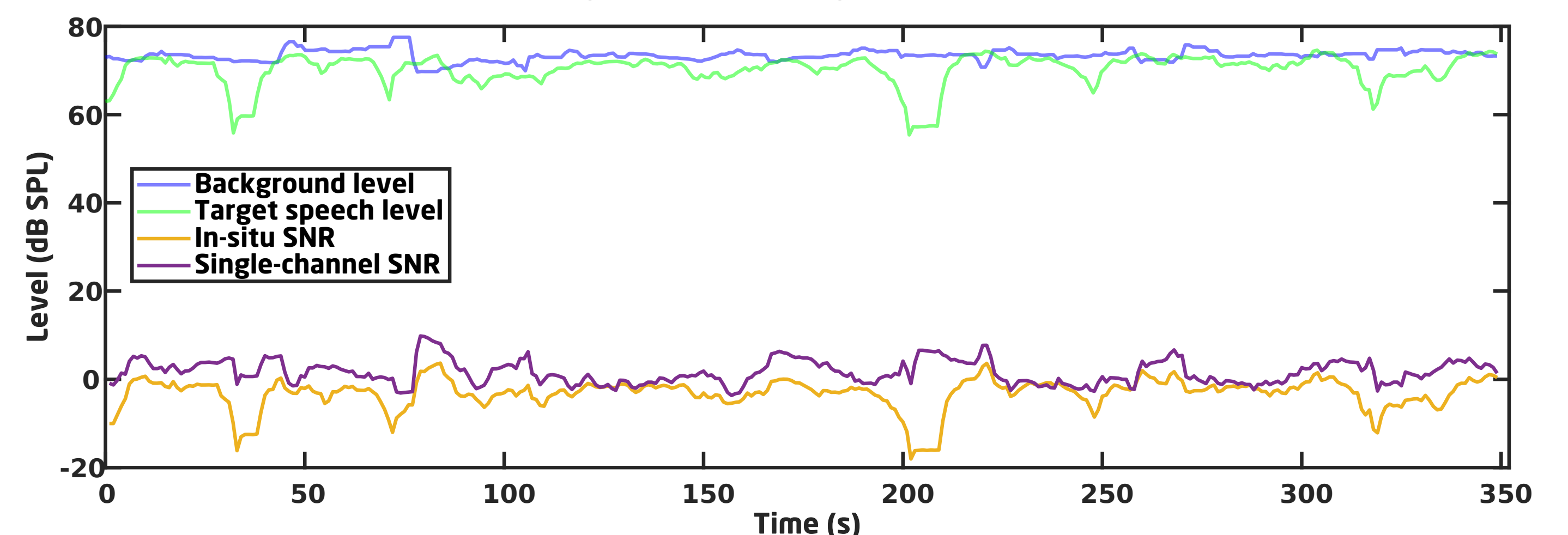


Figure 5. Levels and SNR for a 6-minute workspace meeting recording

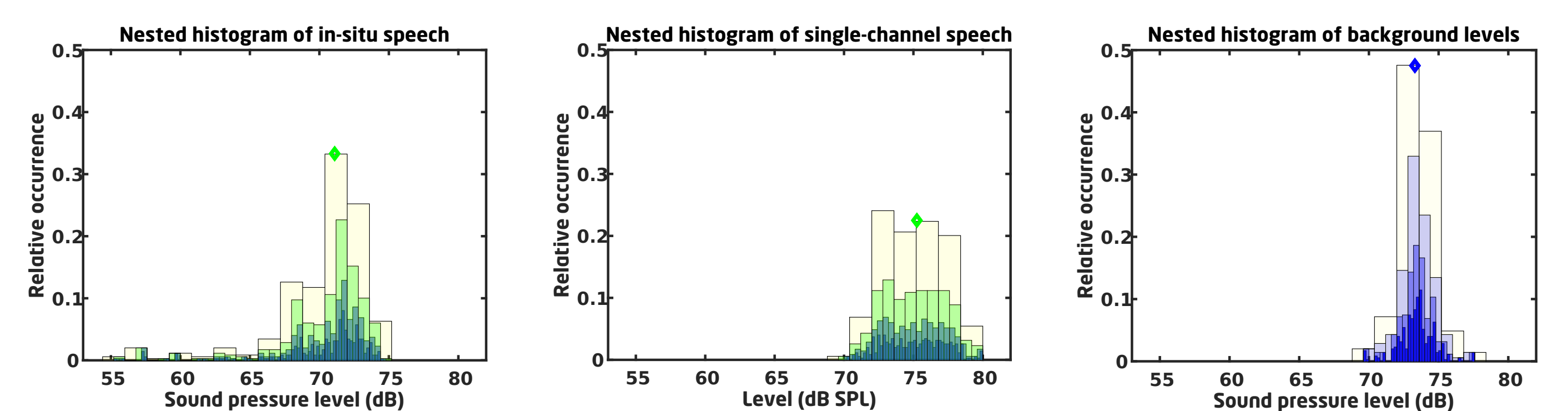


Figure 6. Nested level histograms of in-situ speech, single-channel speech and background levels

### Signal-to-noise ratios

- Computed according to respective in-situ (Eqn. 2) and single-channel (1) SNR equations

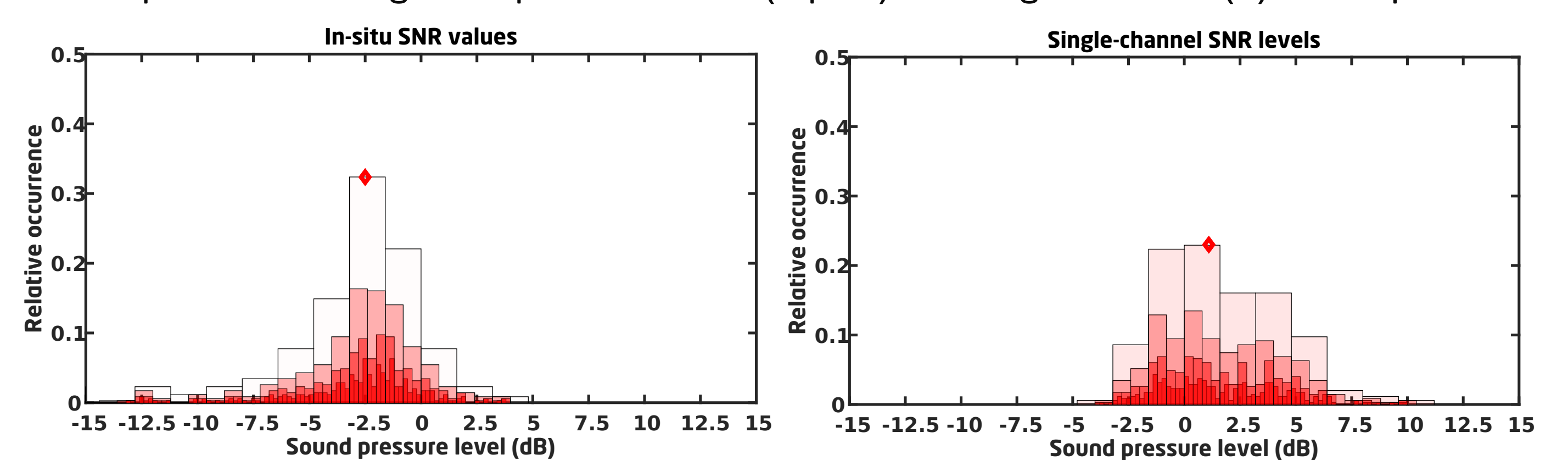


Figure 7. Nested histograms of SNR for in-situ (left) and single-channel (right) method

- Time-averaged dynamic range:

In-situ speech: **55-74 dB** Single-channel speech: **70-80 dB** Background: **70-76 dB**

- Median values, both methods:

In-situ speech: **71 dB** Single-channel speech: **75.2 dB** Background: **73.5 dB**  
 In-situ SNR: **-2.5 dB** Single-channel SNR: **1.7 dB**

## 4. Discussion & Summary

- A high temporal-resolution measurement technique for speech and background levels allows tracking of in-situ conversational SNR, even at negative values.
- The wide dynamic range found for in-situ speech levels is likely due to natural speech pauses and turn-taking during conversation.
- The obtained in-situ SNRs are lower than the corresponding single-channel SNRs, likely due to speech signal being more accurately tracked.
- The 4.2 dB difference between the median in-situ and single-channel SNR indicates a potential overestimation of SNRs with traditional techniques. This is likely due to level effects and correlations between the speech and noise when both are present.
- The in-situ approach requires the availability of a CM signal and suitable RIR recordings.
- The proposed SNR estimation method can accurately characterize in-situ SNRs. This may contribute to understanding how humans communicate in challenging environments, and help improving compensation strategies in hearing instruments.

## References

- Weisser A, Buchholz JM. Conversational speech levels and signal-to-noise ratios in realistic acoustic conditions. The Journal of the Acoustical Society of America 145(1); Jan 2019. p. 349
- Smeds K, Wolters F, Rung M. Estimation of signal-to-noise ratios in realistic sound scenarios. The Journal of the American Academy of Audiology 26(2); February 2015. p. 183-196
- GRAS Sound & Vibration, 45BC Knowles Electronic Manikin for Acoustic Research, Head & Torso with Mouth Simulator, ANSI: S3.36, S3.25, IEC: 60318-7, January 2013
- Kinnunen T, Lib H. An overview of text-independent speaker recognition: From features to supervectors. Speech Communication, 52(1); 2010. p. 12-40

## Contact

**Naim Mansour**  
 E-mail: naiman@dtu.dk  
 Phone: +45 60 56 66 57  
 Web: www.naim-mansour.be