A method for realistic, conversational signal-to-noise ratio estimation

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

Published in:
Journal of the Acoustical Society of America

Link to article, DOI:
10.1121/10.0003626

Publication date:
2021

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

Citation (APA):
A method for realistic, conversational signal-to-noise ratio estimation

Naim Mansour,1,a) Marton Marschall,1,b) Tobias May,1 Adam Westermann,2 and Torsten Dau1,c)

1Hearing Systems Section, Department of Health Technology, Technical University of Denmark, 2800 Kongens Lyngby, Denmark
2Widex Aktieselskab, Lynge, Denmark

ABSTRACT:
The analysis of real-world conversational signal-to-noise ratios (SNRs) can provide insight into people’s communicative strategies and difficulties and guide the development of hearing devices. However, measuring SNRs accurately is challenging in everyday recording conditions in which only a mixture of sound sources can be captured. This study introduces a method for accurate in situ SNR estimation where the speech signal of a target talker in natural conversation is captured by a cheek-mounted microphone, adjusted for free-field conditions and convolved with a measured impulse response to estimate its power at the receiving talker. A microphone near the receiver provides the noise-only component through voice activity detection. The method is applied to in situ recordings of conversations in two real-world sound scenarios. It is shown that the broadband speech level and SNR distributions are estimated more accurately by the proposed method compared to a typical single-channel method, especially in challenging, low-SNR environments. The application of the proposed two-channel method may render more realistic estimates of conversational SNRs and provide valuable input to hearing instrument processing strategies whose operating points are determined by accurate SNR estimates.

© 2021 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/). https://doi.org/10.1121/10.0003626
(Rceived 5 September 2020; revised 20 January 2021; accepted 11 February 2021; published online 5 March 2021)

I. INTRODUCTION

Speech communication is a complex phenomenon that combines auditory, visual, and cognitive processes to enable people to transmit and receive information. Such a conversation often occurs in noisy backgrounds in which a speech source of interest, i.e., the target talker signal, is accompanied by interfering sources (e.g., noise or competing talkers) and reverberation. Levels of conversational speech have been shown to strongly depend on the background noise level as people raise their voices in increasingly loud surroundings to remain intelligible (Lombard, 1911). At the same time, the ratio of the average speech power arriving at the listener to the power of the background noise, i.e., the signal-to-noise ratio (SNR), is known to decrease at a fixed talker distance when the background noise level increases, that is, people do not continue to increase their speech power indefinitely (Weisser and Buchholz, 2019).

Knowledge of the SNR distributions that occur in real-world conversations is important because these SNRs affect a person’s ability to understand speech in noisy environments. Developing more realistic listening tasks, therefore, demands accurate estimates of real-world speech levels and corresponding SNRs. Furthermore, the processing of hearing aids (HAs) strongly depends on the input signal levels. For example, the output SNR of a fast-acting dynamic range compression system depends on the input SNR, potentially impacting HA performance (Naylor and Johannesson, 2009). Accurate conversational SNR estimates would allow a HA to be tailored to the environment of its user (May et al., 2018).

Several studies have focused on the estimation of real-world SNRs. Specifically with regard to broadband, long-term estimates of conversational SNRs, two notable studies exist. In one study, Pearsons et al. (1977) recorded conversations between two normal-hearing (NH) talkers at the ear of one of the participants in a diverse range of conditions, selected by the researchers. In the study by Smeds et al. (2015), HA recordings (Wagener et al., 2008) obtained by HA users in various situations of their daily lives were analyzed. Figure 1 shows the resulting broadband SNR distributions of the two studies (adapted from Wu et al., 2018). The blue and red bars represent the results from Pearsons et al. (1977) and Smeds et al. (2015), respectively. The purple shade indicates areas where the distributions overlap.

Both distributions reveal mostly positive SNRs across listening situations. The Pearsons et al. (1977) distribution is shifted slightly toward lower SNRs compared to the Smeds et al. (2015) distribution, most likely because Pearsons et al. collected data from NH participants who commonly communicate relatively easily at lower SNRs and may, therefore, not avoid such challenging acoustic conditions, unlike the HI participants (even if aided) in the Smeds et al. study.

Although there were differences between the studies in terms of the methodology and hearing status of the participants, the SNRs were estimated in a similar way, using...
recordings made with a single microphone at the receiver position. Specifically, the root-mean-square (RMS) level of the clean speech was estimated by subtracting the average power of the noise-only segments from the average power of the noisy speech. These speech-in-noise and noise-only segments were hand-labeled by a human listener. The SNR was then obtained by dividing the estimated speech power by the noise-only power. This approach assumes that the speech and noise components in the recording are uncorrelated and the estimated noise power in the noise-only segments reflects the noise power in the speech and noise segments. Both assumptions do not necessarily hold in real-world conditions with multiple interacting talkers in fluctuating background noise. Furthermore, it has been shown that at sufficiently negative SNRs when the speech power becomes indistinguishable from the random fluctuations in the noise power, this single-channel approach no longer provides accurate estimates because the SNR distribution essentially reflects the magnitude distribution of those fluctuations (Kim and Stern, 2008). In practice, the method relies on the accurate labeling of speech-in-noise and noise-only segments, which may become inaccurate at low SNRs.

Here, a two-channel method is proposed to estimate real-world, in situ conversational SNRs. The method extends the single-channel approach by introducing a cheek-mounted lavalier microphone to accurately capture the speech-only component of the target talker in addition to the microphone at the receiver. A free-field correction (FFC) and a room impulse response (RIR) convolution were applied to this cheek microphone recording to obtain the target-speech-only signal at the receiving talker. From this signal, the SNR of the target talker at the receiver was derived by division with a noise-only signal, recorded at the ear of a mannequin standing next to the receiver. Accurate target speech labeling was employed based on the high-SNR cheek microphone signal, allowing for a reliable selection of segments where target speech was present even in challenging situations containing speech-on-speech masking. The two-channel method was evaluated in room acoustic simulations of two real-world scenes where theoretical, “true” SNR estimates could be calculated and compared to the single-channel approach of Pearsons et al. (1977) and Smeds et al. (2015). In addition, both methods were evaluated for real-world recordings in the same two scenes.

II. METHODS

A. SNR estimation principle

Figure 2 illustrates the conversational SNR estimation of a speech signal $S$ produced by a target talker $T$ at the location of a receiver $R$ (red icons) in the presence of background noise $N$ (blue rectangle). All signals are expressed in the frequency domain. $S_R$ denotes the speech signal of the target talker at the position of the receiver. The true SNR, $\text{SNR}_{\text{True}}$, is the ratio between the average power of $S_R$, $P(S_R)$, and the receiver noise-only power, $P(N)$,

$$\text{SNR}_{\text{True}} = \frac{P(S_R)}{P(N)}.$$  \hfill (1)

Neither $P(S_R)$ nor $P(N)$ can be measured in a real scene because the target speech is mixed with the background noise by the time it arrives at the receiver. As illustrated in Fig. 2(A), a typical single-channel method uses a single receiver microphone $M_R$ (green circle) to approximate $P(S_R)$ as $\hat{P}(S_R)$ by capturing the noisy target speech power at the receiver $P([S + N]_R)$ and subtracting an estimate of the noise power $\hat{P}(N)$ from it. $\hat{P}(N)$ is obtained by estimating the noise power in speech gaps where the target talker

![FIG. 1. (Color online) Distributions of speech-in-noise SNRs from Pearsons et al. (1977), indicated by the blue bars, and Smeds et al. (2015), indicated by the red bars. The purple shade indicates areas of overlap between the two distributions.](Image 77x593 to 287x738)

![FIG. 2. (Color online) Conversational SNR estimation principle of a target talker $T$ and their speech signal $S$ at a receiver $R$ (red icons) in a real-world containing background noise $N$ (blue rectangle) for a single-channel method yielding $\text{SNR}_{\text{1ch}}$ (A) and the proposed two-channel method yielding $\text{SNR}_{\text{2ch}}$ (B). $S_R$ denotes the speech signal of the target talker at the position of the receiver. $M_C$ and $M_R$ represent a cheek microphone and receiver microphone, respectively. $H_{\text{FFC}}$ denotes the transfer function between $T$ and $R$, made up of the transfer function between $T$ and $M_C$, $M_C$, a free-field correction (FFC) transfer function $H_{\text{FFC}}$, and a room impulse response transfer function $H_{\text{RIR}}$.](Image 327x160 to 537x340)
and receiver are silent. Division of \( \tilde{P}(S_R) \) by \( \tilde{P}(N) \) then yields the single-channel SNR,

\[
SNR_{1ch} = \frac{\tilde{P}(S_R)}{\tilde{P}(N)} = \frac{\tilde{P}([S + N]_R) - \tilde{P}(N)}{\tilde{P}(N)}.
\]  

(2)

The proposed two-channel method, illustrated in Fig. 2(B), estimates \( P(S_R) \) directly by applying the room acoustic transfer function between \( T \) and \( R \), \( H_{TR} \), to \( S \). To account for \( H_{TR} \), a cheek-mounted microphone (green stick) worn by the target talker \( M_{CM} \) was used to capture the target speech (\( H_{CM} \)). Next, a fixed FFC transfer function \( H_{FFC} \), measured at a distance of 0.5 m, was applied to the recorded target speech to correct for near-field and head scattering effects due to the close distance of \( M_{CM} \) to the mouth of the target talker. Finally, convolution with an \textit{in situ} measured RIR, measured between \( T \) and \( R \) and calibrated to account for the attenuation caused by \( H_{FFC} \), resulted in \( S_R(H_{RIR}) \). Division of the average power of \( S_R \) by \( \tilde{P}(N) \), estimated in the same way as for the single-channel method, then yielded the two-channel SNR,

\[
SNR_{2ch} = \frac{P(S_R)}{\tilde{P}(N)} = \frac{P(S \cdot H_{TR})}{\tilde{P}(N)} = \frac{P(S \cdot H_{CM} \cdot H_{FFC} \cdot H_{RIR})}{\tilde{P}(N)}.
\]  

(3)

Assuming that \( M_{CM} \) captures negligible background noise and the speech power is the same at \( R \) and \( M_R \), \( S_R \) can be obtained by the two-channel method. This is the main difference from the single-channel method and implies that the only deviations to \( SNR_{true} \) will be caused by the approximation \( \tilde{P}(N) = P(N) \) if the assumptions for the speech signal, mentioned above, are fulfilled. This approximation for the noise power only holds if \( N \) is isotropic in space between \( R \) and \( M_R \) and stationary over time. In addition, the two-channel method allows for an accurate detection of the target talker speech segments even at low SNRs by using a voice activity detector (VAD) applied to the \( M_{CM} \) signal, which is not possible with the single-channel method.

In the following, each step in the proposed method is outlined in detail. All signals were sampled at a rate of 48 kHz and a resolution of 24 bits. Levels of speech and background noise, as well as SNRs, were derived from their broadband average power in dB.

**B. Microphone measurements and voice activity detection**

The cheek microphone (DPA 4066, DPA Microphones, Lilered, Denmark) used to capture the target speech signal \( S \) was mounted at a 5-cm distance next to the target talker’s mouth, representing \( H_{CM} \). It was assumed that at this distance, the power in the speech signal picked up by \( M_{CM} \) could be entirely attributed to \( S \) and the dynamic range of the signal would be sufficient to accurately separate target speech segments. Energy-based VADs (Kinnunen and Li, 2010) were applied to both the \( M_{CM} \) and \( M_R \) signals. The obtained binary speech detection masks were used to exclude the speech of \( R \) and the noise \( N \) from the signal in \( M_{CM} \) and exclude the speech of \( T \) and \( N \) from the signal in \( M_R \). The VAD applied to \( M_{CM} \) estimated the short-term energy of \( S \) by segmenting the recording into frames of 20 ms duration and subsequently applying a threshold to this short-term energy, relative to its maximum value, to identify frames which contained relevant target activity. This threshold was set to the difference in dB between the 95th and 50th percentiles of the short-term energy to adaptively separate the target speech energy distribution (peaking in the 95th percentile) from the background noise distribution (assumed to be distributed around the 50th percentile). Speech gaps longer than 200 ms (Demol et al., 2007) were not considered to be part of \( T \), ensuring that the estimated speech power would not be affected by silence gaps.

The right ear microphone of a Knowles Electronic Manikin for Acoustic Research (KEMAR, GRAS Sound and Vibration A/S, Holte, Denmark) mannequin with ear canals was used as \( M_R \) to estimate the noise-only signal \( N \) in a way that captures the effects of the head and pinnae shape present in human listening. The receiver speech was then removed using the same VAD applied directly to the \( M_R \) signal but with a fixed threshold energy at 15 dB below the global maximum of the short-term energy, equal to the lower speech range boundary used in the computation of the speech transmission index (Houtgast et al., 1980). A fixed threshold was used in \( M_R \) but not in \( M_{CM} \). The target speech \( S \) contained in \( M_{CM} \) had a larger and more strongly varying dynamic range between frames than the receiver speech in \( M_R \) due to the closer proximity of \( M_{CM} \) to \( T \). This required an adaptive threshold to ensure the proper detection of the target speech. As was verified, applying a fixed threshold to the \( M_{CM} \) signal would have resulted in an underestimation of speech activity. The \( M_{CM} \) and \( M_R \) recordings were time-aligned to compensate for the acoustic delay through cross correlation (Stoica, and Moses, 2005), allowing for the usage of both VAD masks in both microphone signals to remove \( R \) speech and \( T \) speech, respectively.

**C. FFC**

The near-field signal produced by the target talker’s mouth was corrected for free-field conditions using the measurement setup illustrated in Fig. 3. The transfer function \( H_{FFC} \) between the position of \( M_{CM} \) mounted on the KEMAR (mannequin icon) and that of a reference pressure field microphone \( M_{REF} \) (GRAS AG40, GRAS Sound and Vibration A/S, Holte, Denmark), positioned upright at a distance of 0.5 m to the KEMAR, was measured inside an anechoic chamber. The KEMAR mouth simulator produced white noise, recorded by \( M_{CM} \) as \( W_{CM} \), at a sound pressure level (SPL) of 90 dB at \( M_{REF} \)’s position, recorded as \( W_{REF} \). A frequency-domain transfer function \( H_{FFC} \) was then derived from the ratio of the frequency-dependent cross-power spectral density of \( W_{CM} \) and \( W_{REF} \), \( P(\hat{W}_{CM}, \hat{W}_{REF}) \) and the auto-power spectral density of \( W_{REF} \) and \( W_{REF} \), \( P(\hat{W}_{REF}, \hat{W}_{REF}) \).
and applying Hamming windowing to obtain $h$ to avoid overfitting
the critical bands of the human auditory system, to
smoothened magnitude response, consisting of
impulse response (FIR) filter was designed using the
correct for the main features of the transfer function.
compared to evaluate that the chosen filter length was suffi-
tent sinusoidal sweeps from 20 Hz to 20 kHz, played by a
talker with the KEMAR and recording 15-s-long exponen-
tial sinusoidal sweeps from 20 Hz to 20 kHz, played by a
two-way loudspeaker (KEF R3, KEF Audio, Maidstone,
UK) placed in the target talker position (green rectangle).
The sweep was played in a quiet background (interfering
speakers and background were silent) at a level of 90 dB
broadband SPL measured at $R$.

FIG. 4. (Color online) The magnitude response of the transfer function $H_{FFC}$ and its smoothed version $\sqrt{G_S(H_{FFC})^2}$, between 100 Hz and 24 kHz.

The $M_{CM}$-$M_{REF}$ distance of 0.5 m was chosen to ensure a
high dynamic range in WN$_{REF}$ despite the power limitations
of the mouth simulator. This resulted in highly coherent
input signals to both microphones as is necessary for reliably estimating $H_{FFC}$.

D. Real-world measurement setup and RIR measurement

The two-channel SNR measurement setup was realized
in two real-world environments: an office meeting and a
public lunch scenario. Figures 5(A) and 5(B) show a top-
down illustration of the measurement setup. In the office
meeting, 12 NH participants were present, seated and standing
around a large square table, in a typical office conference
room of approximately 25 m$^2$. The participants were co-workers who knew each other well. They were asked to
converse naturally in pairs for a period of 5 min about everyday topics, provided to them on a list, to generate the background noise (blue icons) while the male target $T$ and receiving talker $R$ (red icons) were having the conversation of interest at a distance of 2.4 m. Both the cheek microphone $M_{CM}$ and the right ear of the KEMAR $M_R$ were connected to a sound card (Fireface 800, RME, Haimhausen, Germany) controlled by a laptop. The $M_{CM}$ and $M_R$ inputs were clock-synchronized to sample precision. The setup was similar in the lunch scenario except that the 12 participants were now seated at narrower lunch tables in a large open-plan canteen of approximately 800 m$^2$, and the $T$-$R$ distance was only 1 m. The single-channel SNR estimation method was applied in both scenes as well, using only the $M_R$ recording. However, it used the VAD masks derived by the two-channel method to classify $S_R$ and $N$ segments in the $M_{CM}$ and $M_R$ signals, ensuring manual labeling errors would not affect the classification performance.

For both the single-channel and two-channel SNR analyses, the input recordings were divided into frames of 5 s with a 1-s shift between frames to obtain 294 SNR estimates within the 5-min-long recordings. These values were chosen to ensure a sufficient number of speech and noise samples within a frame and smooth transitions between frames while maintaining the same average frame length that was used in the single-channel reference studies. Frames that contained only speech or only noise samples were excluded from the calculation. The speech and noise stimulus levels were calculated by computing digital RMS values and converting to SPLs.

Because the RIR transfer function $H_{RIR}$ depends on the acoustic surroundings, it was measured in situ in both sound environments. As illustrated in Fig. 5(C), the RIR between $T$ and $R$ (red icons) was obtained by replacing the receiving talker with the KEMAR and recording 15-s-long exponential sinusoidal sweeps from 20 Hz to 20 kHz, played by a two-way loudspeaker (KEF R3, KEF Audio, Maidstone, UK) placed in the target talker position (green rectangle). The sweep was played in a quiet background (interfering speakers and background were silent) at a level of 90 dB broadband SPL measured at $R$.

FIG. 3. (Color online) The FFC measurement setup, including the KEMAR with mounted cheek microphone $M_{CM}$ and reference microphone $M_{REF}$ at a 0.5 m distance inside an anechoic enclosure. WN and WN$_{REF}$ denote the white noise stimulus at the position of $M_{CM}$ and $M_{REF}$, respectively. $H_{FFC}$ denotes the transfer function between $M_{CM}$ and $M_{REF}$.

$$H_{FFC} = \frac{P(W_{CM}, W_{REF})}{P(W_{REF}, W_{REF})}.$$  \hspace{1cm} (4)

$H_{FFC}$ was smoothed in the frequency domain over critical bands using a fourth-order gammatone kernel $G_S$, resembling the critical bands of the human auditory system, to avoid overfitting $H_{FFC}$ to the exact $M_{CM}$ position and head shape that was used in the measurement. The original and smoothed magnitude responses of $H_{FFC}$ are plotted between 100 Hz and 24 kHz in Fig. 4. Finally, a linear-phase finite-impulse response (FIR) filter was designed using the smoothed magnitude response, consisting of $n = 256$ taps and applying Hamming windowing to obtain $h_{FFC}[n]$ as the time-domain representation of $H_{FFC}$.

$$h_{FFC}[n] = \text{FIR} \left[ \sqrt{G_S(|H_{FFC}|^2)} \right].$$  \hspace{1cm} (5)

The target and realized filter magnitude responses were compared to evaluate that the chosen filter length was sufficient to correct for the main features of the transfer function.
FIG. 5. (Color online) (A) The conversation-in-noise recording setup in the office meeting scenario, including the target T and receiving talker R (red), M_CM and M_RT (green), and other participants (blue). (B) The conversation-in-noise recording setup similar to (A) for the public lunch scenario. (C) Illustration of the RIR measurement setup in the office meeting scenario in the presence of all participants (blue and red), with the loudspeaker (top, green), and M_RT (bottom, green) producing and capturing the excitation signal, respectively.

Because the RIR was recorded between T and R, it had to be calibrated to account for the 0.5 m attenuation of S after convolution with $H_{FFC}$. During the calibration stage, the target talker was asked to speak at a conversational level to the receiver [in the same configuration as in Fig. 5(C)] in quiet. In the absence of noise ($N = 0$), the power of the recorded $M_R$ signal, $P(S + N)|_R$, is equal to $P(S)|_R$. A scaling factor $z$ was applied to $H_{RIR}$, set such that the speech levels measured at the receiver $[P(S)|_R]$ and derived from the $M_{CM}$ signal $[P(SH_{CM}H_{FFC}2H_{RIR})]$ were equal.

E. Simulated and real-world validation

To compare the SNR_2ch with SNR_1ch and SNR_True, room acoustic simulations of the two real-world scenes were constructed (further denoted by the suffix “Sim” appended to a variable name). True SNR distributions around a desired median value were established by modeling the target speech with an anechoic source $S$, convolved with the $H_{RIR}$ measured in the two real-world scenes to obtain $S_R$. This $S_R$ signal was scaled and superimposed on an $N$ signal, modeled by the noise-only $M_R$ recordings made in the two real-world scenes, to obtain $[S + N]|_R$. $R$ and $M_R$ were assumed to be in the same position. The target speech source consisted of 30 concatenated, anechoic sentences from the Danish Hearing in Noise Test (HINT) corpus. These male-spoken sentences were, on average, 1.5 s long and separated by silence gaps set to 1 s, the average silence gap length in the real-world version of the target speech. A 5-second frame length and 1-second shift were used to process the signals. The two-channel method was simulated at a median SNR_True by using $S$ and $[S + N]|_R$ as inputs; the single-channel method only had access to $[S + N]|_R$. The two-channel method’s calibration procedure was simulated by setting the $N$ signal in $[S + N]|_R$ to zero.

The simulations assumed $S$ as recorded by $M_{CM}$ to be anechoic (as a result of the use of the HINT corpus) and $N$ to be isotropic (because of the assumption that $M_R$ was in the same position as $R$). Since these assumptions may not entirely hold true in the real world, comparing simulation results to actual measurements is crucial. Whereas SNR_True, by definition, could not be determined in the real-world scenes, differences between SNR_2ch and SNR_1ch were compared between the measurements and simulations. In addition, comparisons were made between the measured SNR_2ch and SNR_1ch and the simulated SNR_2chSim, SNR_1chSim, and SNR_True by matching the measured SNR_1ch distributions to their simulated counterparts SNR_1chSim at their median.

III. RESULTS

The results described below reflect the outcome of the room acoustic simulations, evaluating the performance of the single-channel and two-channel estimation methods compared to the true SNR in the office meeting and public lunch background noise. The in situ measurement results relate the different methods to each other in a real-world application.

A. Room acoustic properties

Table I displays the main room acoustic parameters that characterize the office meeting and public lunch scenarios based on the analysis (Hummersone, 2020) of the early decay characteristics of the measured RIRs: the reverberation time at 1 kHz ($RT_{60}$), the direct-to-reverberant ratio (DRR), the clarity ($C_{50}$), and early decay time at 1 kHz (EDT).

The office meeting room had a dry response (low $RT_{60}$) of 0.4 s with a considerable amount of early reflections (high EDT) and a relatively small direct sound contribution (low DRR) at the receiver position. In contrast, the large public lunch space contained considerable reverberation (high $RT_{60}$) and showed a relatively fast decay of early reflections.
and an increased DRR. These room acoustic parameters reflect the differences in the physical layout of the two scenarios. The office meeting space was a typical conference room with a carpeted floor, two glass walls, and a suspended ceiling, all of which contribute to the low reverberation time. The public lunch took place in a large open-spaced canteen with multiple highly reflective surfaces contributing to increased reverberation. The larger distance of 2.4 m between the target and receiver in the small office meeting room implied that multiple pronounced early reflections reached the receiver at different times after the direct sound, increasing the EDT and subsequently reducing the DRR and C50. Conversely, the target-receiver distance of only 1 m in the public lunch space resulted in a much more prominent direct sound component with sparse early reflections due to the size of the space as evident through the low EDT and increased DRR and C50.

### B. Room acoustic SNR simulations

Figure 6(A) displays box plots of the true SNR distributions (SNR True, red), simulated at specified median SNRs between -16 dB and 10 dB in steps of 2 dB as well as the corresponding SNR distributions obtained by simulating the single-channel (SNR1ch, blue) and the two-channel (SNR2ch, green) methods for the office meeting scenario. Figure 6(B) shows the corresponding simulated distributions for the public lunch scenario. A one-way analysis-of-variance test showed a significant effect of the applied method in both scenes across all SNRs with the single-channel method resulting in significantly increased SNRs compared to both the two-channel method and the true SNR ($p \leq 0.0001$ for all comparisons). The difference increased with decreasing SNRs as the single-channel distributions flattened out around -10 dB SNR. The two-channel distributions were not significantly different from the true SNR distributions ($p = 0.77$ and $p = 0.87$ for the office and public lunch scenario, respectively) but slightly more spread out, especially for the public lunch scenario.

### C. Real-world speech and background levels, SNR

Figure 7(A) shows the $S_R$ distributions obtained with the single-channel ($S_{R1ch}$, blue) and the two-channel ($S_{R2ch}$, green) methods, as well as the common background noise level distribution ($N$, black) for the office meeting, using the left, dB SPL ordinate. The SNRs for the single-channel method (SNR1ch, blue) and the two-channel method (SNR2ch, green) are provided as well, alongside the simulated single-channel SNR distribution (SNR1chSim, blue) matched at the median to SNR1ch and the corresponding simulated two-channel distribution (SNR2chSim, green), using the right, dB SNR ordinate. Finally, the corresponding simulated true SNR is shown (SNRTrue, red). Figure 7(B) shows the corresponding results for the public lunch scenario. The left- and right-hand ordinates were aligned in both panels such that the median noise level in dB SPL corresponded to 0 dB SNR.

In the office meeting scenario, the median of $S_R$ was 76.2 dB SPL for the single-channel method and 71.2 dB SPL for the two-channel method. The median of $N$ was 73.5 dB SPL. The resulting median of $S_{R1ch}$ and $S_{R2ch}$ were -2.5 dB and 2.3 dB, respectively. $S_{R1chSim}$ had a median value of -3.1 dB at a corresponding median SNR1ch of -3.4 dB. In the public lunch scenario, the median of $S_R$ was 79.5 dB SPL in the case of the single-channel method and 75.4 dB SPL for the two-channel method at a median of $N$ of 75.5 dB SPL. The median SNR1ch and SNR2ch were 4.0 dB.

### Table I. Room acoustic parameters for the two real-world scenarios.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>RT60 (s)</th>
<th>DRR (dB)</th>
<th>C50 (dB)</th>
<th>EDT (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Office meeting</td>
<td>0.4</td>
<td>6.6</td>
<td>16.9</td>
<td>0.2</td>
</tr>
<tr>
<td>Public lunch</td>
<td>3.5</td>
<td>16.6</td>
<td>23.5</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Figure 6. (Color online) Room acoustic SNR simulations for the office meeting scene (A) and the public lunch scene (B). The true SNR distributions (SNR True, right, red) around the median and the corresponding SNR distributions obtained with the single-channel (SNR1ch, middle, blue) and two-channel (SNR2ch, right, green) methods are shown.
and −0.6 dB, respectively. SNR_{2chSim} had a median value of 1.2 dB for a median SNR_{True} of 1.5 dB.

A one-way analysis-of-variance test showed that the speech level and SNR distributions were significantly higher for the single-channel method compared to the two-channel method both in the office meeting and the public lunch scenario ($p < 0.0001$ when comparing S_{1ch} to S_{2ch} and SNR_{1ch} to SNR_{2ch}). Also, in both scenarios, the SNR_{2chSim} distribution was significantly lower than the SNR_{1chSim} distribution but not significantly different from either the SNR_{2ch} or the SNR_{True} distributions.

IV. DISCUSSION

The room acoustic simulation results clearly showed that the single-channel method consistently overestimated the true SNR, measured across a range of evaluated SNRs. The two-channel method approximated the true SNR very closely. Because the $N$ signal was estimated in the same way for both methods, the difference was caused by the $S_R$ signal estimations. The single-channel method assumes that speech and noise signals are uncorrelated, which is not the case for the multi-talker babble noise signal used here and, therefore, results in an overestimation of the clean speech power. This challenge did not arise in the two-channel method as $P(S_R)$ was derived directly from the $M_{CM}$ signal. In addition, the single-channel method suffered from saturation at SNRs below −10 dB regardless of the true input SNR. This happens because at low SNRs, $P(S_R)$ becomes small compared to the underlying $P(N)$ such that the SNR distribution essentially reflects the distribution of $P(N)$ during target speech relative to $P(N)$ during speech pauses (Kim and Stern, 2008). The two-channel method’s use of the $M_{CM}$ avoids such saturation. Last, while the implementation of the single-channel method in the present study avoided practical target-speech-segment labeling issues by reusing the two-channel method’s VADs, the hand-labeled data in the reference studies may have been affected by the resulting underrepresentation of low speech levels in the SNR distributions.

Since the simulated two-channel method only differs conceptually from the true SNR in its approximation of $P(N)$ by $P(N)$, its slightly differing estimates occurred because the distribution of $N$ during target speech was not identical to that of $N$ during speech pauses. This was more evident in the public lunch scenario than in the office meeting as the higher DRR and $C_{50}$ values in the public lunch environment reflected a more fluctuating $N$. Nevertheless, the two-channel method approximated the true SNR far more closely than did the single-channel method.

With regard to the real-world measurements, the potential effect of the target speech presence on the noise level, as well as the likely violation of the assumptions of anechoic, noise-free target speech and the isotropic receiver noise, need to be considered. The measured speech, noise, and SNR distributions in the two real-world scenes indicated that although the absolute $S_R$ and $N$ levels, as well as the SNRs, were higher for the public lunch scenario than for the office meeting scenario, the two-channel method provided about 4 dB lower median $S_R$ levels and SNRs compared to the single-channel method in both scenes. These differences were roughly consistent with the corresponding differences between the matched simulated single-channel SNR distributions and their two-channel counterparts even though the widths of the measured two-channel
SNR distributions were narrower than the simulated ones. This reduction in width is due to the more narrow distribution of the real-world recorded speech signal compared to the simulated speech signal. The two-channel method estimated the median of $\text{SNR}_{\text{true}}$ in the office meeting scenario slightly more accurately than in the public lunch. This is likely due to the lower DRR and $C_{50}$ values in the office meeting scenario, indicating a more isotropic and stationary noise field compared to the public lunch, in line with the assumptions pertaining to the $N$ signal. Nevertheless, the two-channel measured SNR distribution’s interquartile range was lower than that of the simulated SNR distribution for both scenarios.

The estimated median SNRs of the two-channel method of $-2.5$ dB and $-0.5$ dB are in line with the SNRs obtained in other realistic scenarios (Culling, 2016) and consistent with the notion that conversational SNRs decrease with increasing talker distance (Weisser and Buchholz, 2019). The width of the $S_N$ level distributions was found to be smaller in the office meeting than in the public lunch scenario for both methods. One explanation for this is that talkers maintained a reasonably constant talking level at a larger fixed distance—where communication is more difficult—compared to when they are close together. This, in turn, affects the widths of the corresponding SNR distributions as well. The distributions for the background noise level were found to be rather symmetric in both scenarios and did not differ between the estimation methods because the noise contribution was calculated in exactly the same way.

Although the two-channel method most likely characterizes conversational SNRs more accurately than the single-channel approach, it has several limitations. The necessity of the cheek microphone signal implies that existing single-channel recordings cannot be reanalyzed such that additional measurements are needed to acquire SNR distributions in scenes other than the two described here. The fact that the RIR needs to be recorded and calibrated at a predefined distance implies that the method is tailored to the fixed talker distance in a specific target-receiver configuration in the scene. Additionally, the FFC applied to the cheek microphone signal was only measured from the front and, thus, did not account for potential head movements of the target talker. The two-channel method implements one specific way of estimating the acoustic path between the target and receiver, aiming to more accurately approximate the true SNR.

Nonetheless, the proposed SNR estimation method captures real-world SNR distributions with an increased degree of accuracy compared to the single-channel approach while also allowing for the dynamical tracking of speech levels and SNRs in real-world scenarios. It can be applied in real-world scenes for both offline data collection, as implemented here, and real-time tracking. This enables applications beyond broadband level estimation, including precise frequency-specific target speech analysis and the accurate temporal characterization of speech rates, turn-taking, and conversational behavior in a realistic way.

V. CONCLUSION

A two-channel method for the SNR estimation of a target talker in conversation was developed based on a room acoustical approximation to the true SNR. With the proper calibration and setup, the method was shown to result in significantly reduced speech levels and downward-shifted SNR distributions compared to a common single-channel reference method. Median values for the two-channel method were more than $4$ dB lower than for the single-channel method, likely due to an overestimation of the level of a noise-correlated speech signal in the single-channel method. As such, the proposed method might provide interesting perspectives on how conversational real-world SNRs can be estimated.

ACKNOWLEDGMENTS

The research was supported by the Centre for Applied Hearing Research (CAHR) and Widex A/S.