Downloaded from orbit.dtu.dk on: Apr 26, 2024

DTU Library

Speech intelligibility in a realistic virtual sound environment

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

Published in:
Journal of the Acoustical Society of America

Link to article, DOI:
10.1121/10.0004779

Publication date:
2021

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. (2021). Speech intelligibility in a realistic virtual sound environment. *Journal of the Acoustical Society of America*, *149*(4), 2791-2801. https://doi.org/10.1121/10.0004779

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.

Speech intelligibility in a realistic virtual sound environment

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

Citation: The Journal of the Acoustical Society of America **149**, 2791 (2021); doi: 10.1121/10.0004779
View online: https://doi.org/10.1121/10.0004779
View Table of Contents: https://asa.scitation.org/toc/jas/149/4
Published by the Acoustical Society of America

ARTICLES YOU MAY BE INTERESTED IN

A method for realistic, conversational signal-to-noise ratio estimation
The Journal of the Acoustical Society of America **149**, 1559 (2021); https://doi.org/10.1121/10.0003626

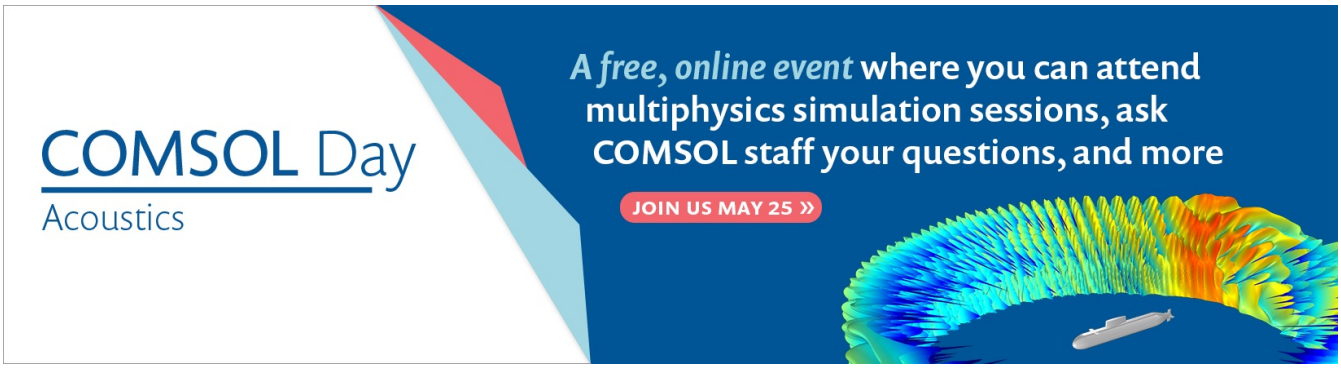
State-space modeling of sound source directivity: An experimental study of the violin and the clarinet
The Journal of the Acoustical Society of America **149**, 2768 (2021); https://doi.org/10.1121/10.0004241

Speech recognition as a function of the number of channels for an array with large inter-electrode distances
The Journal of the Acoustical Society of America **149**, 2752 (2021); https://doi.org/10.1121/10.0004244

Access to semantic cues does not lead to perceptual restoration of interrupted speech in cochlear-implant users
The Journal of the Acoustical Society of America **149**, 1488 (2021); https://doi.org/10.1121/10.0003573

Machine learning in acoustics: Theory and applications
The Journal of the Acoustical Society of America **146**, 3590 (2019); https://doi.org/10.1121/1.5133944

Portable Automated Rapid Testing (PART) for auditory assessment: Validation in a young adult normal-hearing population
The Journal of the Acoustical Society of America **148**, 1831 (2020); https://doi.org/10.1121/10.0002108



JASA ARTICLE

Speech intelligibility in a realistic virtual sound environment

Naim Mansour,[1,a)] Marton Marschall,[1,b)] Tobias May,[1] Adam Westermann,[2] and Torsten Dau[1,c)]
[1]*Hearing Systems Section, Department of Health Technology, Technical University of Denmark, Kongens Lyngby, Denmark*
[2]*Widex A/S, Lynge, Denmark*

ABSTRACT:
In the present study, speech intelligibility was evaluated in realistic, controlled conditions. "Critical sound scenarios" were defined as acoustic scenes that hearing aid users considered important, difficult, and common through ecological momentary assessment. These sound scenarios were acquired in the real world using a spherical microphone array and reproduced inside a loudspeaker-based virtual sound environment (VSE) using Ambisonics. Speech reception thresholds (SRT) were measured for normal-hearing (NH) and hearing-impaired (HI) listeners, using sentences from the Danish hearing in noise test, spatially embedded in the acoustic background of an office meeting sound scenario. In addition, speech recognition scores (SRS) were obtained at a fixed signal-to-noise ratio (SNR) of −2.5 dB, corresponding to the median conversational SNR in the office meeting. SRTs measured in the realistic VSE-reproduced background were significantly higher for NH and HI listeners than those obtained with artificial noise presented over headphones, presumably due to an increased amount of modulation masking and a larger cognitive effort required to separate the target speech from the intelligible interferers in the realistic background. SRSs obtained at the fixed SNR in the realistic background could be used to relate the listeners' SI to the potential challenges they experience in the real world.
© *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/10.1121/10.0004779
(Received 16 September 2020; revised 25 February 2021; accepted 28 March 2021; published online 22 April 2021)
[Editor: Jonas Braasch] Pages: 2791–2801

I. INTRODUCTION

Through their auditory perception, normal-hearing people are able to communicate nearly effortlessly even in challenging acoustic scenarios, such as at a social gathering or in a busy restaurant. In contrast, a person whose hearing is impaired often experiences diminished speech communication ability, hindering many social interactions (Moore, 1996). Hearing aids (HA) aim at compensating for hearing deficits by employing frequency- and level-dependent amplification to restore the wearer's sensitivity to soft sounds, compensate for loudness recruitment and increase overall sound quality. However, despite considerable technological advances in hearing aid technology, hearing aid benefit varies greatly among individual users, particularly in reverberant situations with multiple interfering sound sources.

To appreciate why this occurs, it is necessary to understand how human hearing is currently evaluated in the context of hearing aid applications. Typically considered paradigms, such as loudness perception or speech intelligibility, utilize well-defined, artificially created acoustic stimuli presented over headphones or small sets of loudspeakers. This approach, while having the advantage of being fully controlled and replicable, does not necessarily reflect conditions in the real world.

With respect to the acoustic stimuli employed in speech intelligibility (SI) paradigms, the often used speech-shaped stationary noise (SSN) maskers differ considerably from actual multi-talker acoustic interferers. SSN lacks the typical low-frequency modulations of multi-talker interferers and therefore does not provide the listener with the opportunity to utilize speech "glimpses" in the interferer (Dreschler *et al.*, 2001). The hearing in noise test (HINT), that has been widely used to evaluate speech intelligibility across many languages and in various acoustic conditions, typically uses SSN as its masker, or modulated noise lacking intelligible interferers. This test, along with other approaches like matrix-based sentence tests (Houben *et al.*, 2014; Kelly *et al.*, 2017; Wagener *et al.*, 2003), results in normal-hearing (NH) speech reception thresholds (SRT)—corresponding to 50% speech intelligibility—that are well below a signal-to-noise ratio (SNR) of 0 dB (Nielsen and Dau, 2011; Soli and Wong, 2008; Wagener *et al.*, 2003). In contrast, research trying to categorize real-world SNRs has consistently found a substantially higher range of values in the majority of sound scenarios (Smeds *et al.*, 2015).

The presentation of the stimuli in an SI task is equally problematic in terms of realism. Headphones can "spatialize" the presented sounds using head-related-transfer functions (Wightman and Kistler, 1989), but this approach is limited in accuracy by many factors, including headphone placement and the limited spatial resolution imposed by the angle between the discrete functions. In addition, accounting for head movements and fitting a HA with headphones is

[a)]Electronic mail: naiman@dtu.dk, ORCID: 0000-0001-5673-6840.
[b)]ORCID: 0000-0003-2534-7062.
[c)]ORCID: 0000-0001-8110-4343.

cumbersome in practice. Quadraphonic loudspeaker setups, often used for the spatial evaluation of HA algorithms, physically separate the noise maskers to alleviate these problems, but they generally still do not faithfully reproduce spatially diffuse noise (ITU-T, 2018). In short, an SI task presenting artificial stimuli in a simplified spatial manner might misrepresent the difficulties that both NH and HI people experience when listening to speech in noise in their daily lives. Due to these discrepancies, it has remained unclear how SI performance scores in laboratory settings relate to these real-world difficulties (Culling, 2016).

To more precisely tailor the performance of a hearing aid to the needs of a user, it would be advantageous to utilize an SI testing paradigm that mimics conditions in the real world to the highest possible degree, thereby making it more ecologically valid (Reis and Judd, 2000). One option could be to actually conduct the SI task in the real world (e.g., through field tests). While perfectly realistic, real-world acoustic conditions are highly variable, resulting in outcome measures that would be difficult to interpret and reproduce. Attempting to bring the real world into the lab represents a trade-off between control and realism, both regarding stimulus choice and acoustic presentation. A proper balance of "controlled realism" would have the potential to result in consistent, yet ecologically valid, findings (Best *et al.*, 2015).

A virtual sound environment (VSE) in the form of a spherical loudspeaker array is able to render complex three-dimensional sound fields at its center through higher-order ambisonic (HOA) reproduction techniques (Bertet *et al.*, 2006). Using such an array to present target speech sentences superimposed onto spatial recordings of realistic sound scenarios would ensure the reproduction of acoustic sound field properties within the limitations of the recording setup, allowing for head movements and providing a sense of spatial immersion. VSEs have been used extensively in combination with simulated spatialized maskers based on room-acoustic simulations [e.g., ODEON (2018)] to study aspects of auditory spatial separation (Best *et al.*, 2017a), informational masking (Westermann and Buchholz, 2015), hearing aid performance (Cubick and Dau, 2016; Minnaar *et al.*, 2010), and speech intelligibility (Ahrens *et al.*, 2017; Best *et al.*, 2015; Westermann and Buchholz, 2017). However, this approach is still limited by the number and complexity of sources that can be simulated and has been shown to correlate only poorly with real-world conditions [e.g., Ahrens *et al.* (2019)]. While real-world HOA recordings have become increasingly available [e.g., Weisser *et al.* (2019)], there exists, to the best of the authors' knowledge, no research that utilizes such spatially recorded maskers in VSE-based SI tasks. In addition, for these recorded sound scenarios (e.g., an office meeting or a restaurant visit) to become more ecological valid, they should be selected based on scenarios that users consider critical in their lives and captured in a real-world environment.

In the present study, a speech intelligibility task is presented that aims to increase ecological validity. A set of critical sound scenarios was selected based on a categorization of HA user ecological momentary assessment (EMA) data (Smeds *et al.*, 2018). Out of these scenarios, an office meeting scenario was recorded *in situ* with a spherical microphone array. This recording was subsequently reproduced as a VSE using Ambisonics over a 64-channel, fully spherical loudspeaker array inside an anechoic enclosure. Finally, the reproduced masker was combined with the spatialized speech corpus of the Danish HINT as part of a speech intelligibility task carried out by NH and HI listeners. Adaptive SRTs were captured as well as speech recognition scores (SRS) at a constant SNR of −2.5 dB, corresponding to the median conversational SNR between NH people in the office meeting scenario (Mansour *et al.*, 2019). The hypotheses were that (i) a speech intelligibility paradigm employing realistic, spatialized stimuli would produce higher SRTs compared to those obtained with an artificial approach and (ii) SRSs at a real-world conversational SNR would reflect some of the difficulties HI people experience in the real world.

II. METHODS

A. Sound scenario selection

To increase the relevance and ecological validity of the recorded critical sound scenarios as potential maskers in the speech intelligibility task, the scenario selection was based on EMA data from 281 field reports by HA users, collected by Smeds *et al.* (2018). In EMA, user data are captured in real time by subjects in everyday scenarios. The use of EMA data obtained in this way has become increasingly popular in attempts to describe and characterize which scenes HA users experience (Timmer *et al.*, 2017).

Figure 1 shows the critical sound scenario framework that was developed in this study and which categorized a real-world scenario based on the binary combination of three EMA questionnaire metrics: the reported importance, difficulty, and occurrence of understanding speech in that scenario. Combinations of these three parameters have been

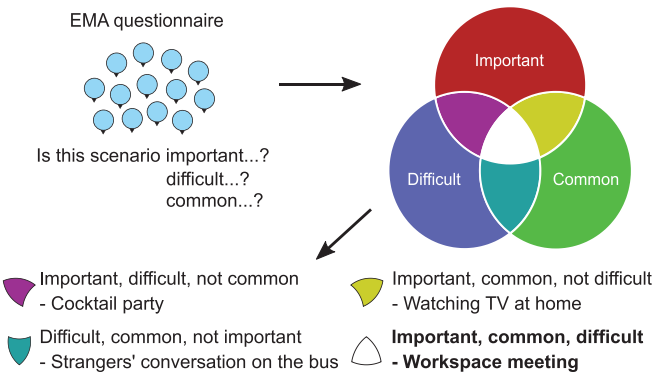


FIG. 1. (Color online) The critical sound scenario framework, developed to categorize HA users' ecological momentary assessment (EMA) field report data based on a binary combination of metrics of reported importance, difficulty, and occurrence of understanding speech in everyday scenarios. Examples of scenarios at the different intersections are given, with the area considered important, common, and difficult displayed in bold.

https://doi.org/10.1121/10.0004779

shown to accurately separate different real-life situations (Wolters *et al.*, 2016). The examples provided in Fig. 1 illustrate different combinations of these metrics, graphically depicted as a Venn diagram. For instance, while watching TV at home is a common and important scenario for HA users, it is generally not considered as difficult. Understanding speech in cocktail party scenarios is important and difficult even for normal-hearing people, but not very common for many HA users. Workspace meetings and having lunch in public are considered important as well as difficult and common. For the scenarios rated as important, the percentages displayed in the four sections of the corresponding circle in the Venn diagram denote their relative occurrences.

In the present study, the subset of scenarios that were simultaneously rated important, difficult and common were chosen for further analysis as they represent conditions in which HA users are challenged the most. This subset was cross-referenced with the EMA reports to reveal the three most prevalent critical sound scenarios: a public lunch, a small festive event (e.g., a family house party), and a workspace meeting. From these, the workspace meeting scenario was selected for further processing.

B. Sound scenario acquisition

Figure 2 illustrates the recording setup used to acquire the office meeting scenario. The scenario was captured during a staged office meeting with a spherical microphone array (em32 Eigenmike, mh acoustics LLC, USA) capable of 4th HOA recording (Bertet *et al.*, 2006) with its spatial aliasing frequency at 9 kHz. In addition, a Knowles Electronic Manikin for Acoustic Research (KEMAR, 2018) with ear canals was used to capture binaural signals. While 12 participants conversed in pairs, seated at and standing around a conference table in the office meeting room [Fig. 2(a)], spatial scene recordings were obtained with the EigenMike and KEMAR in the listener position at the bottom center of the table. Room impulse responses (IR) were captured from the target position (top center of the table) using a mounted loudspeaker producing a series of three repeated 15-s exponential sweeps (Müller and Massarani, 2001) between 20 Hz and 20 kHz, while all participants remained still and quiet to avoid altering the room reverberation [Fig. 2(b)].

To obtain conversational signal-to-noise ratios resulting from two interacting participants seated in the listener position and the target position, respectively [Fig. 2(c)], the method detailed in Mansour *et al.* (2019) was used. In this method, speech produced by the target speaker was recorded via a DPA 4066 cheek microphone. The recorded cheek microphone signal was free-field corrected and convolved with the captured impulse response to obtain an estimate of the target speech at the listener. The free-field correction was obtained as the transfer function between white noise recorded by the cheek microphone mounted on the KEMAR in an anechoic chamber and a reference microphone at 0.5 m distance. The background noise was measured with the right ear of the KEMAR during gaps where neither the target nor the listener was speaking, and the resulting SNR was calculated as the ratio of the speech energy at the receiver and the background noise energy. Energy-based broadband voice activity detection (VAD) (Kinnunen and Li, 2010) was used to separate the target and listener speech from the background. The median value of the SNRs obtained in this manner, −2.5 dB, was used in the constant-SNR SI assessment.

C. Sound scenario reproduction

The spherical microphone recordings made with the Eigenmike were encoded from the 32 raw input channels to a 25-channel Ambisonic fourth-order HOA format. These

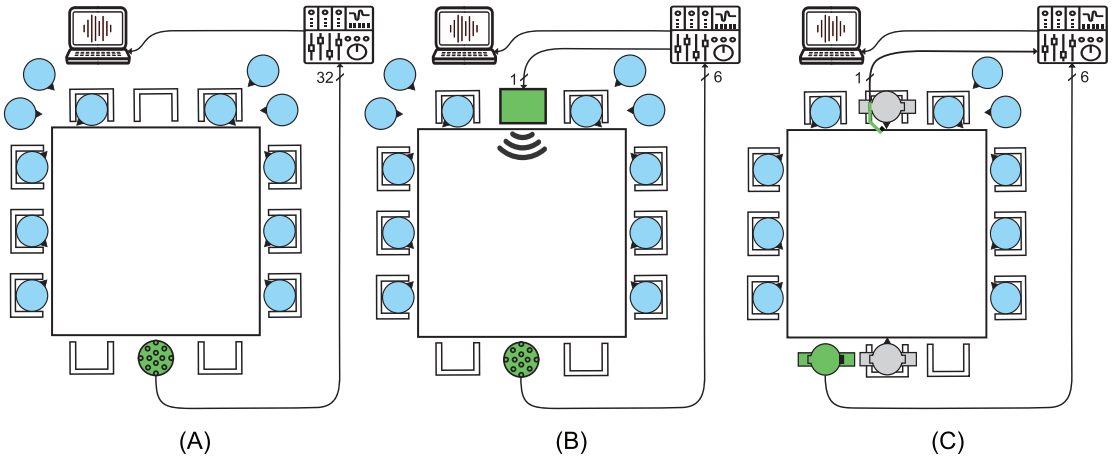


FIG. 2. (Color online) Illustrations of the office meeting scenario acquisition stages, consisting of the background scene recording (A), the room impulse response recording stage (B), and the conversational SNR estimation stage (C). Each panel contains the human participants (blue, round heads) distributed around a large, square conference table, as well as the sound card and recording laptop. The spherical microphone array is depicted in the listener position at the bottom of (A) and (B) (green, dotted circle). Panel (B) additionally includes the loudspeaker in the target position at the top (green, rectangle). The target talker, wearing the cheek microphone (green, line), and the listener are depicted in (C) in the target and listener positions, respectively (gray, heads with shoulders). The KEMAR is positioned to the left of the listener (green, mannequin).

HOA signals were then rendered on the 64-channel spherical loudspeaker array in the AudioVisual Immersion Lab (AVIL) at DTU [Fig. 3(a)] using dual-band (basic, max-rE) decoding, with a crossover frequency of 2400 Hz. HOA auralization was chosen because of its physically faithful rendering of sound fields in the sweet spot at the center of the array (subject to the limitations imposed by the spatial aliasing frequency of the microphone array), ensuring their usability for HAs as well as human ears.

The level of the reproduced masker was required to not fluctuate strongly during its playback, to avoid largely varying SRTs in the SI task. To this end, specific subsections of the raw Eigenmike recording were extracted and concatenated before Ambisonic rendering based on level estimates derived from its front-facing microphone. The 10-min-long recording was segmented into frames of 5 s (with 80% overlap) and level differences (in dB) were calculated between consecutive frames. The upper and lower boundaries for allowed level differences were set to the 5th and 95th percentile of the level difference distribution, respectively. The collection of consecutive frame segments within these boundaries was retained and concatenated in decreasing duration (from 25 s to about 7 s), resulting in a level-equalized recording of 2.5 min. The full 32-channel synchronized version of this reduced recording was rendered to the 64-channel reproduction, calibrated segment by segment to a fixed target sound pressure level (SPL) of 73.5 dB using a B&K 2669 free-field microphone, and cross-faded with a 1-s Hann-windowed overlap for smooth transitions between segments. The target SPL was selected as the median value of the noise level distribution measured during the conversational SNR estimation stage. Finally, the resulting 2-min-long background reproduction was calibrated binaurally inside the loudspeaker array, using a B&K type 4128 Head and Torso Simulator (HATS) with ear canals. This approach ensured that the background reproduction retained its intelligible properties despite having been dynamically stabilized in level.

D. Speech stimuli and interferers

To evaluate speech intelligibility, the Danish HINT (Nielsen and Dau, 2011) was used. The HINT uses brief, mundane, male-voiced, 5-word target sentences presented in speech-shaped stationary noise (SSN) to estimate SRTs using an adaptive, 1-up-1-down, sentence-based scoring procedure. During each trial, consisting of a sequence of 20 sentences, the procedure decreased the SNR of a sentence by 2 dB if the previous sentence was repeated back entirely correctly and increased it by 2 dB otherwise. The initial SNR was set to 0 dB and the first sentence was replayed at increasing SNRs until it was repeated back correctly, before continuing. The SRT of a trial was then calculated as the average of the SNRs across the last 15 sentences. In addition, a non-adaptive procedure, which presented the sequence of 20 sentences at a constant SNR, was implemented to estimate speech reception scores in % correct, testing the second hypothesis that SRSs at real-world SNRs reflect difficulties with speech intelligibility. Three conditions were evaluated for each procedure.

- (HP) The classical HINT reference condition, where anechoic target sentences were presented diotically in SSN over headphones, served as the control condition. In the following, this condition is referred to as the "headphone condition" (HP).
- (RE) The primary spatial condition used spatialized HINT target sentences that were integrated into the reproduced office meeting interferer and presented through the loudspeaker array. The spatialized sentences were obtained by convolving the 60 training and 200 test HINT sentences individually with the spherically recorded IR that was captured between the target and listener positions. Each sentence was calibrated individually to retain the same unique level as that of the single-channel version, measured in the sweet spot of the loudspeaker array. This condition is referred to in the following as the "realistic noise" condition (RE).

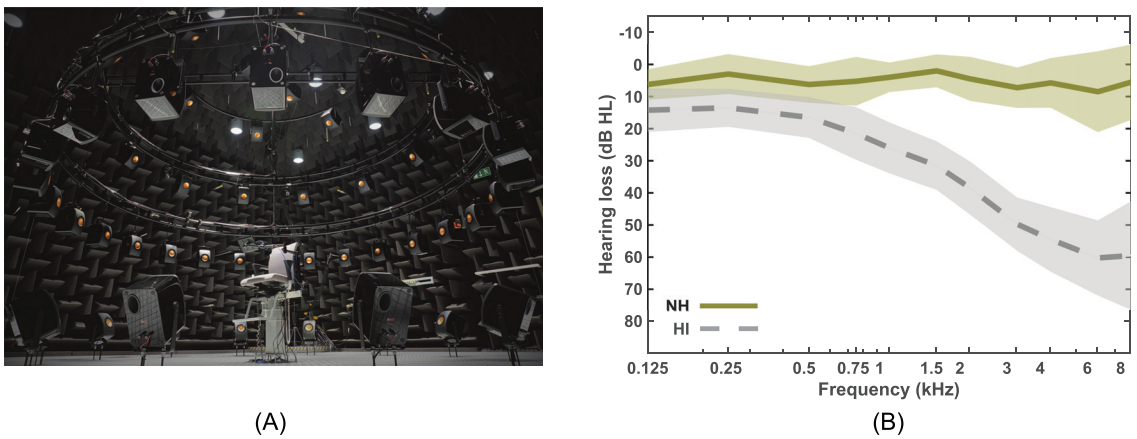


FIG. 3. (Color online) The AudioVisual Immersion Lab (AVIL), serving as the reproduction laboratory, containing a spherical loudspeaker array in an anechoic enclosure (A) and the mean audiograms of the normal-hearing (NH, gold, solid) and the hearing-impaired (HI, silver, dashed) listener groups as well as their standard deviations (B).

https://doi.org/10.1121/10.0004779

- (AR) The secondary spatial condition, with similarly spatialized HINT sentences presented in a decorrelated quadraphonic version of the HINT SSN playing from four loudspeakers in the array at 45°, 135°, 225°, and 315° azimuth, 0° elevation. This "artificial noise" condition (AR) was included primarily to investigate the effect of changing only the type of spatialized background noise on speech intelligibility.

In each condition, speech-to-noise SNRs were established by varying speech sentence levels with respect to the continuously playing, looped background. The fixed SNR used in the non-adaptive procedure was set to the median SNR of −2.5 dB obtained from the conversational assessment (Sec. II B), considered as a representation of a realistic NH conversational SNR.

E. Listeners

Ten NH and ten HI listeners participated in the experiment. The NH listeners were between 21 and 69 years old with a median age of 28, while the HI listeners were between 56 and 75 years old with a median age of 70. The NH listeners had a four-frequency average hearing loss (HL) of maximally 15 dB, while the HI listeners had an average sloping mild (N2) to moderate (N3) hearing loss (Bisgaard *et al.*, 2010). Figure 3(b) shows the individual audiograms as well as their mean (with shaded standard deviations) for the NH (gold, solid) and HI (silver, dashed) listener groups. All HI listeners had a word discrimination score in quiet of at least 92% for both ears, and a left-right-ear HL difference of maximally 10 dB for all frequencies. All listeners provided informed consent and all experiments were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391).

F. Speech intelligibility procedure

Three adaptive training rounds were carried out for target speech in quiet (to ascertain audibility), as well as for the target speech presented in the AR and RE conditions. Then, two evaluation rounds were conducted for all adaptive conditions, and one for all constant-SNR conditions. One HINT round contains a sequence of 20 predetermined sentences, presented in random order. The testing order was randomized over condition (HP-RE-AR) and test list number (1–9) through the use of two 9×9 Latin squares (Bradley, 1958) with two random completions. Within one condition, two adaptive test lists were always followed by one at the constant SNR. The two corresponding SRTs were averaged to obtain a final SRT, and an SRS was established as the percentage of correctly understood words in the constant-SNR list. The SNRs in the adaptive procedure were adjusted based on sentence scoring, where a 5-word target sentence is marked correct only when all 5 words were repeated accurately by the listener. This is the standard way of scoring the Danish HINT (Nielsen and Dau, 2011). The SRSs in the constant-SNR procedure were based on word scoring, where the number of correctly repeated words in every sentence is counted, summed over all sentences in a list and divided by 100. The decision to use word scoring for the constant-SNR procedure was taken to increase the scoring sensitivity of the SI task, thereby avoiding flooring effects, at an SNR where HI listeners were expected to struggle considerably.

All listeners provided informed consent and all experiments were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391). The experiment lasted, on average, one hour and the HINT scoring was carried out by a native Danish speaking audiologist.

G. Questionnaire and statistical analysis

In addition to the objective SI assessment, all listeners were asked to fill out a questionnaire after completion of the experiment. Table I displays the questions that were asked, pertaining to the realism of the sound of the stimuli and the difficulty in understanding speech. The response scale was a 5-point Likert scale, asking the respondent to rate a percept from not at all present (1) to extremely present (5).

To check the data obtained in the different conditions (HP_{NH}, HP_{HI}, RE_{NH}, RE_{HI}, AR_{NH}, and AR_{HI}), a two-way mixed-effects analysis of variance (MANOVA) statistical test was used where the condition HP/RE/AR represented a within-listener factor and the hearing status NH/HI represented a between-listener factor. The normality of each group was verified with the Anderson-Darling and Shapiro-Wilk tests and the similarity in variance between the compared groups required for the MANOVA was evaluated with a Bartlett test. A one-way analysis-of-variance (ANOVA) test was applied to investigate specific paired comparisons between NH and HI listeners, as well as a repeated-measurement ANOVA (RANOVA) to compare between HP/RE/AR conditions within a listener group. In all tests, the significance level was set at 5%.

III. RESULTS

A. Acoustic properties of the stimuli

Figure 4 shows the long-term average spectra (LTAS) (left panel) and the modulation power spectra (right panel) for the speech-shaped noise used by the HINT (dotted), the quadraphonic SSN (dot-striped), the office meeting noise (solid), and a concatenation of all two hundred HINT test sentences, monophonic and spatialized with the office meeting room impulse response (dashed). The LTAS functions were normalized to have the same broadband RMS. The modulation power spectra were obtained by normalizing the calculated power within a modulation band by its respective bandwidth as well as the power in the DC component (Dreschler *et al.*, 2001). The quadraphonic noise, the office meeting noise, and the auralized HINT sentences were recorded binaurally with the HATS inside the loudspeaker array (left-ear spectra shown here).

The LTAS of the HINT SSN (left panel) represents its speech-shaped spectral character. The averaged monophonic

TABLE I. Content of the questionnaire given to all listeners after completing the experiment, as well as the 5-point Likert response scale. The frequency of responses of the normal-hearing (NH) and hearing-impaired (HI) listeners to each possible response are displayed in the two rightmost columns. The most often occurring response in each group is highlighted in bold.

Questions asked to the normal-hearing (NH) and hearing-impaired (HI) listeners	NH					HI				
Response: Not at all (1) - Not that (2) - Somewhat (3) - Very (4) - Extremely (5)	1	2	3	4	5	1	2	3	4	5
How realistic did the office background noise in the experiment sound to you?	0	1	3	**4**	2	0	0	3	**6**	1
How difficult was it to understand the speech in the artificial background noise in the experiment?	1	1	**5**	2	0	0	2	**6**	2	1
How difficult was it to understand the speech in the office background noise in the experiment?	0	1	3	**5**	1	0	0	2	**7**	1
How realistic did the speech that you had to listen to in the experiment sound?	0	1	3	**5**	1	0	0	2	**7**	1

HINT sentence LTAS is identical to that of the HINT SSN, since this noise was originally constructed from the averaged power spectrum of one hundred HINT sentences. For the quadraphonic SSN, the LTAS is increased by up to 10 dB relative to the monophonic version at frequencies above 1100 Hz. The LTAS of the spatialized HINT sentences reflects the effect of the room on the speech-shaped stimuli. Similarly, the LTAS of the office meeting noise shows its speech-like nature, smoothed and altered by the room reverberation.

For the modulation spectra (right panel of Fig. 4), the quasi-stationary nature of the HINT SSN is reflected by its low energy across the displayed modulation frequencies. The modulation power remains low and roughly constant at all considered frequencies. The quadraphonic SSN exhibits the same modulation spectrum as the classical version. The monophonic HINT sentences contain a large amount of modulation power, typical for signals with speech-like envelope fluctuations. Spatializing these signals hardly alters these contributions. The modulation spectrum of the office meeting noise lies in between these two extreme patterns, since it contains various concurrent talkers speaking in a reverberant environment. As such, the HINT SSN reflects the averaged spectral characteristics of the speech stimuli but not their low-frequency modulations, while the office meeting noise is speech-like in both the spectral and the modulation spectral domains.

B. Speech reception thresholds

Figure 5(a) shows individual SRT results for the adaptive HINT procedure (blue diamonds), as well as their mean (red circle) and standard deviations (red squares), median and 25th/75th percentiles (box plot) for the NH and HI listeners in the three conditions HP (left), AR (middle), and RE (right). Each individual SRT result represents the average of two SRT measurements. The mean HP SRTs were obtained at -3.3 dB for the NH listeners as opposed to -1.5 dB for the HI. The mean SRTs in the RE condition were -1.2 dB for the NH listeners and 1.5 dB for the HI listeners. For the AR condition, the mean SRT was -1.1 dB for the NH listeners, compared to -0.3 dB for the HI listeners.

Comparing the HP to the RE condition, a significant effect of condition $[F(1,18) = 41.57, p \leq 0.0001]$ and

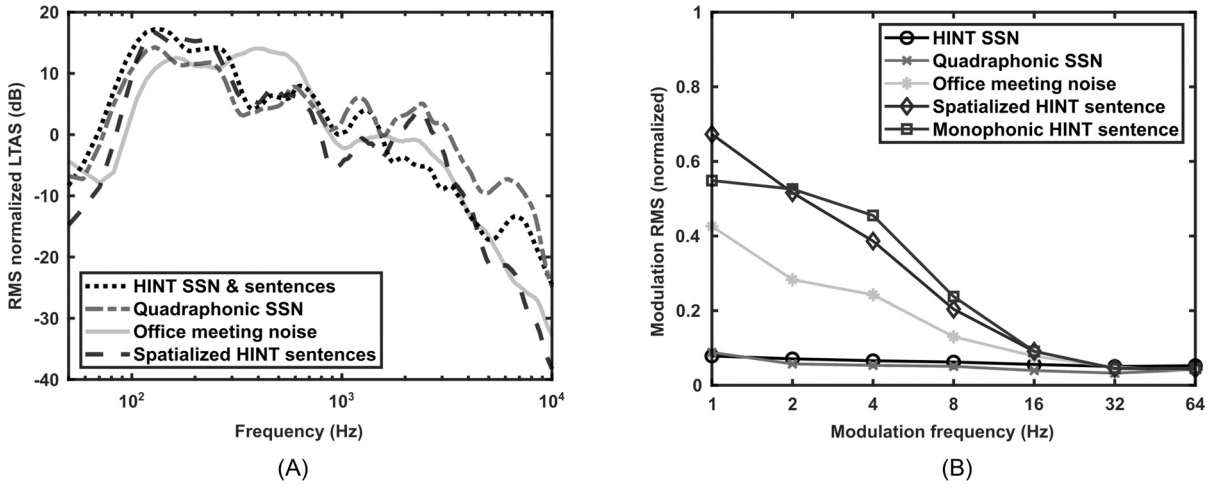


FIG. 4. Root-mean square (RMS) normalized long-term average spectra (A) and RMS normalized modulation frequency spectra (B) of the stimuli used in the speech intelligibility task, specifically the HINT monophonic target sentences and speech-shaped noise (SSN), the quadraphonic SSN, the office meeting noise, and the spatialized HINT target sentences.

https://doi.org/10.1121/10.0004779

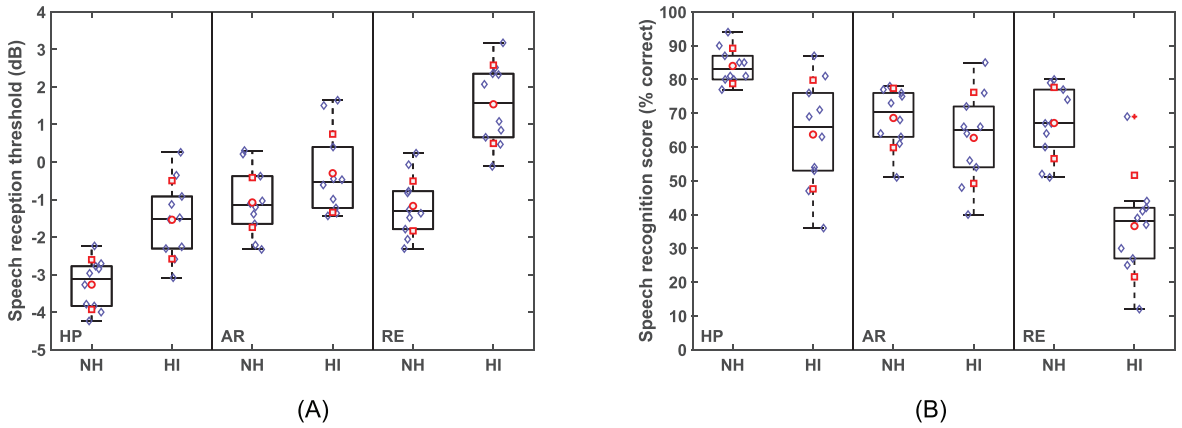


FIG. 5. (Color online) Speech intelligibility results for the adaptive procedure (A) and the percentage correct procedure at the conversational SNR (B). Results are shown for the headphone condition (HP, left), the artificial speech-shaped noise condition (AR, middle) and the office meeting noise condition (RE, right), for normal-hearing (NH) and hearing-impaired (HI) listeners. The box plots show median values and inter-quartile ranges. Individual data points are shown (blue, diamonds) as well as the mean (red circle) and standard deviations (red, squares).

hearing status $[F(1,18) = 205.13, p \leq 0.0001]$ was found, but no significant interaction between the two $[F(1,18) = 2.86, p = 0.1049]$. Similarly, comparing the results obtained for the HP and AR conditions, significant effects for condition $[F(1,18) = 40.26, p \leq 0.0001]$ and hearing status $[F(1,18) = 14.56, p = 0.0013]$ were observed, but no interaction effect $[F(1,18) = 3.08, p = 0.0962]$. Finally, a comparison of the results obtained in the AR and RE conditions revealed significant effects of condition $[F(1,18) = 8.80, p = 0.0083]$ and hearing status $[F(1,18) = 27.57, p \leq 0.0001]$, as well as a significant interaction $[F(1,18) = 10.78, p = 0.0041]$.

Pair-wise comparisons showed that the SRTs for the HI listeners were significantly higher than for the NH listeners in the HP ($p = 0.0003$) and RE ($p \leq 0.0001$) conditions, but not in the AR condition ($p = 0.1056$). Within the NH listeners, SRTs were significantly higher in the RE ($p \leq 0.0001$) and AR ($p \leq 0.0001$) conditions compared to the HP condition, but similar between the RE and AR conditions ($p = 0.801$). For the HI listeners, SRTs were significantly increased in the RE condition compared to the AR ($p = 0.0034$) condition. Last, SRTs in the AR condition were significantly higher than those in the HP condition ($p = 0.034$).

These results demonstrate that speech intelligibility decreased (i.e., SRTs increased) in the RE condition compared to the HP condition in both listener groups, whereas the performance in the RE condition (compared to both the AR and HP conditions) was much more affected in the HI listener group than in the NH listener group.

C. Speech reception scores at the normal-hearing conversational SNR

Figure 5(b) shows the speech reception scores obtained with the constant-SNR HINT procedure, summarizing the word scores of all listeners for the percentage correct evaluation at −2.5 dB SNR, the median SNR in the office meeting recording. The mean HP results corresponded to 84% correct for the NH listeners and 63.7% correct for the HI. The mean RE results corresponded to 67.1% correct for the NH and 36.6% correct for the HI listeners. For the AR condition, mean scores of 68.6% and 62.7% were obtained for the NH and HI listeners, respectively.

A significant effect of condition $[F(1,18) = 47.58, p \leq 0.0001]$ and hearing status $[F(1,18) = 30.72, p \leq 0.0001]$ was found with respect to the HP and the RE conditions, but no significant interaction $[F(1,18) = 2.56, p = 0.1272]$. Significant effects for condition $[F(1,18) = 7.94, p = 0.0114]$ and hearing status $[F(1,18) = 9.13, p = 0.0073]$ as well as a significant interaction effect $[F(1,18) = 6.12, p = 0.0235]$ were found between the HP and AR conditions. Comparing the AR to the RE condition revealed significant effects of condition $[F(1,18) = 14.52, p = 0.0013]$ and hearing status $[F(1,18) = 19.85, p = 0.0003]$, again with a significant interaction $[F(1,18) = 11.53, p = 0.0032]$.[1]

Paired comparisons showed that percentage correct scores for the HI listeners were significantly lower than for the NH listeners in the HP ($p = 0.0013$) and RE ($p \leq 0.0001$) conditions, but not in the AR condition ($p = 0.2614$). Within the NH listeners, percent correct scores were significantly higher in the RE ($p \leq 0.0001$) and AR ($p = 0.0008$) conditions compared to the HP condition, but similar in the RE and AR conditions ($p = 0.7210$). For the HI listeners, percentage correct scores were significantly decreased in the RE condition compared to the HP ($p = 0.0003$) and AR ($p = 0.0018$) conditions, but were not significantly different between HP and AR conditions ($p = 0.8429$).

The SI scores showed the same trend as in the case of the adaptive SRT estimation procedure: (1) Speech intelligibility at −2.5 dB SNR was consistently poorer in the RE condition than in the HP condition for all listeners; (2) when comparing the RE to the AR condition, the HI listeners showed substantially poorer performance than the NH listeners. Overall, the spread of percentage correct responses for NH and HI listeners across conditions showed that

neither ceiling nor flooring effects occurred, and that the RE condition resulted in the greatest separation between NH and HI performance.

D. Questionnaire results

Table I displays the results of the questionnaire given to all listeners. For both the NH and HI listeners, answers were accumulated per question and per response to produce the number ranges in the rightmost two columns. The highest frequency response within each response group is highlighted in bold. The results indicate that all listeners rated the background noise in the RE condition as mostly very realistic sounding, while the HI listeners experienced the speech in the RE condition as overall more realistic and difficult to understand.

IV. DISCUSSION

A. Speech reception thresholds

SRTs for both listener groups were found to be 2–3 dB higher in the RE condition compared to the HP condition. This effect was likely caused by several factors. First, the HP condition used anechoic target sentences presented over headphones, as opposed to the reverberant ones presented over loudspeakers in the RE condition. Thus, in the RE condition, the target speech direct-to-reverberant energy ratio decreased considerably for the same broadband SNR, which has been shown to lead to decreased speech intelligibility (Roman and Woodruff, 2013). Second, the modulation spectra in Fig. 4(b) indicate the presence of modulation energy in the office meeting noise, in contrast to the stationary (and thus less modulated) speech-shaped HINT noise. These modulations were a consequence of the mixture of speech sources in the RE noise, but were less prominent than for the monophonic HINT target speech due to the room effect and the number of interfering talkers (Dreschler *et al.*, 2001). Still, this specific type of speech-like noise can lead to energetic speech-on-speech masking of the target in both the spectral (Brungart *et al.*, 2006) and the modulation spectral (Jørgensen and Dau, 2011) domains. Third, the many interfering talkers in the RE noise were intelligible and distributed throughout the frontal plane of the listener position. This may have produced informational masking (Westermann and Buchholz, 2015), with a detrimental effect on SI, especially since the male gender of the target talker matched that of 10 out of 12 interfering talkers in the room (Helfer and Freyman, 2008). The overall higher variance in the obtained data for the HI listeners compared to the NH listeners was expected and was most likely caused by the differences in hearing loss across the HI listeners.

While the transition from the AR condition to the RE condition led to an increase in SRTs for the HI listeners, this was not the case for the NH listeners. This may have resulted from a combination of effects. Comparing the LTAS [Fig. 4(a)] for the quadraphonic SSN in the AR condition to the LTAS of the office noise in the RE condition, there is a considerable decrease in spectral energy above 1 kHz for the AR noise. This lower amount of high-frequency noise of the office noise might reduce its speech masking effect for the NH listeners, while the HI listeners would not benefit to the same extent due to their increasing hearing loss at higher frequencies. However, this effect could be ruled out by testing a version of the AR noise that was spectrally matched to the office noise, rendering a similar NH and HI performance to the one reported here. Instead, one likely explanation is that the modulated maskers in the RE condition allowed for dip-listening, aiding the phonemic restoration of noisy speech (Warren, 1970) and leading to increased speech intelligibility (Peters *et al.*, 1998). Such dip-listening ability is commonly reduced in HI listeners, negatively affecting their SI performance [e.g., Takahashi and Bacon (1992)]. In addition, the effect of better-ear glimpsing on spatial release from energetic masking, a strategy used by NH listeners to increase their SI performance (Glyde *et al.*, 2013), has been shown to be limited in HI listeners, potentially due to reduced audibility, even at increased target-to-masker ratios (Best *et al.*, 2017b). Finally, the presence of realistic, meaningful speech in the RE condition may have increased the difficulty of the SI task to a greater extent in the HI listener group than in the NH listener group. This was evidenced by spontaneous and unanimous testimony by the HI listeners, who noted that the most challenging (and recognizable) aspect of performing the SI task in the RE condition was to not get distracted by the content of the background conversations. The NH listener group did not report these difficulties.

The increase in SRTs for both the NH and HI listeners between the HP condition and the AR condition occurred despite the fact that speech intelligibility typically increases when the target and the masker become spatially separated (Licklider, 1948), binaurally unmasking the speech from the noise (Durlach, 1963). However, the transition from diotic, anechoic target speech to spatialized, reverberant speech simultaneously decreased its spatial separation and its intelligibility. Spatial release from masking probably played a smaller role in the RE condition, since the background noise consisted of a large number of similar, interfering talkers (Freyman *et al.*, 2001).

Besides the overall increases in SRTs and decreases in SRSs obtained in the office meeting condition (RE), the results for the HI listeners differed considerably from NH performance in this condition compared to the corresponding results in the HP and AR conditions. The virtual sound environment, combined with more realistic target and masker stimuli, therefore might reflect some of the described hearing deficits in HI listeners.

B. Speech reception scores at the normal-hearing conversational SNR

While the adaptive SRTs results inform on changes in transitioning from an artificial to a more realistic SI task paradigm, speech reception scores at a constant SNR that represents normal-hearing conversation may provide insight into how the SI paradigm relates to the real world. For all

https://doi.org/10.1121/10.0004779

conditions and both listener groups, the SRSs followed the same trend as the SRTs obtained with the adaptive procedure. The word scoring procedure successfully avoided flooring effects, but the SRSs need to be corrected in order to compare them directly to the SRTs obtained with the adaptive, sentence-based scoring procedure. This is necessary because the word score of a HINT sentence between zero and four translates to a sentence score of zero, creating a non-linear negative bias of the sentence score versus the word score that increases with increasing word score. The distributions of the difference between the SRSs computed as word scores and those same SRSs computed as sentence scores are shown in Fig. 6. For all listeners, the word scores were consistently about 20%–25% higher than the sentence scores across conditions, indicating that most listeners still repeated 2–3 words correctly in a sentence marked as incorrect by sentence scoring.

The RE condition shows that, on average, the HI listeners correctly received just over one word out of two, while the NH listeners correctly received two words out of three. Thus, the HI listeners understood about half as many words as the NH listeners did. While this relative comparison is irrespective of absolute performance, the SRSs for the NH listeners were only at a level of 66% for an SNR where their ability to communicate was close to 100% in the real office meeting scene. SRSs that reflect real-world SI might be used as target percentage correct scores for NH listeners when conducting an adaptive SI tasks in other VSE-based critical sound scenarios to relate the obtained SRTs back to real-world SNRs necessary for proper speech communication. A percentage correct task at these SNRs would then reveal the comparable HI performance. Once an appropriate SNR for real-world NH SI has been established, this method could be used in any SI task to relate the performance of NH listeners to that of HI listeners.

Last, Fig. 7 shows the psychometric functions of the HP condition (left panel) and the RE condition (right panel) for the NH listeners (solid line) and the HI listeners (dashed line), derived by fitting a cumulative normal distribution to the pooled percentage correct scores per discrete SNR data

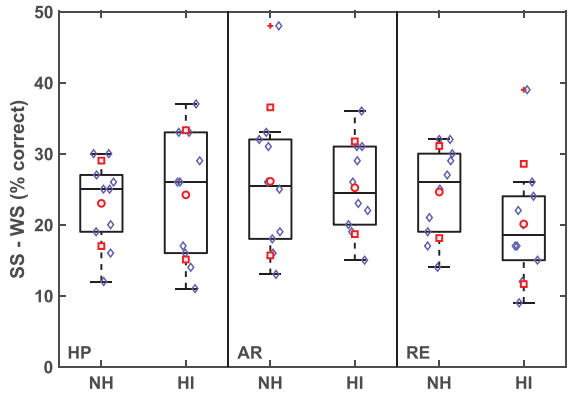


FIG. 6. (Color online) Distributions of the difference between the speech reception scores when calculated based on sentence scores (SS) and based on word scores (WS) in the percentage correct procedure at −2.5 dB SNR. Results are shown for the headphone condition (HP), the artificial speech-shaped noise condition (AR), and the office meeting noise condition (RE), for normal-hearing (NH) and hearing-impaired (HI) listeners.

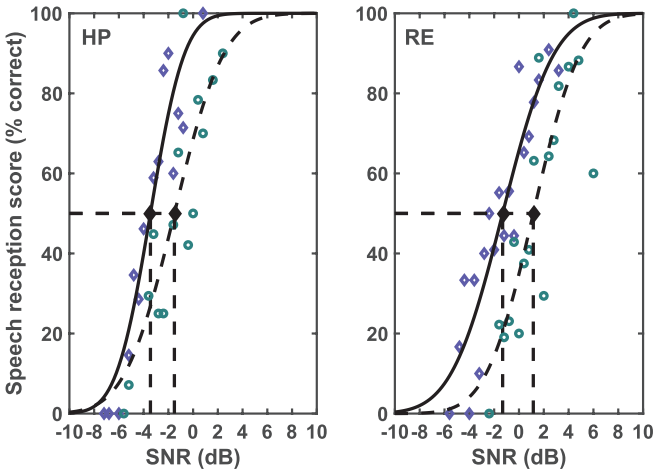


FIG. 7. (Color online) Psychometric functions for the HP condition (left panel) and RE condition (right panel) for the NH (solid line) and HI (dashed line) listeners. The NH and HI aggregated percentage scores are shown as diamonds (blue) and circles (green), respectively. The straight dashed lines that intersect the diamonds (black) relate the SNRs for both listener groups and conditions to the 50% correct point on the psychometric function.

point for the adaptive SRT procedure. These functions thus represent performance based on sentence scoring. The mean SRTs, corresponding to the 50% correct point on the psychometric function, are indicated by straight dashed lines intersecting black diamonds, and the aggregated NH and HI percentage scores are represented by blue diamonds and green circles, respectively. The RE condition resulted in narrow, steeply sloped psychometric functions for both listener groups, comparable to those obtained with in the HINT HP condition. The realistic VSE therefore seems to provide sensitive as well as stable SI outcome measures.

The SI task implemented in the office meeting VSE still remained limited in realism in several ways. No visual stimuli were presented in the laboratory environment alongside the auditory signals. In the HP condition, the absence of visuals matched the HINT procedure it represented, since no visual stimuli were used there either. It has been shown that speech reception scores can increase by 20% or more when the face of the target speaker is visible to the listener (Neely, 1956), an effect which becomes especially important at negative SNRs (Sumby and Pollack, 1954) and high background noise levels (Hadley *et al.*, 2019).

With regard to the acoustical reproduction accuracy of the SI stimuli, the VSE condition remains limited by the applied HOA recording and reproduction methods. The spatial aliasing frequency of the microphone array reduces the acoustic reliability of the office meeting recording at frequencies beyond 10 kHz. The Ambisonic reproduction order of 4 used by the loudspeaker array guarantees a sufficiently large sweet spot for the listener, but might not supply enough spatial accuracy to accurately replicate narrow acoustic sources. However, this reproduction error is offset by the presence of reverberation in the reproduced environment (Oreinos and Buchholz, 2015).

Furthermore, only a limited number of conditions were considered in this study, due to limitations in the size of the

SI speech corpus as well as time limitations in listener participation. It may be valuable to consider a condition with an unintelligible, phase-scrambled version of the office meeting background noise to assess the relative impact of informational masking and cognitive effort on SI performance, or to evaluate conditions with anechoic target speech in the spatialized maskers.

Last, while the experimental setup considered in this study was elaborate, it is not given that this level of sophistication is required to capture real-world SI performance. However, developing laboratory environments that approximate reality with increasing accuracy is a worthwhile endeavor, increasingly enabling the assessment of psychoacoustic phenomena beyond SI in an empirical way. A more qualitative argument in support of increasing realism in SI paradigms is the juxtaposition of experimental realism to mundane realism, as defined in psychology. Mundane realism refers to experimental conditions that mimic those of the real world as closely as possible, whereas experimental realism indicates the extent to which listeners actually experience those conditions as realistic (Aronson *et al.*, 1990). Therefore, to obtain meaningful results from a listener, his or her perception of realism may be just as important as its objective realization. Despite the mentioned limitations of the proposed SI paradigm, the questionnaire results from Table I confirmed that the experimental realism experienced by all listeners with respect to the sound of both the office meeting background noise as well as the speech stimuli was high. It was interesting to observe that, despite slight numerical differences, the overall distributions of difficulty and realism ratings were very similar for both listener groups. While the NH listeners achieved lower SRTs than the HI listeners, they rated the task in the realistic environments as similarly difficult as the HI listeners because the 50%-correct, adaptive HINT procedure presented both listeners groups with target speech sentences at similarly challenging SNRs.

V. SUMMARY AND CONCLUSION

A speech intelligibility task was designed and implemented, aiming to increase ecological validity and experimental realism with respect to the nature and presentation of the acoustic stimuli. It was shown that both NH and HI SRTs obtained in an HOA-reproduced office meeting critical sound scenario were, on average, 2–3 dB higher compared to the headphone-based HINT reference condition. These differences were found to be mainly due to the spatialization of the background noise (causing reverberation), the presence of speech-like modulations (causing speech-on-speech modulation masking) and the intelligibility of the interfering talkers (causing informational masking). Comparison with a spatialized artificial noise condition revealed that the HI listeners were more negatively affected by the realism in the VSE than the NH listeners, likely due to their reduced ability to use better-ear listening and listening in the dips, as well as due to an increased cognitive effort to focus on the target speech in the presence of intelligible, interfering speech-like noise. SRSs provided a way to relate SI performance to potential difficulties experienced by HI listeners in the real world, by evaluating SI at a constant SNR at which NH communication ability was close to 100%. The approach presented in this study might be valuable for investigations into the effects of hearing loss and hearing aid benefit on SI in simulated real-world environments and could be extended by providing visual information to increase the realism of the simulated environment.

ACKNOWLEDGMENTS

The research was supported by the Technical University of Denmark and the Centre for Applied Hearing Research (CAHR). We would like to thank Jens-Bo Nielsen for providing the graphical user interface for the headphone-based Danish HINT and Rikke Sørensen for screening and scoring the participating listeners.

[1]Contrary to the SRT results, the Bartlett test rejected the null-hypothesis of equal variance between the groups ($p = 0.0337$), but given the balanced group size and the borderline significance, the MANOVA was still valid (Stevens, 2012).

Ahrens, A., Marschall, M., and Dau, T. (**2017**). "Measuring speech intelligibility with speech and noise interferers in a loudspeaker-based virtual sound environment," J. Acoust. Soc. Am. **141**(5), 3510–3510.

Ahrens, A., Marschall, M., and Dau, T. (**2019**). "Measuring and modeling speech intelligibility in real and loudspeaker-based virtual sound environments," Hear. Res. **377**, 307–317.

Aronson, E., Carlsmith, J. M., and Ellsworth, P. C. (**1990**). *Methods of Research in Social Psychology* (McGraw-Hill, New York).

Bertet, S., Daniel, J., and Moreau, S. (**2006**). "3D sound field recording with higher order ambisonics-objective measurements and validation of spherical microphone," in *Audio Engineering Society Convention 120*, Audio Engineering Society.

Best, V., Buchholz, J. M., and Weller, T. (**2017a**). "Measuring auditory spatial perception in realistic environments," J. Acoust. Soc. Am. **141**(5), 3692–3692.

Best, V., Keidser, G., Buchholz, J. M., and Freeston, K. (**2015**). "An examination of speech reception thresholds measured in a simulated reverberant cafeteria environment," Int. J. Audiol. **54**(10), 682–690.

Best, V., Mason, C. R., Swaminathan, J., Roverud, E., and Kidd, G., Jr. (**2017b**). "Use of a glimpsing model to understand the performance of listeners with and without hearing loss in spatialized speech mixtures," J. Acoust. Soc. Am. **141**(1), 81–91.

Bisgaard, N., Vlaming, M. S., and Dahlquist, M. (**2010**). "Standard audiograms for the IEC 60118-15 measurement procedure," Trends Amplif. **14**(2), 113–120.

Bradley, J. V. (**1958**). "Complete counterbalancing of immediate sequential effects in a Latin square design," J. Am. Stat. Assoc. **53**(282), 525–528.

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**(6), 4007–4018.

Cubick, J., and Dau, T. (**2016**). "Validation of a virtual sound environment system for testing hearing aids," Acta Acust. Acust. **102**(3), 547–557.

Culling, J. F. (**2016**). "Speech intelligibility in virtual restaurants," J. Acoust. Soc. Am. **140**(4), 2418–2426.

Dreschler, W. A., Verschuure, H., Ludvigsen, C., and Westermann, S. (**2001**). "ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment: Ruidos icra: Señates de ruido artificial con espectro similar al habla y propiedades temporales para pruebas de instrumentos auditivos," Audiology **40**(3), 148–157.

Durlach, N. I. (**1963**). "Equalization and cancellation theory of binaural masking-level differences," J. Acoust. Soc. Am. **35**(8), 1206–1218.
Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2001**). "Spatial release from informational masking in speech recognition," J. Acoust. Soc. Am. **109**(5), 2112–2122.
Glyde, H., Buchholz, J., Dillon, H., Best, V., Hickson, L., and Cameron, S. (**2013**). "The effect of better-ear glimpsing on spatial release from masking," J. Acoust. Soc. Am. **134**(4), 2937–2945.
Hadley, L. V., Brimijoin, W. O., and Whitmer, W. M. (**2019**). "Speech, movement, and gaze behaviours during dyadic conversation in noise," Sci. Rep. **9**(1), 1–8.
Helfer, K. S., and Freyman, R. L. (**2008**). "Aging and speech-on-speech masking," Ear Hear. **29**(1), 87.
Houben, R., Koopman, J., Luts, H., Wagener, K. C., Van Wieringen, A., Verschuure, H., and Dreschler, W. A. (**2014**). "Development of a Dutch matrix sentence test to assess speech intelligibility in noise," Int. J. Audiol. **53**(10), 760–763.
ITU-T (**2018**). Recommendation ITU-T P.570: Artificial Noise Fields under Laboratory Conditions (International Telecommunication Union), http://handle.itu.int/11.1002/1000/13624 (Last viewed August 15, 2020).
Jørgensen, S., and Dau, T. (**2011**). "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," J. Acoust. Soc. Am. **130**(3), 1475–1487.
Kelly, H., Lin, G., Sankaran, N., Xia, J., Kalluri, S., and Carlile, S. (**2017**). "Development and evaluation of a mixed gender, multi-talker matrix sentence test in Australian English," Int. J. Audiol. **56**(2), 85–91.
KEMAR (**2018**). GRAS 45BC KEMAR Head & Torso with Mouth Simulator, GRAS Sound & Vibration A/S, Holte, Denmark.
Kinnunen, T., and Li, H. (**2010**). "An overview of text-independent speaker recognition: From features to supervectors," Speech Commun. **52**(1), 12–40.
Licklider, J. (**1948**). "The influence of interaural phase relations upon the masking of speech by white noise," J. Acoust. Soc. Am. **20**(2), 150–159.
Mansour, N., Marschall, M., May, T., Westermann, A., and Dau, T. (**2019**). "A method for conversational signal-to-noise ratio estimation in real-world sound scenarios," J. Acoust. Soc. Am. **145**(3), 1873–1873.
Minnaar, P., Favrot, S., and Buchholz, J. M. (**2010**). "Improving hearing aids through listening tests in a virtual sound environment," Hear. J. **63**(10), 40–42.
Moore, B. C. (**1996**). "Perceptual consequences of cochlear hearing loss and their implications for the design of hearing aids," Ear Hear. **17**(2), 133–161.
Müller, S., and Massarani, P. (**2001**). "Transfer-function measurement with sweeps," J. Audio Eng. Soc. **49**(6), 443–471.
Neely, K. K. (**1956**). "Effect of visual factors on the intelligibility of speech," J. Acoust. Soc. Am. **28**(6), 1275–1277.
Nielsen, J. B., and Dau, T. (**2011**). "The Danish hearing in noise test," Int. J. Audiol. **50**(3), 202–208.
ODEON (**2018**). *ODEON User Manual*, version 15 (ODEON A/S, Kgs. Lyngby).
Oreinos, C., and Buchholz, J. M. (**2015**). "Objective analysis of ambisonics for hearing aid applications: Effect of listener's head, room reverberation, and directional microphones," J. Acoust. Soc. Am. **137**(6), 3447–3465.
Peters, R. W., Moore, B. C., and Baer, T. (**1998**). "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," J. Acoust. Soc. Am. **103**(1), 577–587.
Reis, H. T., and Judd, C. M. (**2000**). *Handbook of Research Methods in Social and Personality Psychology* (Cambridge University Press, Cambridge, UK).
Roman, N., and Woodruff, J. (**2013**). "Speech intelligibility in reverberation with ideal binary masking: Effects of early reflections and signal-to-noise ratio threshold," J. Acoust. Soc. Am. **133**(3), 1707–1717.
Smeds, K., Wolters, F., Larsson, J., Herrlin, P., and Dahlquist, M. (**2018**). "Ecological momentary assessments for evaluation of hearing-aid preference," J. Acoust. Soc. Am. **143**(3), 1742–1742.
Smeds, K., Wolters, F., and Rung, M. (**2015**). "Estimation of signal-to-noise ratios in realistic sound scenarios," J. Am. Acad. Audiol. **26**(2), 183–196.
Soli, S. D., and Wong, L. L. (**2008**). "Assessment of speech intelligibility in noise with the hearing in noise test," Int. J. Audiol. **47**(6), 356–361.
Stevens, J. P. (**2012**). *Applied Multivariate Statistics for the Social Sciences* (Routledge, London), p. 249.
Sumby, W. H., and Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. **26**(2), 212–215.
Takahashi, G. A., and Bacon, S. P. (**1992**). "Modulation detection, modulation masking, and speech understanding in noise in the elderly," J. Speech Lang. Hear. Res. **35**(6), 1410–1421.
Timmer, B. H., Hickson, L., and Launer, S. (**2017**). "Ecological momentary assessment: Feasibility, construct validity, and future applications," Am. J. Audiology **26**(3S), 436–442.
Wagener, K., Josvassen, J. L., and Ardenkjær, R. (**2003**). "Design, optimization and evaluation of a Danish sentence test in noise: Diseno, optimización y evaluación de la prueba danesa de frases en ruido," Int. J. Audiol. **42**(1), 10–17.
Warren, R. M. (**1970**). "Perceptual restoration of missing speech sounds," Science **167**(3917), 392–393.
Weisser, A., Buchholz, J. M., Oreinos, C., Badajoz-Davila, J., Galloway, J., Beechey, T., and Keidser, G. (**2019**). "The ambisonic recordings of typical environments (ARTE) database," Acta Acust. Acust. **105**(4), 695–713.
Westermann, A., and Buchholz, J. M. (**2015**). "The influence of informational masking in reverberant, multi-talker environments," J. Acoust. Soc. Am. **138**(2), 584–593.
Westermann, A., and Buchholz, J. M. (**2017**). "The effect of nearby maskers on speech intelligibility in reverberant, multi-talker environments," J. Acoust. Soc. Am. **141**(3), 2214–2223.
Wightman, F. L., and Kistler, D. J. (**1989**). "Headphone simulation of free-field listening. I: Stimulus synthesis," J. Acoust. Soc. Am. **85**(2), 858–867.
Wolters, F., Smeds, K., Schmidt, E., Christensen, E. K., and Norup, C. (**2016**). "Common sound scenarios: A context-driven categorization of everyday sound environments for application in hearing-device research," J. Am. Acad. Audiol. **27**(7), 527–540.