The effect of nearby maskers on speech intelligibility in reverberant, multi-talker environments

Adam Westermann and Jörg M. Buchholz

Citation: The Journal of the Acoustical Society of America **141**, 2214 (2017); doi: 10.1121/1.4979000

View online: http://dx.doi.org/10.1121/1.4979000

View Table of Contents: http://asa.scitation.org/toc/jas/141/3

Published by the Acoustical Society of America

The effect of nearby maskers on speech intelligibility in reverberant, multi-talker environments[a]

Adam Westermann[b] and Jörg M. Buchholz
National Acoustic Laboratories, Australian Hearing, 16 University Avenue, Macquarie University, New South Wales 2109, Australia

(Received 30 November 2016; revised 10 February 2017; accepted 8 March 2017; published online 30 March 2017)

The extent to which informational masking (IM) is involved in real-world listening is not well understood. In the literature, IM effects of more than 8 dB are reported, but these experiments typically used simplified spatial configurations and speech materials with exaggerated confusions. Westermann and Buchholz [(2015b). J. Acoust. Soc. Am. **138**, 584–593] considered a simulated cafeteria and found only substantial IM effects when the target and maskers were colocated and the same talker. The present study further investigates the relevance of IM in real-world environments, specifically distractions by nearby maskers and the effect of hearing impairment. Speech reception thresholds (SRTs) were measured with normal hearing (NH) and sensorineural hearing impaired (HI) listeners in a simulated cafeteria environment. Three different masker configurations were considered: (1) seven dialogues distributed in the cafeteria, (2) two monologues presented close to the listener with varying angular separation, and (3) a combination of (1) and (2). The contribution of IM was measured as the difference in SRTs between speech maskers and unintelligible vocoded maskers. No significant IM was found with the seven dialogues alone. Including nearby maskers resulted in substantial IM for both NH and HI listeners, suggesting that such maskers might result in IM in real-world environments. © *2017 Acoustical Society of America*.

[http://dx.doi.org/10.1121/1.4979000]

[AKCL] Pages: 2214–2223

I. INTRODUCTION

For years, researchers have investigated the auditory mechanisms related to understanding speech in reverberant multi-talker environments, often described as the "cocktail-party effect" (e.g., Cherry, 1953; Bronkhorst, 2000). In such challenging conditions, it has been shown that the auditory system can take advantage of talker characteristics, such as differences in fundamental frequency, spatial location and fluctuations in maskers to better understand a target talker (Brungart *et al.*, 2001; Freyman *et al.*, 1999; Festen and Plomp, 1990). When speech is masked by speech, the concepts of informational masking (IM) and energetic masking (EM) are often applied (for a review see Kidd *et al.*, 2007). While EM describes masking effects that occur because of spectro-temporal overlap in the auditory periphery, IM is often related to more central masking effects which interfere either with auditory object formation or stream selection (Shinn-Cunningham, 2008). Here, we define this type of IM as resulting from *confusions* due to the uncertainty about which signal component belongs to the target and which to the masker. Another type of IM results from maskers which *distract* listeners by drawing their attention. This could be maskers discussing a topic which interests the listener or the maskers saying the listeners name (Wood and Cowan, 1995).

Several studies have tried to quantify the influence of IM. Generally, these studies employ a reference condition with a high level of target-masker confusions and a method that enables segregation of the target and masker signals in order to measure the reduction, or release from, IM. Such reference conditions normally include target and masker in the same location (colocated) (Freyman *et al.*, 1999), and speech corpora with many inherent confusions (Bolia *et al.*, 2000). Thereby, confusions are often created by using the same talker to realize target and masker speech as well as using speech material that has a very similar, synchronized sentence structure for the target and maskers. Perceptual segregation has been introduced by changing the angular separation or distance between a target and maskers (Freyman *et al.*, 1999; Westermann and Buchholz, 2015a), or the gender and talker characteristics (Brungart *et al.*, 2001). Other studies have estimated the influence of IM by comparing intelligibility with a speech masker and a speech-modulated noise-masker (Best *et al.*, 2013; Brungart *et al.*, 2001). Across studies it has been suggested that IM effects can result in differences of up to 8 dB in measured SRTs. However, the reference condition applied in all of these studies relies on speech corpora and spatial configurations with exaggerated confusions.

Westermann and Buchholz (2015b) investigated the influence of IM in a simulated cafeteria environment presented via a three-dimensional loudspeaker array. They measured SRTs in a background of dialogues consisting of the same talker as the target, different talkers or unintelligible noise vocoded

[a] Aspects of this work were presented at the Association for Research in Otolaryngology (ARO) MidWinter meeting 2014 and at the International Hearing Aid Research Conference (IHCON) 2014.

[b] Also at: Department of Linguistics, Macquarie University, Building C5A, Balaclava Road, North Ryde, New South Wales 2109, Australia. Electronic mail: adwe@widex.com

talkers that were either colocated with the target or distributed throughout the simulated room. Overall, they found no contribution of IM in conditions that were representative of real-life, i.e., when the masking talkers were spatially distributed and different from the target. Furthermore, they argued that in conditions where limited cues were available to segregate the target from the maskers the contribution of IM was dependent on the level of the target talker compared to the level of each individual masker, known as the target-to-masker ratio (TMR) (Brungart *et al*., 2001). Since the TMRs were predominantly positive in their simulated cafeteria (as in most real-world environments: Smeds *et al*., 2014), they concluded that the influence of IM is low in realistic environments. However, they mainly considered confusion-based IM and did not consider the effect of maskers close to the listener, which typically provide rather low TMRs and due to their salience are also more likely to provide distraction-based IM. They also did not consider the effect of a hearing impairment, which may decrease the quality of auditory cues and thus, may affect the susceptibility to either form of IM. Moreover, since hearing impairment is often age-related, cognitive decline with increasing age may further influence the observed IM.

Few studies have looked into the effect of IM with masker that are closer to the listener than the target. Lavandier and Culling (2007) employed a number of conditions with nearby maskers to measure the effect of the direct-to-reverberant ratio on spatial release from masking (SRM), but they always kept target and masker with an angular separation of 65° which likely resolved IM. Westermann and Buchholz (2015a) investigated the effect of differences in distance on SRM when target and masker were directly in front of the listener with normal hearing (NH), and later, with hearing impaired (HI) listeners (Westermann and Buchholz, 2017). They found that placing a masker further away in distance from the target resolves IM and leads to improved SRTs. However, in the opposite case, when the masker was closer to the listener than the target, they observed a substantial IM effect. This effect was especially pronounced with HI listeners. Analyzing the errors the listeners made when the masker was closer to the target, they found few target-masker confusions and thereby concluded that the involved IM may be related to the distraction by the maskers (i.e., affecting selective attention) rather than confusions. However, their experiment applied highly confusing speech corpora and a colocated reference condition, both limiting the ecological validity of their results.

The current study presents a speech intelligibility test in a simulated cafeteria that combines the increased ecological validity of Westermann and Buchholz (2015b) with the nearby maskers of Westermann and Buchholz (2015a). The test was specifically designed to estimate IM effects with nearby maskers in realistic environments, and it was conducted on both normal hearing NH and HI listeners.

II. METHODS

A. Subjects

In this study 16 NH (12 female, six male) and 16 HI (six female, 10 male) native Australian English speaking subjects participated. The mean age of the NH subjects was 29.2 yr, and the subjects had pure-tone audiometric thresholds ≤20 dB hearing loss (HL) at audiometric frequencies from 250 Hz to 8000 kHz. These subjects were either employees of the National Acoustic Laboratories or students at Macquarie University. The HI listeners had a mean age of 72.5 yr and all had symmetric (threshold differences between ears of ≤10 dB), mild to moderate sensorineural hearing losses. Individual audiograms and mean and standard deviation of the audiometric thresholds are shown in Fig. 1. All HI subjects had extensive experience with psychoacoustic experiments. Before the study, all participants gave written consent and subjects not associated with National Acoustic Laboratories (NAL) were given a gratuity for their participation. Ethical clearance was received from the Macquarie University Human Research Ethics Committees (Reference No: 5201300165).

B. Stimuli

The experiments used sentences from the Bamford–Kowal–Bench (BKB) corpus (Bench *et al*., 1979). This speech corpus contains 336 sentences with an approximate length of 1.5 s and simple syntactical structure (e.g., "The girl lost her doll"). Sentences are spoken by a native Australian–English male speaker, sampled at 44.1 kHz and divided into 21 lists. As in Westermann and Buchholz (2015b) a 512-tap finite impulse response (FIR) filter was applied to the original BKB material to make its long-term spectra match the long-term spectrum of a long (65 min) anechoically recoded monologue spoken by the original speaker. This filtering stage removed the original spectral shaping which was applied to the BKB corpus to match the long-term average speech spectrum (LTASS) defined in Byrne *et al*. (1994) and made the sentences sound more natural in the cafeteria background described in Sec. II D.

Two different maskers were realized to estimate the influence of IM: (1) a *speech masker* comprised of talkers different than the target talker and (2) a *vocoded masker* which was a vocoded version of the speech masker. The influence of IM was defined as the difference in SRTs between the two maskers. The maskers were approximately 5 min in length and were continuously looped throughout the experiment.

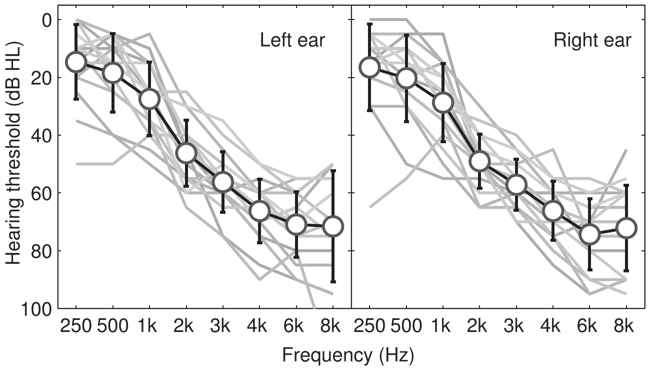

FIG. 1. Individual (gray lines), mean and ±1 standard deviation of the pure-tone audiometry for the 16 HI participants.

The speech masker consisted of a mixture of seven background two-talker dialogues, as used in Westermann and Buchholz (2015b), and two nearby maskers containing monologues. The dialogues and monologues were taken from the International English Language Testing System (IELTS) and recorded in the anechoic chamber at the National Acoustic Laboratories with six male and eight female, native Australian English speakers. The level of each individual talker was equalized using the speech level calculation methodology of the Speech Transmission Index (STI; IEC 60268-16). This processing was mainly applied to disregard the long speech pauses due to the turn-taking in the dialogues. Due to the limited number of recorded talkers, the monologues were spoken by two male talkers that also appeared in the dialogues.

To separate the effects of EM and IM, a vocoded version of each of the speech maskers was implemented (same as used in Westermann and Buchholz, 2015b). The aim of the vocoder processing was to make sure that the combined, multi-talker background speech was completely unintelligible while maintaining the spatial percept of multiple noise-sources distributed around the listener. This was realized by preserving a high temporal resolution, thereby maintaining transients (as much as possible), and in turn, strongly smoothing the spectra. The short-time Fourier transform (STFT), with window length of 20 ms and 75% overlap, was used to convert each of the anechoic speech maskers to the time-frequency domain. The resulting time-frequency representation was then spectrally smoothed across either one octave for the seven background dialogues or two octaves for the nearby maskers. The additional smoothing of the nearby masker was applied to assure that the vocoded speech was unintelligible. The vocoded signal was reconstructed using the inverse short-time Fourier transform by combining the smoothed magnitude spectrum with the phase-spectrum from white noise.

C. Equipment

The speech testing was conducted in a spherical loudspeaker array in the anechoic chamber at the National Acoustic Laboratories. Outside the anechoic chamber, a PC running MATLAB generated and played the sound files. The PC was fitted with a RME MADI sound card connected to two RME M-32 D/A converters. The analog output of the converters was amplified by eleven 4-channel Yamaha XM4180 amplifiers whose output was fed into the anechoic chamber through an acoustically dampened passage and connected to each individual loudspeaker in the array. The loudspeaker array consisted of 41 Tannoy V8 loudspeakers arranged on a sphere with a radius of 1.85 m. The subject was seated on a height-adjustable chair such that their head was in the center of the loudspeaker array. To reproduce the direct sound component of the nearby maskers, four 8080 Genelec monitor loudspeakers were suspended inside the array in level with the primary ring of 16 loudspeakers at a distance of 0.85 m from the center of the array. These small speakers were placed between the array loudspeakers at ±11.25° and ±56.25° and hung only with thin strings to minimize acoustical shadow. Since the Genelec monitors contained amplifiers, they were connected directly to the RME M-32 D/A converter through a balanced cable.

D. Spatialization of sounds

The loudspeaker array described in Sec. II C was used to reproduce the cafeteria environment shown in Fig. 2. First, a model of the room was created in ODEON software (Rindel, 2000), which included various surfaces such as tables, chairs and windows, each with their individual inherent absorption coefficients. The room was 15 m long by 8.5 m wide by 2.8 m high and had a reverberation time of $T_{30} \approx 0.6$ s. The sources were placed as shown in Fig. 2 and realistic talker directivity was included by applying ODEON's directivity

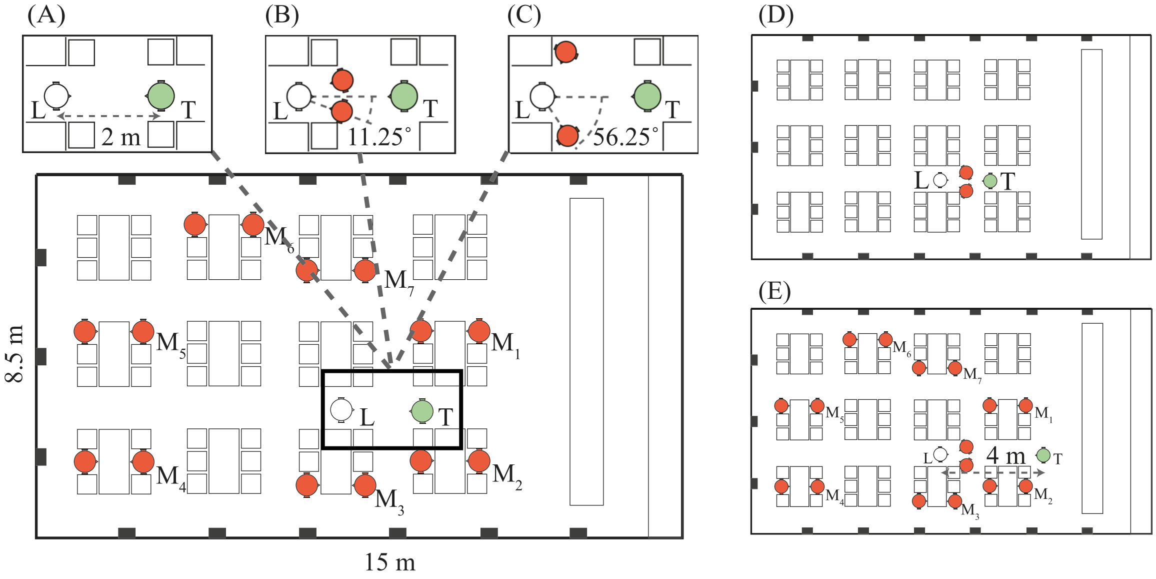

FIG. 2. (Color online) Top-down view of cafeteria simulated in ODEON for each of the measured conditions (A)–(E). The listener, L, faces the target, T, at either 2 m (A)–(D) or 4 m (E) distance. The masking dialogues were distributed in the room (M_1,…,M_7) and the nearby maskers were separated from the target by either ±11.25° or ±56.25°.

file Tlknorm_natural.so8. For each source the acoustic path to the listener, as captured by the room impulse response (RIR), was calculated with the loudspeaker-based room auralization (LORA) toolbox (Favrot and Buchholz, 2010) using reflectograms and decay curves provided by ODEON. Within the LORA toolbox, the reflectograms are used to map the direct sound and specular early reflections (here up to third order) to the nearest loudspeaker in the playback loudspeaker array. The late reverberation was added by applying the frequency- and direction-dependent decay envelope to uncorrelated noise. This resulted in 41 impulse responses (IRs), corresponding to each channel in the loudspeaker array, for each sound source (excluding the nearby maskers).

Four of the conditions included nearby maskers [Fig. 2(B)–2(E)] which were substantially closer to the listener than the loudspeakers in the array. The direct sound component of the nearby maskers was presented from the additional loudspeakers suspended in the array, whereas the remaining part of the RIR was reproduced using the 41-channel loudspeaker array. Thereby for each of the nearby maskers, a 42-channel IR was derived.

In order to spatialize each sound source in the simulated cafeteria (Fig. 2), the anechoic speech recordings described in Sec. II B were convolved with the corresponding multi-channel IRs. To reduce individual variation of loudspeaker sensitivity and to compensate for the difference in arrival time of the near loudspeakers, equalization filters were designed for each loudspeaker and applied to all stimuli.

The maskers were fixed to an ecologically appropriate sound pressure level (SPL). In addition to room acoustical information about each source, ODEON also estimates its level at the receiver, assuming a given vocal effort. Here "normal" vocal effort was applied, and the overall level of the different cafeteria backgrounds was calculated by summing the predicted power of all involved sources. For the conditions shown in Fig. 2 the SPLs were 59.4, 65.4, 65.4, 62.3, and 65.4 dB(A) for Figs. 2(A)–2(E), respectively.

In order to (partially) restore audibility for the HI participants, the NAL-RP linear amplification scheme was adopted (Dillon, 2001). To remove effects and variability related to hearing aid processing and to ensure externalization of sound sources (i.e., localizing sources outside the head, rather than inside) the hearing loss compensation was performed at the loudspeakers rather than at the listeners' ears. However, this removed the ability to apply different prescriptions across ears, and as a result, symmetrical hearing loss was a recruitment requirement. The measured hearing level (HL) was entered in the NAL-RP formula and an insertion gain in third-octave channels was calculated and applied to a series of 20 third-octave linear-phase finite impulse response (FIR) filters covering a frequency range from 100 Hz to 8 kHz. The weighted third-octave filters were summed to provide a single, subject-specific FIR filter which was then applied to all 45 loudspeakers (including the four nearby loudspeakers) to realize individual amplification according to NAL-RP.

E. Procedures

Testing of each participant was completed in one appointment. For both NH and HI listeners, an audiometric screening was fist conducted in a double-walled booth. While in the booth, the HI listeners also completed a computerized reading span test (RST) and Stroop test (Sec. IV C). These tests took approximately 15 min. Subsequently, participants were taken to the anechoic chamber and seated in the center of the loudspeaker array. The height of the chair was adjusted to ensure that the subject's head was at level with the center of the array. The participants were instructed on the task and fitted with a lavalier microphone to communicate with the test administrator outside the chamber. For the HI listeners, the measured audiogram was entered in a MATLAB script which in turn computed and applied the NAL-RP equivalent loudspeaker gain.

Preceding the main speech test, a short training session was conducted to familiarize the participants with the task and interaction with the test administrator outside the chamber. Here, one of the 21 lists was presented in the cafeteria background without the nearby maskers [Fig. 2(A)]. The results obtained from the familiarization were discarded. During the test 10 SRTs were measured, i.e., Figs. 2(A)–2(E) each with the speech and vocoded masker (Sec. II B). The masker level was kept constant while the target level varied according to an adaptive, one-up one-down staircase method to estimate the signal-to-noise ratio (SNR) which yielded 50% correct performance. The testing procedure was implemented following Keidser *et al.* (2013), requiring for each speech reception threshold (SRT) a minimum of 16 presentations with decreasing step sizes of 5, 2, and 1 dB. A run ended either when the standard error fell below 0.8 dB or when 32 sentences were presented. Since each list of the BKB material contained 16 sentences, two randomly selected lists were combined for each measured SRT. Across participants the order of presentation was randomized. The entire test took approximately 1.5 h, and halfway through the participants were given a short break.

III. RESULTS

The top panels of Fig. 3 shows the mean and 95% confidence intervals of the measured SRTs in reference to free field level. The triangles and squares show the results for the NH and HI listeners, respectively, and the gray symbols denote the speech masker and black symbols denote the vocoded masker. The bottom panels of Fig. 3 show the estimated IM, which is the difference between the speech and vocoded masker SRT, calculated individually per subject. Two-way repeated measures analysis of variance (ANOVA) were applied separately to the NH and HI data. For the NH listeners, it showed significance for condition [$F(2.48, 37.26) = 79.29$, $p < 0.001$], type of masker [$F(1, 15) = 226.07$, $p < 0.001$] and interaction between the two [$F(4, 60) = 13.52$, $p < 0.001$]. Note, the Greenhouse–Geisser correction was applied to the condition effects to ensure that a violation of the sphericity assumption did not influence the significance of the observed effects. For the

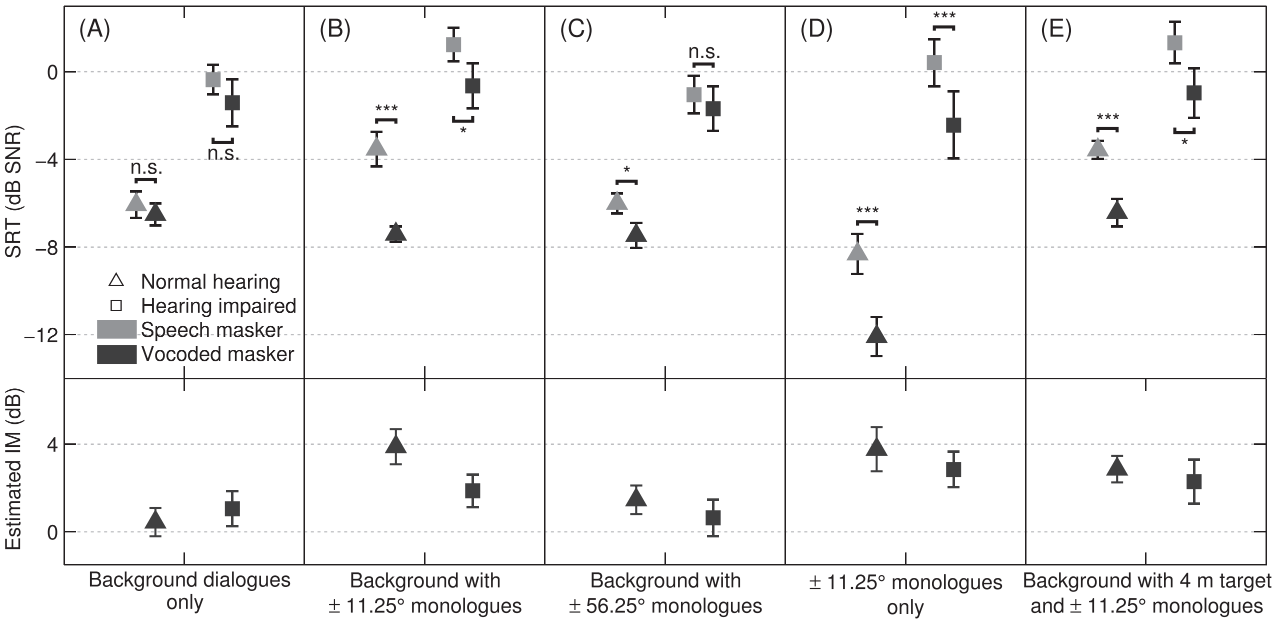

FIG. 3. Top panels: Mean and across-subject 95% confidence interval of SRTs for conditions A through E according to Fig. 2. Speech and vocoded maskers are indicated by gray and black symbols, respectively, and the NH participants by triangles and HI by squares. Stars indicate level of significance between conditions (i.e., $*$, $**$, and $***$ correspond to $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively). Bottom panels: Mean and across-subject 95% confidence intervals of the vocoder benefit (i.e., the difference between the speech masker SRT and the vocoded masker SRT).

HI listeners the two-way ANOVA also showed significance for condition [$F(3,45) = 9.74$, $p < 0.001$], masker type [$F(1,45) = 36.93$, $p < 0.001$] and interaction [$F(3,45) = 6.62$, $p < 0.001$]. *Post hoc t-tests* with Bonferroni correction were applied to compare the speech and vocoded masker SRT and results are indicated in Fig. 3 (i.e., *, **, and *** correspond to significance of $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively).

A. Normal hearing listeners

Overall, with only the background cafeteria [Fig. 3(A)] the difference in SRTs between the speech and vocoded masker was not significant for the NH listeners. This is a repeated condition from Westermann and Buchholz (2015b) and confirms their findings.

When the nearby maskers are introduced in Figs. 3(B)–3(E), substantial and significant differences between speech masker and vocoded masker SRTs of up to 4 dB can be observed. This suggests that IM effects can indeed be observed in realistic environments, but only when nearby maskers are present. The increased separation when the nearby maskers were shifted from ±11.25° to ±56.25° [Figs. 3(B) and 3(C)] substantially reduced the speech masker SRTs, but did not change the vocoded SRTs. This indicates a spatial release from IM as angular separation increases while EM remains fairly constant.

In the condition with only the masker at ±11.25° and no background maskers [Fig. 3(D)], the SRTs were considerably lower than when the background maskers were included. This indicates that considerable dip-listening is available to the NH listeners. However, the estimated IM for the two conditions is very similar, confirming that IM is introduced by the two nearby maskers and not disturbed by the cafeteria background.

There was not a big change in SRTs when the target was moved from 2 to 4 m distance [between Figs. 3(B) and 3(E)]. Only the SRT for the vocoded masker increased slightly, which resulted in a small reduction in IM of less than 1 dB.

B. Hearing impaired listeners

The overall behavior of the results for the HI subjects, shown as squares in Fig. 3, is similar to the NH listeners, except that SRTs are about 4 dB higher on average for the HI listeners. Higher thresholds were expected because of the reduced audibility, frequency selectivity and temporal resolution of the impaired auditory system. In addition, the overall benefit of removing the background cafeteria dialogues [between Figs. 3(A) and 3(D)] was smaller for the HI than the NH listeners. However, such reduced ability to effectively use masker fluctuations (or "listen in the dips") following a hearing impairment has been reported in multiple studies (e.g., Festen and Plomp, 1990; Bernstein and Grant, 2009).

Similar to NH listeners, also no significant involvement of IM was observed for the HI listeners in the cafeteria background without nearby maskers [Fig. 3(A)], although the SRT for the speech masker showed a tendency towards higher values than the vocoded masker. This could indicate that HI listeners are more susceptible to IM than NH listeners when multiple, partially intelligible masking talkers are involved. Also similar to NH subjects, when the nearby maskers were moved from ±11.25° to ±56.25° [comparing Figs. 3(B) and 3(C)], the SRT for the speech masker decreased significantly but not for the vocoded masker and thus, highlighting a substantial spatial release from IM.

Somewhat surprising, the HI listeners showed smaller differences in SRTs between speech and vocoded maskers when the nearby maskers were present [Figs. 3(B)–3(E)], demonstrating that the involved IM is still significant but about 1–2 dB smaller than in NH listeners. Generally, it has been argued that IM effects are independent of hearing loss (Agus *et al.*, 2009; Helfer and Freyman, 2008), and as far as the authors are aware no studies have suggested reduced IM following a hearing impairment. Brungart *et al.* (2001) argued that the occurrence of IM is dependent on the TMR, and diminishes at positive TMRs. The latter conclusion is further supported by a number of related studies, including Helfer and Freyman (2008) and Agus *et al.* (2009). To illustrate the TMRs involved in this experiment, in particular between the target and the nearby maskers, the SRT data shown in Fig. 3 is replotted in Fig. 4, but this time the SPL of the target at the SRT is shown instead of the SNR at the SRT. Additionally, the overall level of the masker is shown for each condition by an asterisk (∗), and in the conditions with a nearby masker, the level of each individual nearby masker is shown by a circle (○). The effective TMR in relation to each of the nearby maskers is the difference between the target SPL value and the SPL of the nearby masker. In conditions where the nearby maskers were present [Figs. 4(B), 4(C), and 4(E)], the NH TMRs were close to 0 dB, while the HI TMRs were around +5 dB. In the nearby masker alone condition [Fig. 4(D)], the TMRs were around −5 to −8 dB for NH and around +1 to +4 dB for HI subjects. Hence, the NH and HI subjects were tested at different TMRs, which might explain the difference in the observed IM.

IV. DISCUSSION

A. Effect of informational masking

Westermann and Buchholz (2015b) did not find any significant IM in a simulated cafeteria environment when the maskers were different talkers from the target and spatially distributed. Their conclusion was confirmed in the present study, considering the same cafeteria background masker [Fig. 3(A)]. However, in all other conditions [Figs. 3(B)–3(E)], in which nearby maskers were introduced, a significant IM effect of up to 4 dB was observed.

Examples of such conditions in real-life could be classrooms in which a student is trying to listen to a teacher while nearby students are talking, or at the dinner table when trying to listen to a conversation across the table. However, it is not known how and if the IM effect reported in this work translates to real-life listening; especially considering that the effect was only pronounced when the maskers were located in a similar direction as the target, i.e., at ±11.25° with the target at 0°. Moreover, Helfer and Freyman (2005) showed that the inclusion of visual cues, allowing lip-reading, significantly reduces the contribution of IM. Although visual cues are often available in real life, they require the presence of sufficient light and line of sight to the talker of interest.

Westermann and Buchholz (2015a) found substantial IM effects when applying a speech masker that was closer to the listener than the target. Because of the properties of the Coordinate response measure (CRM) corpus, they were able to measure the number of target-masker confusions, i.e., the masker sentence reported instead of the target sentence. They observed that with nearby maskers, the subjects rarely reported the color-number combination of the maskers (5.7%), which in turn happened more often in the colocated condition (14.8%). In this way, they argued that this IM was not caused by confusions but rather by the maskers distracting the listener and thus, hindering attention steering. Since a setup with some similar features is realized here, it is likely that mainly distraction- or attention-based IM is involved. However, since the employed speech corpus does not allow counting target-masker confusions, it is difficult to draw

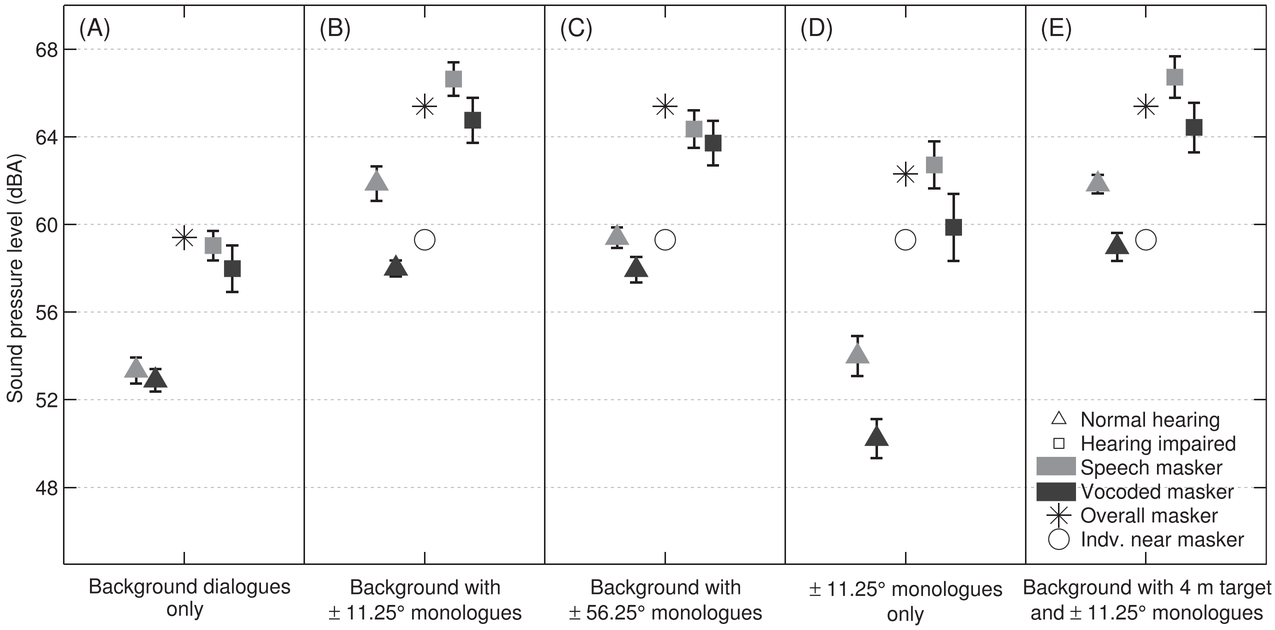

FIG. 4. As Fig. 3, but now the value shows the SPLs, in reference to free-field level, at the measured SRTs. In addition, the overall SPL of the masker and the level of each nearby masker is shown by the star and circle, respectively.

definitive conclusions. When listening to the conditions with nearby maskers, the target and maskers sounded very different and seemed rather difficult to confuse, but it was difficult to ignore the maskers. This was supported by the subjects' comments, that the conditions with nearby maskers were the most "*annoying*" and that "*it was hard to block out the person in the hanging speaker*." Further testing with a closed-set speech corpus that can effectively measure target-masker confusions (such as the CRM corpus) could be applied; however, applying a corpus that exaggerates confusions contradicts the goal of this study to better understand the effect of IM in more realistic environments.

There was almost no difference in SRTs when the target was at 4 m [Fig. 3(E)] compared to the similar configuration with the target at 2 m [Fig. 3(B)]. Since the target level is adjusted adaptively, the distance-dependent level changes are not included. Thereby, the main difference between the target signal at these two distances is a change in the direct-to-reverberant ratio (DRR). Zahorik (2002) showed that DRR just-noticeable differences (JNDs) were around 5–6 dB for NH listeners, and Akeroyd *et al.* (2007) later demonstrated that such JNDs were much higher for HI listeners. As the difference in target position applied in this study represents a doubling of distance (maximally changing the DRR by 6 dB), and furthermore, as reverberation resulting from the target is masked by the background dialogues, it is likely that the subjects did not even perceive the change in target distance.

B. The region of informational masking

As already mentioned in Sec. III B, a number of studies have shown that the influence of IM in speech-on-speech masking depends on the TMR (e.g., Westermann and Buchholz 2015b). Brungart *et al.* (2001) reported that the effect of IM is strongest at a TMR of around 0 dB, and diminishes at positive TMRs due to loudness cues. Best *et al.* (2013) have shown that in conditions where sufficient cues are available to segregate the target from the masker (such as separation by ±90°), IM is largely resolved and speech intelligibility is limited by EM. Agus *et al.* (2009) utilized a speech-in-speech-in-noise masking task to measure the effect of IM as a function of SNR and reported that IM reaches a maximum at a TMR of just under 0 dB and falls off towards lower and higher TMRs. Hence, evidence is provided for the existence of a region of IM, a concept that is further illustrated in Fig. 5. Figure 5(D) illustrates the effect of IM as a function of TMR and Figs. 5(A)–5(C) illustrate possible underlying psychometric functions, whereby each panel refers to a different measurement condition. Solid lines indicate the psychometric functions for a speech masker, the dashed lines for a corresponding noise (or noise-vocoded speech) masker, and the gray area indicates the involved IM. Figure 5(A) refers to the case in which sufficient cues are available to resolve most of the IM and speech intelligibility is mainly limited by EM. In Fig. 5(B), the TMR is equal to zero and the IM is maximal. In Fig. 5(C), loudness cues limit the effect of IM. The IM shown in Fig. 5(D) is derived by plotting the distance between the dashed and solid lines at the 50% correct value of the different psychometric functions, i.e., considering the SRT as measured throughout the present study.

Even though the "region of IM" is a straight forward concept, the detailed behavior will depend on acoustic factors (e.g., type and number of masking talkers and their spatial distribution), auditory factors (e.g., hearing loss limiting temporal, spectral, and spatial cues as well as reduced sensitivity to loudness), the applied speech material (e.g., temporal and semantic structure; context information), and maybe cognitive factors (e.g., executive function ability). Thereby, the psychometric functions are often steeper in conditions where IM is involved as compared to only EM (Brungart *et al.*, 2001; Agus *et al.*, 2009) and can be even non-monotonic (Brungart *et al.*, 2001).

Having defined such a region, two requirements can be formulated for IM to occur: (1) the target and masker must

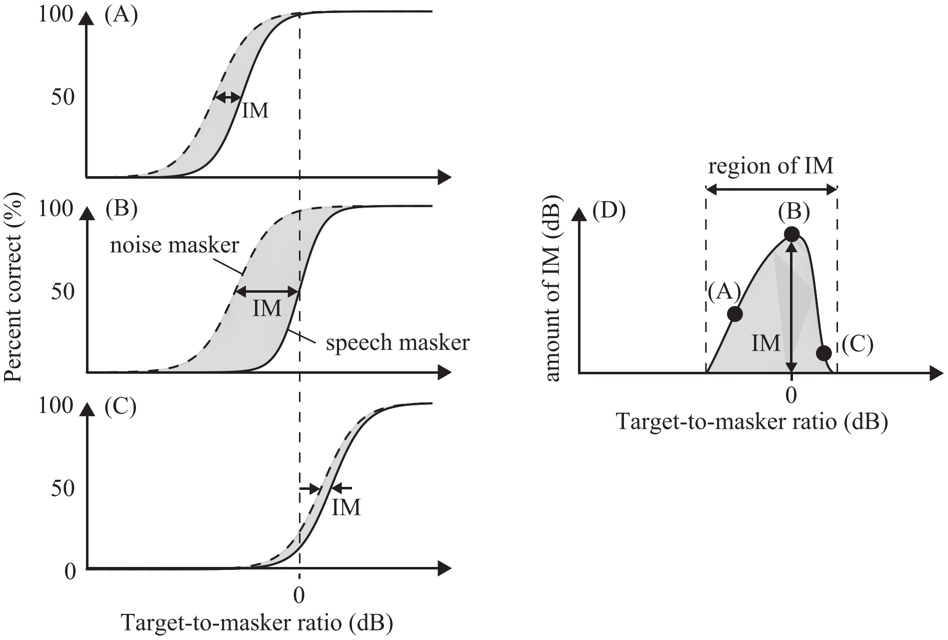

FIG. 5. Illustration of fictitious psychometric functions related to different experiment conditions and thus to different TMRs. Panel (A) presents a condition that is mainly limited by EM. Panel (C) illustrates a condition where the test is conducted at positive TMR, where loudness cues limit IM. Panel (B) refers to a condition that results in maximal IM. Gathering panel (A)–(C) as a function of TMR results in the "region of IM" in panel (D), which defines a range of TMRs that facilitate IM. Further details are given in the text.

not be fully segregated perceptually (for further discussion see Ihlefeld and Shinn-Cunningham, 2008) and (2) the TMR must fall into the region of IM. However, these rules are based on confusion-based IM, and it has been shown that perceptually segregated (non-confusable) maskers can still cause distraction-based IM (Westermann and Buchholz, 2015a). Even though, the limitations of distraction-based IM are not well known, a concept such as a region of IM might still apply. Intuitively, the main difference between the two forms of IM may be at low TMRs, i.e., when the target becomes noticeable softer than the masker. Assuming that EM is not the limiting factor, loudness cues will help to resolve confusion-based IM (i.e., listening to the softest talker) whereas distraction-based IM may be at its strongest.

In relation to the results from this study, Figs. 4(B), 4(C), and 4(E) illustrate that the nearby maskers in this study created TMRs for the NH listeners around 0 dB. As illustrated in Fig. 5(B), conditions with TMRs around 0 dB fall into the middle of the region of IM. Thus, it is not surprising that NHs listeners are especially affected by IM (Sec. III A). The reduced IM observed with the HIs listeners (Sec. III B) could, therefore, also be explained by the concept of a region of IM. Figure 4 shows that for SRTs measured with the HI listeners, the SPLs of the individual nearby maskers (shown as the ○) are substantially lower than the target SPL. Hence for the HI listeners, the TMRs were greater than zero and thereby pushed outside the "region of IM." This is as illustrated in Fig. 5(C). Finally, the observation that the effect of IM was not reduced towards very low (negative) TMRs of around −5 dB, as seen in Fig. 4(D) for NH listeners, may confirm that the nearby-maskers mainly introduced distraction-based IM and not confusion-based IM, which could be resolved by listening to the softer talker in the mixture.

Overall, to further evaluate these TMR-related effects, conditions with fixed SNRs (and thereby also fixed TMRs) across the NH and HI listeners could be included. However, since NH and HI SRTs with the vocoded masker are more than 6 dB SNR apart it might be difficult to find an appropriate SNR which would be intelligible for all HI listeners while not reaching a ceiling of 100% intelligibility for NH listeners.

C. Cognition and informational masking

Informational masking is often linked with auditory cognition (Kidd *et al.*, 2007; Helfer and Freyman, 2008; Glyde *et al.*, 2012) and some even go so far as calling it "cognitive masking." However, as far as the authors are aware no studies have successfully shown a strong relationship between cognitive measures and susceptibility to IM or release from IM. Glyde *et al.* (2012) measured SRM with varying degrees of IM as a function of hearing loss and age and applied the COGSTAT questionnaire to measure individual cognitive ability, but found no correlation between the results.

During the testing of the HI listeners two common cognitive tests were included: (1) a computerized RST (Daneman and Carpenter, 1980) and (2) a Stroop test (Golden and Freshwater, 2002). The RST assesses effective working memory capacity and the Stroop test aims to capture executive function abilities. These two tests were chosen here since they seemed to be most related to the cognitive task involved in suppressing (nearby) distracters while focusing on target speech (Sec. IV A). A Pearsons linear regression found no significant correlation between the resulting two scores measured for each participant ($r^2 = 0.1$ and $p = 0.2$).

Table I summarizes results of a linear regression analysis between the four-frequency average hearing loss (4FAHL), age, RST score, and Stroop test score on SRTs and SRT differences between speech and noise maskers, i.e., the estimated amount of IM. Here only the data for the conditions A and B (Fig. 2) are presented, but the other conditions with nearby maskers showed very similar results to condition B. Overall, there was no correlation between RST scores or age on any of the measures. There was a weakly significant correlation between the 4FAHL and the measured absolute SRTs but not for the estimated amount of IM. This correlation between HL and intelligibility in noise is in agreement with many other studies (e.g., Agus *et al.*, 2009; Glyde *et al.*, 2012). Moreover, in condition A there was a weakly significant correlation between the Stroop score and the estimated amount of IM, indicating that the susceptibility to IM might be linked to executive function ability. However, this dependency was not observed in any of the other conditions. This is surprising, because condition A was the only one that did not show any significant IM.

While these results are in no way conclusive, they suggest that mainly measures which include executive function and inhibition are of relevance to IM. This would also be in line with the observation that the IM involved in this study is most likely due to the nearby maskers distracting the subjects from attending to the target speech (Sec. IV A). To establish sufficient statistical power more subjects are needed and it could be interesting to compare the result with

TABLE I. Linear regression analysis for condition A and B for hearing loss, age, RST, and Stroop score ($n = 16$).

		Regression results	
Measure	Predictor	r^2	p
(A) Vocoded SRT	4FAHL	0.6	<0.001
	Age	0.03	0.5
	RST	0.01	0.8
	Stroop	0.00	0.9
(A) Estimated IM	4FAHL	0.00	1
	Age	0.01	0.7
	RST	0.00	0.9
	Stroop	0.3	<0.05
(B) Vocoded SRT	4FAHL	0.4	<0.001
	Age	0.09	0.3
	RST	0.01	0.8
	Stroop	0.01	0.7
(B) Estimated IM	4FAHL	0.05	0.4
	Age	0.00	0.9
	RST	0	1
	Stroop	0.03	0.5

normative data from the young NH listeners, which unfortunately was not measured here.

Other studies have employed a dual-task paradigm, where participants are scored both on the speech experiment and on a secondary task (e.g., Helfer *et al.*, 2010). The hypothesis is that while intelligibility might be comparable between conditions, effort, or cognitive load, is different and can be measured by performance on a secondary task. It could be interesting to include a secondary task and to measure its interaction with IM. This might even be different for confusion-based or distraction-based IM.

D. Perspectives

When only the background cafeteria was present (condition A), the HI listener showed a slight increase in SRTs, though not significant, when measured with speech maskers rather than vocoded maskers. In this background cafeteria condition Westermann and Buchholz (2015b) established that all TMRs at SRT are significantly above 0 dB and thus this condition should be out of the "region of IM." The fact that HI listeners still show signs of IM could indicate that they are either more susceptible to IM then NH listeners in general or their "region of IM" is extended to higher TMRs. Hence, they might simply be more distracted by the intelligible conversations around them. All in all, it could be interesting to develop a method for quantifying the "region of IM." If known, a possible application could be in an auditory model which also accounts for IM.

The conditions and locations of the nearby maskers were chosen heuristically. To minimize long-term better ear SNR effects, the nearby maskers were placed symmetrically around the listener. Also, to reduce extensive pauses only monologues were used for the nearby maskers. However, the effect of dialogues between nearby maskers could be significant. In addition, other configurations might be of interest, e.g., nearby maskers behind the listener, which might be particularly interesting when the subjects wear hearing aids, which can introduce significant front-back confusions (Best *et al.*, 2010).

V. SUMMARY AND CONCLUSION

This study investigated the ecological relevance of IM by considering a simulated cafeteria environment, thereby expanding the study of Westermann and Buchholz (2015b) by including nearby distracting maskers, HI listeners and cognitive measures. Generally the results showed:

(1) In contrast to Westermann and Buchholz (2015b), who did not find any IM in conditions where target and masker were spatially separated and different talkers, this study demonstrated that IM can occur when near masking talkers are introduced. However, the resulting IM was most likely not due to target-masker confusions as most commonly considered, but rather due to the nearby maskers distracting the listeners from attending to the target speech.

(2) As the nearby maskers were spatially separated from ±11.25° to ±56.25°, the contribution of IM decreased. This spatial release from IM demonstrates that, even if nearby maskers are present, they need to be located in a similar direction as the target to introduce substantial IM effects.

(3) The HI listeners appeared to be less susceptible to IM than the NH listeners. However, it was discussed that this was mainly a consequence of their higher SRTs, which resulted in TMRs that were above 0 dB and significantly higher than for the NH listeners. These higher TMRs shifted the HI listeners out of the "region of IM" and thereby provided loudness cues that partially resolved IM.

(4) Cognition has often been linked to IM. However, the cognitive measures applied here could not explain susceptibility to IM on an individual subject level. The RST showed no correlation with any of the data. And the Stroop test, which addresses executive function, was only weakly correlated with the amount of IM measured individually, but only in the cafeteria background without nearby maskers. Generally, more work needs to be done to tie the potential link between cognition and IM.

Overall, this study suggests that real-life listening can involve IM when nearby maskers are present. However, in most other real-life listening conditions, IM seems to be of rather low relevance.

ACKNOWLEDGMENTS

This work was supported by an International Macquarie University Research Excellence Scholarship (iMQRES) and Widex A/S. The authors would also like to thank Margot McLelland and Katrina Freeston for their help with data collection.

Agus, T. R., Akeroyd, M. a., Gatehouse, S., and Warden, D. (**2009**). "Informational masking in young and elderly listeners for speech masked by simultaneous speech and noise," J. Acoust. Soc. Am. **126**, 1926–1940.

Akeroyd, M. A., Gatehouse, S., and Blaschke, J. (**2007**). "The detection of differences in the cues to distance by elderly hearing-impaired listeners," J. Acoust. Soc. Am. **121**, 1077–1089.

Bench, J., Kowal, A., and Bamford, J. (**1979**). "The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children," Br. J. Audiol. **13**, 108–112.

Bernstein, J. G. W., and Grant, K. W. (**2009**). "Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **125**, 3358–3372.

Best, V., Kalluri, S., McLachlan, S., Valentine, S., Edwards, B., and Carlile, S. (**2010**). "A comparison of CIC and BTE hearing aids for three-dimensional localization of speech," Int. J. Audiol. **49**, 723–732.

Best, V., Thompson, E. R., Mason, C. R., and Kidd, G. (**2013**). "An Energetic Limit on Spatial Release from Masking," J. Assoc. Res. Otolaryngol. **14**, 603–610.

Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (**2000**). "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**, 1065–1066.

Bronkhorst, A. (**2000**). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," Acta Acust. Acust. **86**, 117–128.

Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (**2001**). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **110**, 2527–2538.

Byrne, D., Dillon, H., and Tran, K. (**1994**). "An international comparison of long-term average speech spectra," J. Acoust. Soc. Am. **96**, 2108–2120.

Cherry, E. (**1953**). "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**, 975–979.

Daneman, M., and Carpenter, P. (**1980**). "Individual differences in working memory and reading," J. Verbal Learn. Verbal Behav. **19**, 450–466.

Dillon, H. (**2001**). *Hearing Aids* (Thieme, New York), pp. 450–466.

Favrot, S., and Buchholz, J. (**2010**). "LoRA: A loudspeaker-based room auralization system," Acta Acust. Acust. **96**, 364–375.

Festen, J. M., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725–1736.

Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (**1999**). "The role of perceived spatial separation in the unmasking of speech," J. Acoust. Soc. Am. **106**, 3578–3588.

Glyde, H., Cameron, S., Dillon, H., Hickson, L., and Seeto, M. (**2012**). "The effects of hearing impairment and aging on spatial processing," Ear Hear. **34**, 15–28.

Golden, C. J., and Freshwater, S. M. (**2002**). *The Stroop Color and Word Test* (Sterling, Wood Dale, IL).

Helfer, K., and Freyman, R. (**2005**). "The role of visual speech cues in reducing energetic and informational masking," J. Acoust. Soc. Am. **117**, 842–849.

Helfer, K. S., Chevalier, J., and Freyman, R. L. (**2010**). "Aging, spatial cues, and single- versus dual-task performance in competing speech perception," J. Acoust. Soc. Am. **128**, 3625–3633.

Helfer, K. S., and Freyman, R. L. (**2008**). "Aging and speech-on-speech masking," Ear Hear. **29**, 87–98.

Ihlefeld, A., and Shinn-Cunningham, B. (**2008**). "Spatial release from energetic and informational masking in a selective speech identification task," J. Acoust. Soc. Am. **123**, 4369–4379.

Keidser, G., Dillon, H., Mejia, J., and Nguyen, C.-V. (**2013**). "An algorithm that administers adaptive speech-in-noise testing to a specified reliability at selectable points on the psychometric function," Int. J. Audiol. **52**, 795–800.

Kidd, G., Mason, C. R., Richards, V. M., Gallun, F. J., and Durlach, N. I. (**2007**). "Informational masking," in *Auditory Perception of Sound Sources* (Springer, New York), pp. 143–189.

Lavandier, M., and Culling, J. F. (**2007**). "Speech segregation in rooms: Effects of reverberation on both target and interferer," J. Acoust. Soc. Am. **122**, 1713–1723.

Rindel, J. (**2000**). "The use of computer modeling in room acoustics," J. Vibroeng. **3**, 219–224.

Shinn-Cunningham, B. G. (**2008**). "Object-based auditory and visual attention," Trends Cognit. Sci. **12**, 182–186.

Smeds, K., Wolters, F., and Rung, M. (**2014**). "Estimation of signal-noise ratios in realistic sound scenarios," J. Am. Acad. Audiol. **26**, 183–196.

Westermann, A., and Buchholz, J. (**2017**). "The effect of hearing loss on source-distance dependent speech intelligibility in rooms," J. Acoust. Soc. Am. **141**, EL140–EL145.

Westermann, A., and Buchholz, J. M. (**2015a**). "The effect of spatial separation in distance on the intelligibility of speech in rooms," J. Acoust. Soc. Am. **137**, 757–767.

Westermann, A., and Buchholz, J. M. (**2015b**). "The influence of informational masking in reverberant, multi-talker environments," J. Acoust. Soc. Am. **138**, 584–593.

Wood, N., and Cowan, N. (**1995**). "The cocktail party phenomenon revisited: How frequent are attention shifts to one's name in an irrelevant auditory channel?," J. Exp. Psychol. **21**, 255–260.

Zahorik, P. (**2002**). "Direct-to-reverberant energy ratio sensitivity," J. Acoust. Soc. Am. **112**, 2110–2117.