Trends in Hearing
Volume 21: 1–13
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/2331216517716844
journals.sagepub.com/home/tia
SAGE

The Influence of Noise Reduction on Speech Intelligibility, Response Times to Speech, and Perceived Listening Effort in Normal-Hearing Listeners

Maj van den Tillaart-Haverkate[1,2], Inge de Ronde-Brons[1], Wouter A. Dreschler[1], and Rolph Houben[1,2]

Abstract

Single-microphone noise reduction leads to subjective benefit, but not to objective improvements in speech intelligibility. We investigated whether response times (RTs) provide an objective measure of the benefit of noise reduction and whether the effect of noise reduction is reflected in rated listening effort. Twelve normal-hearing participants listened to digit triplets that were either unprocessed or processed with one of two noise-reduction algorithms: an ideal binary mask (IBM) and a more realistic minimum mean square error estimator (MMSE). For each of these three processing conditions, we measured (a) speech intelligibility, (b) RTs on two different tasks (identification of the last digit and arithmetic summation of the first and last digit), and (c) subjective listening effort ratings. All measurements were performed at four signal-to-noise ratios (SNRs): -5, 0, $+5$, and $+\infty$ dB. Speech intelligibility was high ($>97\%$ correct) for all conditions. A significant decrease in response time, relative to the unprocessed condition, was found for both IBM and MMSE for the arithmetic but not the identification task. Listening effort ratings were significantly lower for IBM than for MMSE and unprocessed speech in noise. We conclude that RT for an arithmetic task can provide an objective measure of the benefit of noise reduction. For young normal-hearing listeners, both ideal and realistic noise reduction can reduce RTs at SNRs where speech intelligibility is close to 100%. Ideal noise reduction can also reduce perceived listening effort.

Keywords

speech perception, noise reduction, listening effort, speech in noise, response time

Date received: 6 December 2016; revised: 9 May 2017; accepted: 23 May 2017

Introduction

Challenging listening situations require extra cognitive effort for speech understanding (Gatehouse & Gordon, 1990). In noisy environments, a listener needs to separate the speech from intrusive background noise. To ease listening in noisy environments, devices such as mobile phones, hearing aids, and cochlear implants contain noise-reduction algorithms. The majority of the currently marketed digital hearing aids contain single-microphone noise-reduction algorithms. The aim of the noise-reduction algorithm is to improve the signal-to-noise ratio (SNR) by adjusting the gain in each time-frequency region according to the estimated SNR in that specific region (Bentler & Chiou, 2006). Hearing-aid users have been shown to prefer noise reduction (Bentler, 2005; Boymans & Dreschler, 2000; Ricketts & Hornsby, 2005). Researchers have been trying to understand this preference and to develop objective measures to quantify the benefit of noise reduction.

Several studies examining the objective effect of noise reduction have focused on speech intelligibility (SI). These studies revealed that existing single-microphone noise-reduction algorithms do not improve SI in noise

[1]Clinical and Experimental Audiology, Academic Medical Center, Amsterdam, The Netherlands
[2]Pento Audiological Center, Amersfoort, The Netherlands

Corresponding author:
Maj van den Tillaart-Haverkate, Pento Audiological Center, Zangvogelweg 150, 3815 DP Amersfoort, The Netherlands.
Email: m.vandentillaart@pento.nl

and may even worsen it (Alcántara, Moore, Kühnel, & Launer, 2003; Bentler, Wu, Kettel, & Hurtig, 2008; Boymans & Dreschler, 2000; Desjardins & Doherty, 2014; Neher, Grimm, & Hohmann, 2014; Ricketts & Hornsby, 2005; Sarampalis, Kalluri, Edwards, & Hafter, 2009). Therefore, it seems that noise-reduction algorithms either do not adequately eliminate the noise that masks speech or they remove the noise resulting in speech distortion (Jorgensen & Dau, 2011; Houben, Dijkstra, & Dreschler, 2012; Lunner, Rudner, & Rönnberg, 2009).

The evaluation of noise-reduction algorithms has recently shifted from a focus on SI toward a focus on the objective assessment of cognitive measures, particularly listening effort. As yet there is no standard definition of listening effort. The British Society of Audiology recently proposed the following definition in a white paper: "the mental exertion required to attend to, and understand, an auditory message" (McGarrigle et al., 2014, p. 434). Currently, there is also no standard method of measuring listening effort. In fact, for several known methods, it is not known whether they in fact correspond to listening effort (McGarrigle et al., 2014). Listening effort has been indirectly measured using response times (RTs), for instance, in a dual-task paradigm where the secondary task is nonauditory (Desjardins & Doherty, 2014; Downs, 1982; Neher et al., 2014; Pals, Sarampalis, & Baskent, 2013; Sarampalis et al., 2009). Sarampalis et al. (2009), for example, used a dual-task paradigm to evaluate the effect of noise reduction and found significantly better recall of words and faster reaction times as a benefit of noise reduction at the lower SNRs (−2 and −6 dB SNR). They hypothesized that noise-reduction algorithms could reduce the noise in a way that is comparable with the ability of the auditory and cognitive systems in the brain to ignore the noise (Sarampalis et al., 2009). Noise reduction could support this function of the brain, not by improving the SI, but rather by relieving the cognitive load, thereby resulting in a perceived improvement in listening comfort and a decrease in listening effort (Brons, Houben, & Dreschler, 2013; Huckvale & Frasi, 2010; Lunner et al., 2009; Marzinzik, 2000; Sarampalis et al., 2009). In contrast, noise-reduction algorithms might also introduce signal distortions, leading to a reduction in perceived listening comfort or listening effort (Lunner et al., 2009; Ng, Rudner, Lunner, Pedersen, & Rönnberg, 2013).

A few researchers have attempted to measure listening effort objectively using RTs to auditory stimuli alone without the addition of a nonauditory secondary task (e.g., in a dual task; Gustafson, McCreery, Hoover, Kopun, & Stelmachowicz, 2014; Houben, van Doorn-Bierman, & Dreschler, 2013; Huckvale & Frasi, 2010; Huckvale & Leak, 2009). Houben et al. (2013) used triplets of spoken digits presented in a background noise to assess differences in listening effort between different SNRs. They found that RTs increased with decreasing SNRs, even when SI was optimal. The authors hypothesized that the addition of background noise increased the listening effort required to maintain the same intelligibility scores.

In summary, there is an ongoing search for objective measures that quantify the effects of noise-reduction algorithms on listening effort, even or especially when SI is unaffected. In the present study, we investigated whether measuring RTs to digits triplets in noise with the method introduced by Houben et al. (2013) is appropriate for this purpose. This method, which makes use of existing auditory-only stimuli, has been shown to be sensitive to the effects of noise, even when the noise does not affect SI.

In the study by Houben et al. (2013), normal-hearing participants performed two tasks: identification (ID) of the last digit in each triplet (ID task) and addition of the first and last digit (arithmetic [AR] task). Because the effect of noise reduction on these two specific tasks is as yet unknown, we decided to include both tasks in our experiments. We also included an SI test to verify whether noise-reduction processing affected SI. Finally, we included a subjective rating scale to assess perceived listening effort. This rating scale has previously been used by several authors for this purpose (Brons, Houben, & Dreschler, 2012; Luts et al., 2010; Marzinzik, 2000).

The aim of this study was to assess whether the benefit of noise-reduction algorithms is reflected in a change in RTs during tests of speech understanding in noise as well as in perceived listening effort, at SNRs where SI is optimal. In other words, can RT be used to assess listening effort when listening to unprocessed speech in noise and speech in noise processed with a noise-reduction algorithm, when this speech in noise is highly intelligible?

We evaluated the influence of noise-reduction algorithms on four different outcomes: SI, RT to speech stimuli in two different tasks, and subjective listening effort. We compared unprocessed stimuli with stimuli that were processed with one of two noise-reduction algorithms: an ideal binary mask (IBM; Wang, 2005) and a minimum mean square error estimator (MMSE; Ephraim & Malah, 1984). The IBM algorithm receives speech and noise separately and thus has a priori knowledge on the real input SNR. Although this unrealistic algorithm cannot be used in hearing aids, it provides a useful tool to investigate the maximum achievable effect of noise reduction. IBM can improve SI in noise (Brons et al., 2012; Wang, Kjems, Pedersen, Boldt, & Lunner, 2009), unlike realistic single-microphone noise reduction. The second noise-reduction algorithm, MMSE, is a realistic algorithm that estimates the SNR from the mixed input signal. This algorithm is not expected to improve SI, but may affect listening effort.

Previous studies have shown that noise-reduction processing depends on the input SNR (Bentler & Chiou, 2006). Consequently, the perceptual effects of noise reduction may depend on SNR (Brons et al., 2013). We therefore included four SNR levels (-5, 0, +5, and $+\infty$ dB) at which all speech is expected to be fully intelligible, even for the unprocessed condition. We selected these levels because they were expected to lead to optimal SI and because they are relevant for many daily listening situations (Luts et al., 2010; Olsen, 1998). At the selected SNRs, the noise reduction does not improve SI because intelligibility is already optimal. In addition, we verified the effect of IBM and MMSE on SI at a suboptimal SNR of -10 dB in a separate experiment. This extension to an SNR with an expected suboptimal (<100%) SI was included because it has been shown that IBM can improve SI, in contrast to more realistic noise-reduction methods. We wanted to verify whether this effect in intelligibility indeed occurs and to see how reaction times are influenced at an SNR level of -10 dB.

Methods

Participants

Twelve normal-hearing listeners participated in the current study that was approved by the Medical Ethical Committee of Academic Medical Centre Amsterdam on 29 May 2013 (reference number NL44143.018.13). Participants were recruited via a poster on a billboard at the Medical Faculty of the University of Amsterdam. All 12 participants (2 male and 10 female) were normal-hearing adults and native Dutch speakers. Pure tone thresholds were equal to or better than 20 dB Hearing Level at octave frequencies 250 to 8000 Hz. Listeners ranged in age from 19 to 34 years, with a mean of 24 years ($SD = 4.15$).

The required sample size was estimated using a power analysis based on the study by Houben et al. (2013). Our aim was to identify a within-subject effect size similar to the smallest effect size resulting in a change of 5 dB SNR in their study. Houben et al. found the smallest effect in the ID task, with a change in SNR from -1 to $+4$ dB; the difference in RT between these SNRs was 0.028 s, with a between-participant standard deviation of 0.023 s (for $n = 12$). Power analysis revealed that for a significance level of .05 and a power of 80%, a sample size of 12 was required (two-tailed paired t test). For this reason and due to the need to balance 12 conditions (four SNRs and three processing conditions), we included 12 normal-hearing listeners in this study.

Stimuli

We used spoken digits from the newest version of the Dutch digit triplet test (Smits, Goverts, & Festen, 2013). The triplets in this test contain digits from zero to nine spoken by a male speaker, and each triplet contains a unique combination of three different digits. The digit triplets include silent intervals of 150 ms between the digits and do not include natural coarticulation or prosody (Smits et al., 2013). We selected all triplets that allowed the AR task to be conducted with a single key press (i.e., sum of first and third digit < 10; resulting in 60 of the 120 available triplets).

The stationary long-term average speech spectrum noise of the digit triplet test (Smits et al., 2013) was used to obtain four different SNRs: -5, 0, +5, and $+\infty$ dB (i.e., in quiet, thus no noise added). The level of the noise was kept constant at 65 dB(A) for all SNRs except at $+\infty$ dB, where the average level of the speech was 65 dB(A). The noise had a symmetrical onset- and offset-ramp (Hann-window, 200 ms-ramp), starting 250 ms before the first digit and ending 200 ms after the final digit.

Noise-Reduction Algorithms

For each SNR, the triplets in noise were processed with two noise-reduction algorithms: the IBM (Wang, 2005) and an MMSE (Ephraim & Malah, 1984; Huckvale & Frasi, 2010).

IBM can be considered an ideal noise-reduction algorithm because it receives noise and speech as separate inputs and does not require estimation of either noise or speech from the mixed signal. The SNR of the signal in a specific time-frequency unit determines whether this unit is preserved (if the SNR is above threshold) or eliminated (if the SNR is below threshold). The MATLAB implementation of the IBM algorithm used in this study was provided by Loizou and was previously used in several studies (Brons et al., 2012; Hu & Loizou, 2008; Li & Loizou, 2008). We used the same fixed threshold of 0 dB SNR as was done before. In Figure 1 (right-hand panels), the physical effect of IBM on a digit triplet in noise is presented at four SNRs. Both the effects on the time signal and the attenuation as a function of time and frequency are shown. The time domain signals show the input signals in dark gray and the signals after IBM processing in light gray. The spectrogram-like plots show the attenuation as a function of time and frequency, with the time-frequency units that were removed (infinite attenuation) in black and those that were retained in white. For lower input SNRs, the number of noise-dominated time-frequency units increases, and thus more signal parts were removed by the IBM (i.e., number of white pixels decreases).

The realistic noise-reduction algorithm, MMSE, assumes that both the speech and the noise in the

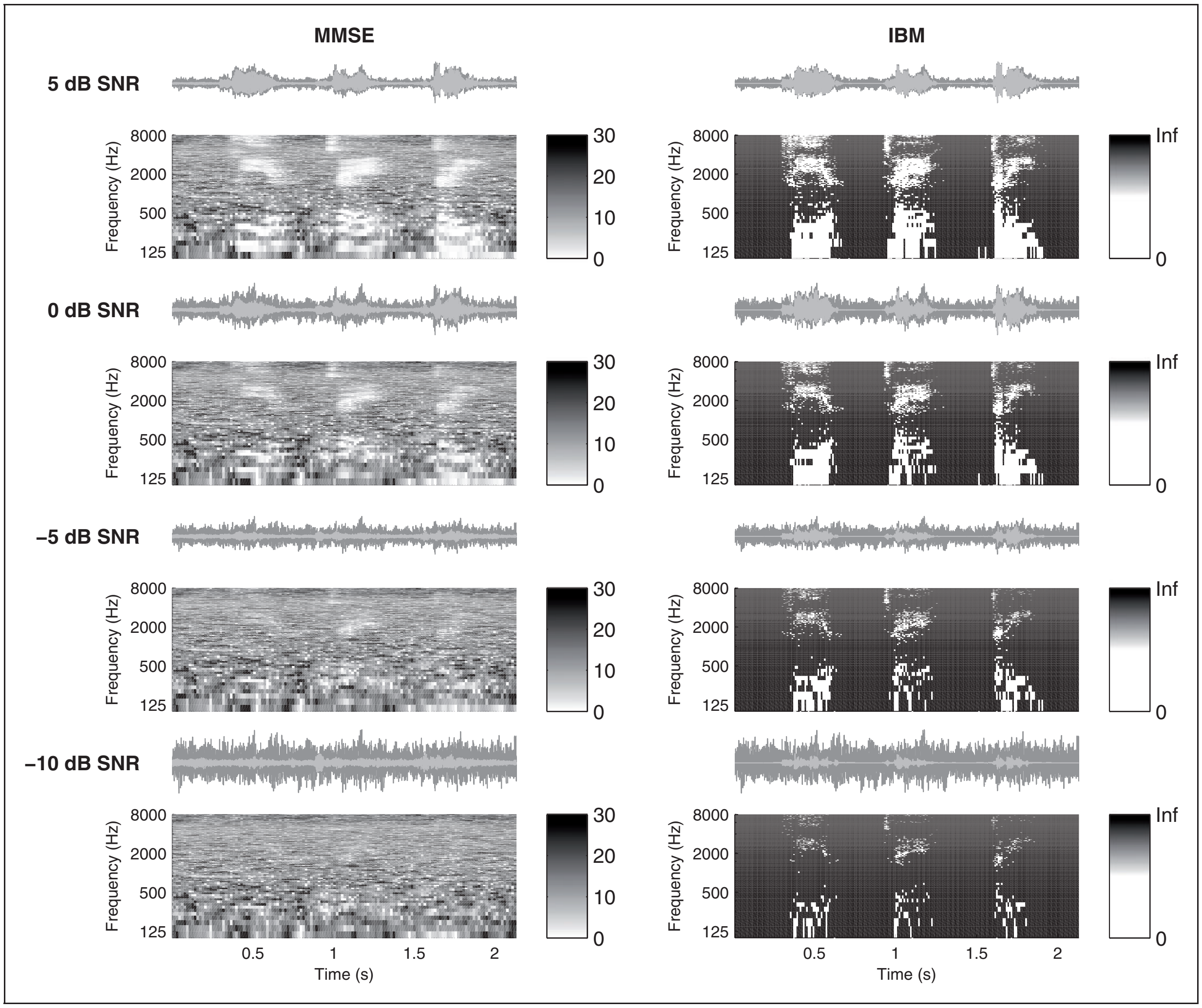

Figure 1. Acoustical effects of the two noise-reduction algorithms, IBM (right-hand panel) and MMSE (left-hand panel), on a digit triplet in noise at four SNRs. In the time domain signals, dark gray indicates the input signal, and light gray indicates the time signal after noise-reduction processing. In addition, the corresponding changes in gain as a function of time and frequency are shown in the spectrogram-like plots. For IBM, black pixels indicate noise-dominated time-frequency units that were removed (infinite attenuation), and white pixels are speech-dominated units that were retained. For MMSE, the attenuation is color coded from white (0 dB) to black (30 dB).
IBM = ideal binary mask; MMSE = minimum mean square error estimator; SNR = signal-to-noise ratio.

combined signal are independent Gaussian samples. It derives the MMSE of the clean speech spectrum, based on the noise amplitude, and attenuates the gain primarily at frequencies where the SNR is estimated to be the poorest (Sarampalis et al., 2009). We applied the MMSE algorithm implemented by Brookes (1997), previously used by Huckvale and Frasi (2010). Figure 1 (left-hand panels) shows the effects of MMSE on a digit triplet in noise and the resulting attenuation in the time-frequency domain, at four SNRs. White time-frequency units were retained, whereas gray time-frequency units were attenuated, with stronger attenuation for darker pixels. At +5 dB input SNR, the noise-reduction algorithm was able to recognize the speech from the noise and to retain speech-dominated time-frequency units (i.e., the pattern of white pixels is comparable with that of the IBM). For lower SNRs, however, it gradually becomes more difficult to separate the speech from the noise, and most of the signal was attenuated by the MMSE algorithm.

The three processing conditions (unprocessed, IBM, and MMSE) and four SNRs (−5, 0, +5, and +∞ dB) resulted in a total of 12 conditions.

Equipment

Participants performed the experiment in a soundproof testing room with stimuli presented diotically through headphones (Sennheiser HDA200, Sennheiser, Wedemark, Germany), which were connected to a computer (Dell precision T3500) via an audio interface (RME Fireface 800). The presentation of the stimuli was controlled by software in MATLAB (version 7.14.0.739, The Mathworks, R2012a), and response data were collected using MATLAB's Psychophysics Toolbox extensions (Version 3; Brainard, 1997; Brookes, 1997). Participants responded by using the numerical keypad of a keyboard that was placed close to the hand with which the participants responded. The experimental set up was calibrated using an artificial ear (B&K 4153, Brüel & Kjaer, Naerum, Denmark) equipped with a flat-plate adaptor, connected to a sound level meter (B&K 2260 Investigator, Brüel & Kjaer, Naerum, Denmark).

Procedure

We included two RT tasks: an ID task and an AR task. In the ID task, the participants had to identify the final digit of each triplet, and in the AR task, the participants had to add the initial and final digit of a triplet. Participants were instructed to respond as fast as possible using the numeric keypad.

The experiment was divided into two visits per subject. The first visit started with pure tone audiometry. Each visit contained an RT task (either the ID or the AR task), an SI task, and a listening effort rating (LEr) task. The two RT tasks were divided among the two visits; half of the participants performed the ID task in the first visit and the AR task during the second visit, while for the other half of the participants, the order was reversed. The duration of a visit was approximately 1.5 hr, including short breaks. All participants signed an informed consent form prior to starting the experiment. We collected data from a total of 1,920 triplet presentations per subject, distributed over the ID, AR, SI, and LEr task.

Each visit started with a practice session, in which a total of 20 triplets were presented divided among four different conditions containing all three processing conditions and all four SNRs. This allowed participants to practice with the ID or AR task and to familiarize themselves with the equipment and the processing conditions. In addition, the practice session was used to verify that the participants fully understood the instructions.

For both the ID and AR tasks, the 60 triplets were split into two sets. These sets and the two tasks were balanced across participants, and across the two visits, to avoid order effects. During the course of the two visits, all participants performed both tasks using both sets in all conditions. Conditions were balanced across participants based on a Latin square design (Wagenaar, 1969) to minimize possible training effects on the group data. Triplets were presented in random order, in blocks of 30 triplets per condition. Participants started the RT (either ID or AR) task in 12 blocks (i.e., conditions). These blocks consisted of the first set of 30 triplets. After a short break, the RT task was repeated using the 12 blocks in the same order, but with the other set of 30 triplets. The RT task was followed by a pause.

After the pause, participants performed the SI and LEr tasks combined. The SI and LEr tasks were done per condition: first, for Condition 1, SI followed by LEr, then for Condition 2, SI followed by LEr, and so forth. The instruction for the SI task was to enter all three stimulus digits, without any time constraint. The 12 conditions were balanced across participants, presented once per visit, and each condition contained 20 randomly presented triplets. During the SI task, each condition was followed by the LEr task, in which participants were asked to rate their perceived listening effort for the preceding block of 20 triplets. The following question had to be answered: "How much effort did it take to understand the last 20 triplets?" The participants scored their perceived listening effort on a 9-point rating scale (based on ITU 1996; Brons et al., 2012; Luts et al., 2010; Marzinzik, 2000), ranging from *no effort* (1) to *extremely high effort* (9). This combined measurement of the SI and LEr tasks was performed during each visit. The order of tasks in the two visits is summarized in Table 1.

Table 1. Summary of the Content of the Two Visits.

	Task
Visit 1	
1	Pure Tone Audiometry
2	Practice RT task
3	RT task (either ID or AR)
	Break
4	Repeat RT task
	Break
5	SI and LEr task combined
Visit 2	
1	Practice RT task
2	RT task (either ID or AR task, whichever was not performed in Visit 1)
	Break
3	Repeat RT task
	Break
4	SI and LEr task combined

Note. RT = response time; ID = identification; AR = arithmetic; SI = speech intelligibility; LEr = listening effort rating.

Additional SNR

At SNRs lower than the SNRs included in our design, intelligibility is expected to worsen for the unprocessed and MMSE conditions, but not for IBM. IBM can improve intelligibility at low SNRs because of its a priori knowledge on the signal (Wang, 2005). To verify whether this improvement in intelligibility indeed occurs, and to see how reaction times are influenced in this situation, we included an additional SNR level of −10 dB in an extra session. The same group of participants performed all four tasks at two SNR levels: at −10 and +∞ dB. The inclusion of the +∞ dB SNR condition allowed us to compare this supplementary data to the main results. However, we did not include these data in the statistical analysis of the main results for two reasons: (a) they were gathered in a separate experiment and were thus not included in the same balanced design as the other conditions and (b) they do not fulfill the requirement of maximum SI.

Data Analysis

The data analysis was performed using MATLAB (version 7.14.0.739, The Mathworks, R2012a). A response for the SI measurement was considered correct only if all three digits of a triplet were repeated accurately. The percentage of appropriate responses was calculated per condition for each subject. To satisfy the homoscedasticity criterion for statistical analysis, this percentage was transformed to rationalized arcsine units (Studebaker, 1985), and the LEr scores were arcsine transformed.

For the ID and AR task, the RTs were the main outcome variables. The time between the end of the waveform of the third digit and the subsequent key press by the participant was defined as the RT. Data of the practice sessions were discarded. Only correct responses were included in the analyses. Unrealistically long RTs can be discarded by selecting a fixed cut-off value. This application of a fixed cut-off value to remove spuriously long RTs has been shown to be suitable for RT data (Ratcliff, 1993; Whelan, 2008). For each task, 1.25% of all RTs was removed with cut-off values for the ID and AR task of 0.80 s and 3.34 s, respectively. The value of 1.25% was chosen to be the same as that used by Houben et al. (2013).

To examine the effect of SNR and processing condition, we calculated changes in response time (ΔRT) by subtracting the RT at +∞ dB SNR from the RT at all other SNRs per processing condition, for each task-subject-triplet combination. As a result, by definition, the condition at +∞ dB SNR has both ΔRT values and confidence intervals of zero. Thus, ΔRT reflects the effect of added noise and signal processing on the RT. Positive ΔRT values denote prolonged RTs compared with the condition without noise (i.e., at +∞ dB SNR).

We used mixed-effect models as suggested by Baayen and Milin (2010) to analyze RT data. Mixed-effect models offer better modeling of the dependency on participant and triplets, by using multiple random effects. In this way, the mixed model can be regarded as a repeated measures model that contains two (subject, triplet) rather than one random effect (subject). Thus, a mixed-effect model has the advantage that differences between both participants and triplets can be accurately modeled (Houben et al., 2013). This is important because the stimuli were originally developed to be equally intelligible (Smits et al., 2004) rather than having equal RT (Houben et al., 2013). Because a mixed model can also be used to analyze data with a single random effect (repeated measures design), we used a mixed-model analysis of variance (ANOVA) on the intelligibility and LEr data as well (with only a single random effect). The significance level for all statistical tests was set at .05. Post hoc pairwise comparisons for multiple testing were performed with Bonferroni correction, including calculation of the effect size *r*.

Results

Speech Intelligibility

The intelligibility scores for all conditions within the SNR region of interest (i.e., −5 to +∞ dB), averaged over all 12 participants, were equal to or above 97.7% correct, as shown in Figure 2. We performed a mixed-model ANOVA on the rationalized arcsine unit-transformed intelligibility scores. Processing condition,

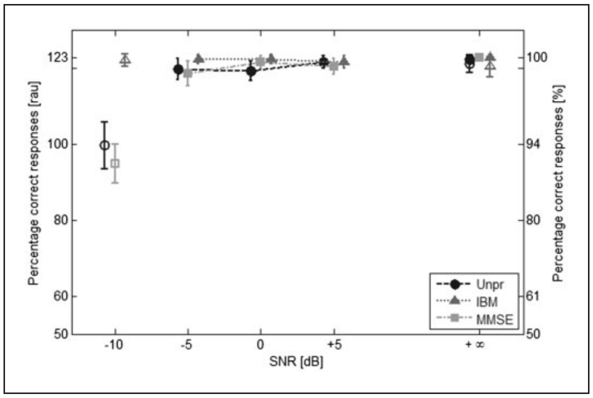

Figure 2. Speech intelligibility results averaged over all subjects. Closed markers present the performance on the speech intelligibility task for the SNR region of interest, in RAU (left ordinate) and percentage (right ordinate) correct responses. Error bars denote 95% confidence intervals between subjects. Open markers depict the result of the additional SNR.

IBM = ideal binary mask; MMSE = minimum mean square error estimator; SNR = signal-to-noise ratio; Unpr = unprocessed; RAU = rationalized arcsine unit.

SNR, and the interaction of these two variables were considered fixed factors, and participant a random factor. The model demonstrated that the effect of processing condition was significant, $F(2, 121) = 3.7$, $p < .05$. Pairwise comparisons with Bonferroni correction ($n = 3$), considering the data of all SNRs together per processing condition, showed a significantly better performance of 0.68% on SI of IBM versus unprocessed stimuli, $t(47) = -2.56$, $p < .05$, $r = .35$. ANOVA showed that the main effect of SNR was also significant, ($F(3, 121) = 4.17$, $p < .01$. The post hoc Bonferroni corrected ($n = 6$) pairwise comparisons, considering data of all processing conditions together per SNR, resulted in a significantly lower score at both −5 and 0 dB SNR compared with +∞ dB SNR—$t(35) = -3.36$, $p < .05$, $r = .49$, difference is 1.25%; $t(35) = 2.26$, $p < .05$, $r = .36$, difference is 0.63%, respectively. The interaction of processing condition and SNR was not significant, $F(6, 121) = 1.67$, $p = .13$.

The extra data points at −10 dB SNR in Figure 2 show that SI for MMSE and unprocessed stimuli was indeed considerably lower at poorer SNRs, while this reduction in performance was not present for the stimuli processed with IBM.

ID Task

To test whether the participants were able to perform the tasks reliably, in spite of instructions to do it quickly, we assessed the percentage of correct responses of the ID task first. The percent correct responses averaged over all conditions was 98.7%. Percent correct data were analyzed using another ANOVA, with processing condition and SNR as fixed factors and participant as random factor. No significant effects of processing condition, $F(2, 121) = 2.73$, $p = .07$, and SNR, $F(3, 121) = 1.34$, $p = .26$, were found. For analysis of the ΔRTs, we used only the correct responses; considering all data did not change the statistical outcomes.

On average, the participants responded more than twice as fast in the ID task (0.31 s) compared with the AR task (0.82 s). For the presentation in the figures and for statistical analysis, RT data were calculated relative to the condition at +∞ dB SNR per processing condition, SNR, subject, and triplet, and subsequently averaged across participants. This relative RT is denoted by ΔRT.

Figure 3 shows the result of the ΔRTs for the ID task. The influence of processing condition and SNR on ΔRT in the ID task was analyzed with a mixed-model ANOVA. Processing condition, SNR, and their interaction were the fixed variables, whereas participant and triplet were entered as random factors. Thus, prior to analysis, RT data were not averaged across participants or triplets. Because ΔRT at +∞ dB SNR is zero by definition, this SNR was not included in the ANOVA. SNR thus had three levels: −5, 0, +5 dB. Post hoc comparisons can easily be done relative to a fixed value (e.g., zero), and thus +∞ dB SNR was included in the post hoc analysis. The analysis showed a significant effect of SNR, $F(2, 6124) = 8.15$, $p < .001$, whereas the effect of processing condition, $F(2, 6124) = 0.75$, $p = .47$, and the interaction of processing condition and SNR, $F(6, 6124) = 0.54$, $p = .71$, were not significant. We also

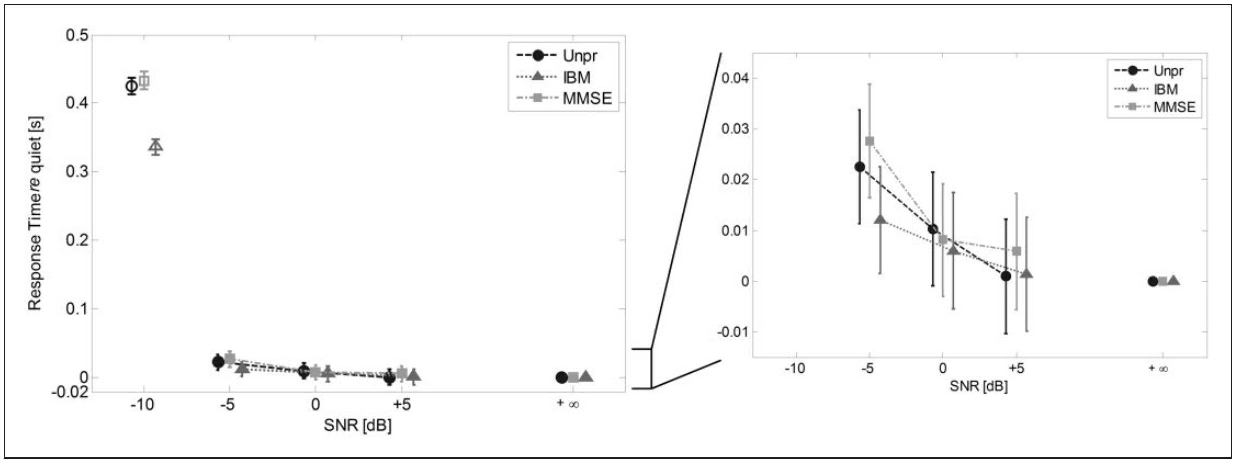

Figure 3. Response time results for the identification task averaged across all subjects. The panel on the left presents the data for all measured SNRs, while the one on the right zooms in on the region of interest. Closed markers show ΔRT (i.e., the response times to the identification task, relative to +∞ dB SNR) for the SNR region of interest and open markers for −10 dB. Error bars denote 95% confidence intervals between subjects.

IBM = ideal binary mask; MMSE = minimum mean square error estimator; SNR = signal-to-noise ratio; Unpr = unprocessed; ΔRT = changes in response time.

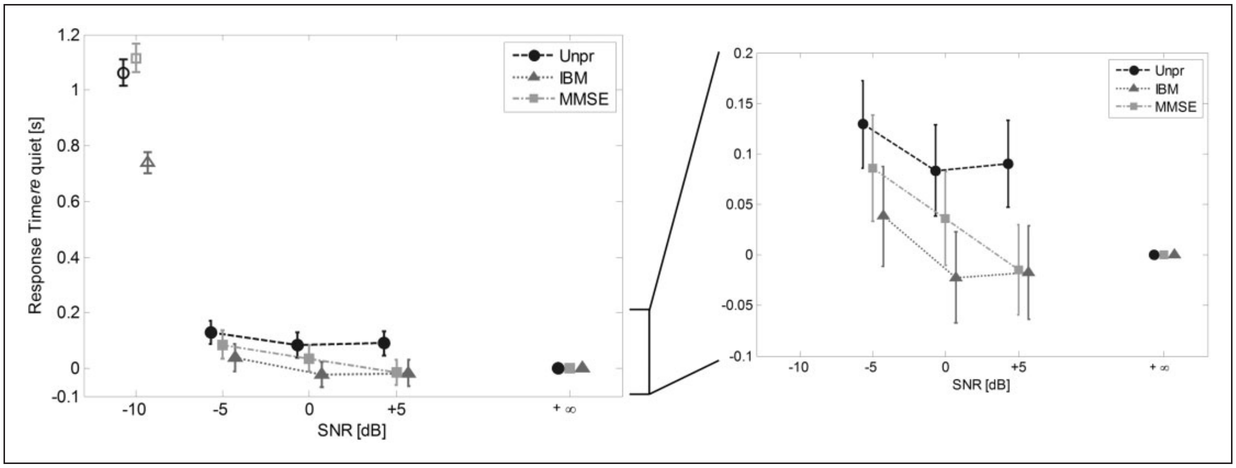

Figure 4. Response time results for the arithmetic task averaged across all subjects. The left panel shows the data for all measured SNRs, while the right panel zooms in on the region of interest. Closed markers show ΔRT (i.e., the response times to the identification task, relative to $+\infty$ dB SNR) for the region of interest and open markers for -10 dB SNR. Error bars denote 95% confidence intervals between subjects.
IBM = ideal binary mask; MMSE = minimum mean square error estimator; SNR = signal-to-noise ratio; Unpr = unprocessed; ΔRT = changes in response time.

conducted pairwise comparisons with Bonferroni correction ($n = 6$) on all combinations of SNRs. This analysis revealed that the ΔRTs at the lowest SNR (-5 dB), averaged over processing conditions, were significantly longer than the ΔRTs at all other SNRs (0 dB, $t(1976) = 2.95$, $p < .005$, $r = .07$; $+5$ dB, $t(1981) = 3.49$, $p < .001$, $r = .08$; $+\infty$ dB, $t(2064) = 6.53$, $p < .001$, $r = .14$).

At the supplementary SNR of -10 dB, the ΔRTs were on average 390 ms longer than in the ceiling region (SNRs $\geqslant -5$ dB). Moreover, IBM significantly reduced the ΔRT compared with the unprocessed stimuli by 89 ms, $t(1270) = 11.53$, $p < .001$, $r = .31$, and compared with stimuli processed with MMSE by 97 ms, $t(1251) = 11.45$, $p < .001$, $r = .31$. The ΔRT values for MMSE did not differ significantly from the unprocessed stimuli, $t(1207) = 1.42$, $p = .16$, $r = .04$.

AR Task

Averaged over all conditions, the percent correct responses for the AR task was 96.8%. Percent correct data were analyzed using another ANOVA, with processing condition and SNR as fixed factors and participant as random factor. This analysis showed no significant effects of either processing condition, $F(2, 121) = 0.05$, $p = .96$, or SNR, $F(3, 121) = 0.69$, $p = .56$.

Figure 4 shows the results of the ΔRTs for the AR task. We analyzed the AR data with a mixed-model ANOVA analogous to the results of the ID task. This analysis revealed a significant effect of processing condition, $F(2, 5884) = 13.74$, $p < .001$, and SNR, $F(2, 5884) = 6.87$, $p < .001$. The interaction between processing condition and SNR was not significant, $F(4, 5884) = 0.51$, $p = .73$.

Post hoc pairwise comparisons between the three processing conditions reveal that the mean ΔRT of the unprocessed condition was significantly (125 ms) longer than the ΔRTs of IBM, $t(1838) = 5.09$, $p < .001$, $r = .12$, and 28 ms longer than the ΔRTs of MMSE (245 ms), $t(1832) = 3.53$, $p < .001$, $r = .08$. Comparison of the different SNRs primarily demonstrated that on average, the ΔRT at the lowest SNR (-5 dB) was significantly longer than the ΔRTs at $+ 5$ dB SNR, $t(1836) = 4.23$, $p < .001$, $r = .10$, and $+\infty$ dB SNR, $t(1980) = 5.96$, $p < .001$, $r = .13$.

Figure 4 shows that the ΔRTs at the supplementary point of -10 dB SNR are on average 939 ms longer than at the other SNRs. At -10 dB SNR, IBM again had a significantly reduced ΔRT compared with unprocessed stimuli with 324 ms, $t(1409) = 10.11$, $p < .001$, $r = .26$, and compared with MMSE with 378 ms, $t(1389) = 11.47$, $p < .001$, $r = .29$. The ΔRT values for MMSE and unprocessed stimuli were not significantly different from each other, $t(1387) = 1.68$, $p = .09$, $r = .05$.

Subjective Listening Effort

The results of LErs are presented in Figure 5. To satisfy the ANOVA criteria, we transformed LErs with an arcsine transformation. A mixed-model ANOVA was accomplished with SNR, processing condition, and their interaction as fixed factors and participant as a random factor. We found significant effects of processing condition, $F(2, 265) = 23.27$, $p < .001$; SNR, $F(3, 265) = 149.89$, $p < .001$; and the interaction

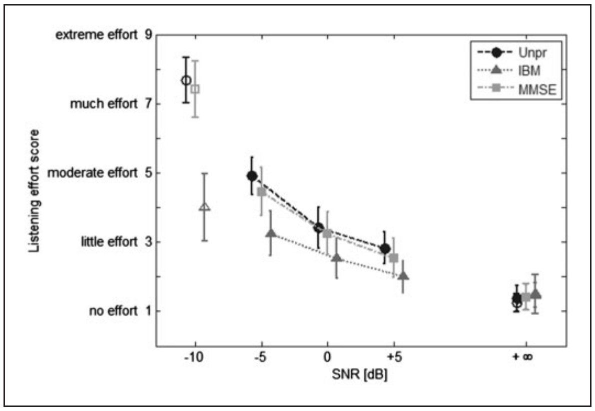

Figure 5. Listening effort ratings averaged across all subjects. Closed markers show the perceived listening effort for the SNR region of interest, open markers for −10 dB SNR. Error bars denote 95% confidence intervals between subjects.
IBM = ideal binary mask; MMSE = minimum mean square error estimator; SNR = signal-to-noise ratio; Unpr = unprocessed.

between processing condition and SNR, $F(6, 265) = 3.94$, $p < .001$.

Table 2 shows the significant pairwise comparisons of processing condition and SNR after Bonferroni correction ($n = 3$ for processing conditions, $n = 6$ for SNRs, and $n = 27$ for their interactions). These results show that unprocessed and MMSE-processed digit triplets were rated as requiring significantly more effort than the triplets processed with IBM. Zooming in on SNR, this significant result was present at−5 dB, but at +5 dB only unprocessed stimuli were rated significantly more effortful (score: 2.8) than IBM (score: 2.0). At the additional SNR of −10 dB, the effect of processing condition seems to persist; participants rated less effort for IBM compared with unprocessed stimuli with 3.6 points, $t(47) = 5.20$, $p < .001$, $r = .60$, and compared with MMSE with 3.4 points, $t(47) = 5.25$, $p < .001$, $r = .61$.

The main effect of SNR indicates that the LEr decreased significantly over all SNRs. Considering the combinations of SNR per processing condition (i.e., the interaction), this decrease held for both the unprocessed and MMSE processed conditions, except for the combination of 0 dB versus +5 dB SNR. For IBM, only three combinations of SNRs showed significantly different ratings, as presented in Table 2.

Table 2. Significant Pairwise Comparisons of Listening Effort Ratings.

	df	*t*	*p*	*r*
Processing conditions				
Main effect				
IBM < Unpr	95	6.69	<.001	.75
IBM < MMSE	95	5.27	<.001	.48
Interactions				
SNR = −5 dB				
IBM < Unpr	23	6.63	<.001	.81
IBM < MMSE	23	5.22	<.001	.74
SNR = + 5 dB				
IBM < Unpr	23	4.77	<.001	.71
SNRs				
Main effect				
−5 dB > 0 dB	71	8.00	<.001	.69
−5 dB >+5 dB	71	11.45	<.001	.81
−5 dB >+∞ dB	71	17.42	<.001	.90
0 dB >+5 dB	71	4.33	<.001	.46
0 dB > +∞ dB	71	11.65	<.001	.81
+5 dB > +∞ dB	71	8.78	<.001	.72
Interactions				
Processing condition = Unpr				
−5 dB > 0 dB	23	7.53	<.001	.84
−5 dB > +5 dB	23	9.68	<.001	.90
−5 dB > +∞ dB	23	15.95	<.001	.96
0 dB > +∞ dB	23	11.99	<.001	.93
+5 dB > +∞ dB	23	7.74	<.001	.85
Processing condition = MMSE				
−5 dB > 0 dB	23	4.46	<.001	.68
−5 dB > +5 dB	23	6.23	<.001	.79
−5 dB > +∞ dB	23	11.66	<.001	.93
0 dB > +∞ dB	23	8.74	<.001	.88
+5 dB > +∞ dB	23	5.19	<.001	.73
Processing condition = IBM				
−5 dB > +5 dB	23	5.10	<.001	.73
−5 dB > +∞ dB	23	7.08	<.001	.83
0 dB > +∞ dB	23	3.61	<.01	.60

Note. The *p* values are Bonferroni corrected. Insignificant contrasts were omitted from the table. Overall, three comparisons of processing conditions, six comparisons of SNRs, and 27 comparisons of interactions were made. SNR = signal-to-noise ratio; IBM = ideal binary mask; MMSE = minimum mean square error estimator; Unpr = unprocessed.

Discussion

Speech Intelligibility

We measured SI to verify whether noise reduction affected intelligibility at the SNRs in our region of interest (−5 dB to +∞ dB). As expected, all intelligibility scores at −5 dB SNR and higher were close to 100% (⩾ 97.7%). Nevertheless, statistical analysis revealed small but significant differences between processing conditions and between SNRs. SI was significantly higher in the IBM condition (99.8%) than in the MMSE (99.1%) and unprocessed (99.1%) conditions. IBM is known to improve the speech reception threshold (i.e., the SNR corresponding to 50% correct) by approximately 13 dB for speech-shaped noise (Brons et al., 2012; Wang et al., 2009). Thus, in our study, for decreasing SNRs, the intelligibility scores for IBM will remain close to 100%, where scores for unprocessed speech in noise start to decrease. Indeed, the results at the additional reference point of −10 dB SNR show that SI scores with IBM processing were still close to 100%, whereas performance for the MMSE and unprocessed condition had decreased to approximately 94%. SI was significantly higher for speech at +∞ dB SNR (99.9%) than for the two lowest SNRs (−5 and 0 dB, 98.7% and 99.3%, respectively) in the region of interest. Differences between successive SNRs were nonsignificant. Hence, in line with our goal, the included SNRs represent the ceiling levels for SI for the speech materials used.

ID and AR Task

For objective evaluation of the influence of noise reduction, we measured RTs to speech in noise for the ID and AR tasks.

We found no significant effect of processing condition on the ΔRTs for the ID task. In the AR task, however, both noise-reduction algorithms significantly reduced the ΔRTs in comparison with the unprocessed condition. This indicates a benefit of noise reduction at SNRs where there is no loss of intelligibility. Previous studies did not find a significant effect of noise reduction on RTs to speech stimuli (Huckvale & Frasi, 2010; Huckvale & Leak, 2009) using a dual task (Sarampalis et al., 2009) at SNRs where intelligibility reached ceiling level. Note that at the same SNRs, our results showed either no improvement (MMSE) or only a small improvement (IBM) in intelligibility. Gustafson et al. (2014) found improved verbal RTs to nonwords due to digital noise reduction of two hearing aids in children, but at SNRs with an average SI of about 62%. In the study of Huckvale and Leak (2009) and Huckvale and Frasi (2010), RTs to digits in noise did not change due to processing with MMSE. The RT method used in those studies is comparable with the ID task used in this study, in which we also did not find any significant effect of noise reduction. The probable reason for the higher sensitivity of the AR task will be discussed in more detail at the end of this section.

In both the ID and AR task, at the supplementary SNR of −10 dB, MMSE did not significantly change the ΔRT compared with unprocessed stimuli, whereas IBM did significantly reduce the ΔRT compared with both the unprocessed and MMSE condition. These results indicate an interaction between processing condition and SNR, which was not found at higher SNRs. For the AR task, this finding suggests that realistic noise-reduction algorithms provide a benefit, that is, less listening effort, at high SNRs, and that there is no benefit at poorer SNRs ($\leqslant -10$ dB SNR). This is in line with the visual impression from Figure 1, where at higher input SNR, the MMSE was able to preserve parts of the speech signal while attenuating the noise, but at low input SNR, the speech could no longer be separated from the noise. The IBM, on the contrary, used information on the clean speech and noise input signal and was therefore able to preserve parts of the speech better, even at −10 dB SNR.

Noise reduction is known to involve a trade-off between speech distortion and noise removal (Brons et al., 2013; Houben et al., 2012; Lunner et al., 2009; Luts et al., 2010). For IBM, the signal distortions are mainly caused by quick changes in gain (Wang, 2008), and at lower SNR also by removing parts of the speech signal (see Figure 1). MMSE causes distortions because of errors in separating the speech and noise from the mixed input signal. These distortions may increase the cognitive load (Lunner et al., 2009), resulting in longer ΔRTs in our results. Apparently, for IBM, the positive effect of reducing noise dominates over the negative effect of distortions on all input SNRs in our experiments, whereas for MMSE, the negative effect of distortions is stronger at lower SNR. However, for IBM, the negative effect of distortions was also visible in the ΔRT, which increased as the input SNR decreased. If IBM were able to isolate the speech from the noise without distortions, reaction times could be expected to remain as low as for speech in quiet.

In summary, our results reveal that at high SNRs, realistic noise reduction can provide an objective benefit for listeners in terms of reduced cognitive load, as reflected by shortened RTs. This beneficial effect was not present at the lower SNRs (see the extra condition at SNR = −10 dB). Ideal noise-reduction processing (IBM) can successfully reduce RTs even at −10 dB SNR.

In the AR task, some data points in Figure 4 are negative for IBM and MMSE, indicating that participants were faster at those SNRs compared with +∞ dB SNR (i.e., when no noise was added). This can probably be attributed to a residual learning effect that may have persisted despite the balancing between the 12 conditions. Although we started with a training session before the main experiment, RTs may have a residual learning effect in the first few conditions of the main experiment. This learning effect may be larger in the more complex AR task compared with the ID task. In addition, the average RTs varied considerably across participants. Balancing the conditions is only effective if all participants experience the exact same learning effect, which may not have been the case here. To reduce the effect of learning on the data in future experiments, the length of the training session should be extended.

The effect of SNR on ΔRT was significant for both tasks, indicating that at poorer SNRs, participants took more time to respond. Pairwise comparison revealed that

ΔRT for -5 dB SNR was higher than for all other SNRs. Houben et al. (2013) found that ΔRT differed significantly between all successive SNRs included. This difference in the effect of SNR on ΔRTs could be explained by comparing the SI data of the two studies. SI results show that the intelligibility performance in Houben et al. (2013) reached > 97% at higher SNR (-1 dB SNR, their Figure 1) than for our materials (-5 dB SNR). This difference implies that our measurements were done at SNRs with higher intelligibility, resulting in smaller differences in performance and ΔRT between successive SNRs, because the ceiling effect was stronger in our region of interest.

One possible reason for higher intelligibility in the same SNR region in our results compared with the results of Houben et al. (2013) is that they used different speech materials. The speech material used by Houben and colleagues was an older version of the digit triplet test (Smits et al., 2004), uttered by a female speaker and using a band-limited speech spectrum (300 Hz to 3.4 kHz). Our stimuli were uttered by a male speaker and used the full speech spectrum (50 Hz to 16 kHz) and thus contained more speech information. Even when the speech is (partially) masked by the noise, the additional information contained in the frequencies that were not present in the previous band-limited stimuli may add to SI. For the same increase in SNR (e.g., 1 dB higher SNR), additional speech information will be added with broadband stimuli (i.e., the information contained in the frequencies below 300 Hz and above 3.4 kHz). Hence, our stimuli produced results in data that showed a ceiling effect for intelligibility scores that started at a lower SNR. Similarly, the effect of the addition of noise on the RT will be different for the previous band-limited stimuli and our broadband stimuli. We observed smaller differences in ΔRTs between the successive SNRs in our experiment than those observed by Houben et al. (2013), and this might be related to the stronger ceiling effect in our region of interest.

We observed that the RTs were approximately 0.5 s longer for the AR task than for the ID task. This was expected because the required calculation in the AR task makes this task more complex than the ID task (Houben et al., 2013). In addition, the complexity of the AR task led to a lower performance or percentage of correct responses (96.8%) than in the ID task (98.7%). To evaluate whether this performance was at ceiling level during the ID and AR task, we analyzed the percentage of correct responses for both tasks. This analysis revealed no significant effects of processing condition or SNR, confirming that the measurements were at maximum performance, or ceiling level.

With respect to the two different tasks, the AR task appears to be more sensitive in the evaluation of noise reduction than the ID task. This difference between the two tasks is in line with the findings of Houben et al. (2013), who concluded that the AR task is more affected by the noise level than the ID task, which suggests that noise reduction may influence the AR task more as well. Sarampalis et al. (2009) explained that communication is a complex process, involving more than just physiological auditory functions, for instance, selectively attending to sound sources, storing information in memory, and generating quick appropriate responses. These processes were more involved in the AR task than in the ID task. Moreover, the results are consistent with the assumption that noise reduction may release attention resources in the brain, which could then be used for other, simultaneous tasks (Lunner et al., 2009; Rudner, Lunner, Behrens, Thorén, & Rönnberg, 2012; Sarampalis et al., 2009) such as AR.

Subjective Listening Effort

LErs revealed that listening to triplets processed with IBM required significantly less effort than listening to unprocessed stimuli and stimuli processed with MMSE. Thus, ideal noise reduction (IBM) significantly reduces subjective listening effort, while realistic noise reduction (MMSE) does not. This concurred with the observations made by Brons et al. (2012). However, the results of other studies on the influence of realistic noise reduction on subjective listening effort differ. Some studies found that LErs were significantly lower due to noise reduction (Bentler et al., 2008; Luts et al., 2010), while other studies found no difference in LEr for noise reduction and no processing (Alcántara et al., 2003; Brons et al., 2013; Desjardins & Doherty, 2014). The reason for these differing results is unknown but could lie in the different noise-reduction algorithms, SNRs, and presentation of stimuli used in the studies.

The results of the AR task and LEr task are very similar in that both showed the best results (ΔRT and ratings) for the IBM, the worst results for the unprocessed condition, and for both tasks, the MMSE results lay in between the results for IBM and the unprocessed stimuli. However, the reduction in ΔRTs was significant for both noise-reduction algorithms compared with the unprocessed condition, while the reduction in LEr was significant only for the stimuli processed with IBM.

The previously mentioned hypothesis that noise reduction relieves the cognitive load (Sarampalis et al., 2009) suggests that listening effort is reduced by noise-reduction processing. The measured differences in ΔRTs may be considered an objective measure of change in cognitive load. In other words, we argue that the RTs can capture effects related to noise reduction that are real and objective but that nevertheless may not be found in the subjective rating task. Our results imply that the objective AR task may be more sensitive in detecting the influence of noise reduction on cognitive load than the subjective rating task. Alternatively, it could be that the measured ΔRT and LEr index reflect different things. RTs are believed to be related to speech processing rate, representing auditory processing speed (McGarrigle et al., 2014). On the other hand, subjective LErs are thought to measure perceived listening effort, and both measures (RT and LEr) might be determined by different underlying mechanisms. Neher et al. (2014) evaluated the benefit from different binaural noise reduction settings with a dual-task paradigm (speech recognition and visual reaction time) and a measure of perceived listening effort using a 9-point scale. They found that reaction time and perceived listening effort scores were unrelated and state that this outcome implies that the two measures capture different perceptual aspects. In our study, we found a consistent positive effect of MMSE relative to unprocessed stimuli (see Figure 5). However, this effect was nonsignificant and smaller than the effect of IBM. The finding that MMSE caused a significant reduction in RT, but not in LEr, may be attributed to speech distortion as mentioned in the previous section, which is more pronounced in MMSE than in IBM. Although in terms of RTs the distortions did not completely cancel out the positive effect of reduced noise by MMSE, it seems that in rating the perceived effort the participants put more weight on this negative effect of noise reduction than on the positive effect of reduced noise.

Conclusions

We measured objective benefit for both noise-reduction algorithms in that they reduced the RTs to digit triplets at high SNRs (SI > 97%). This effect of noise reduction was found only for the more complex AR task and not for the ID task. In other words, the measurement results suggest that RT to an AR task can provide an objective measure of the benefit of noise reduction. Subjectively, rated listening effort showed a significant benefit of IBM only. The finding that MMSE led to significantly improved RTs suggests that our method using RTs is more sensitive than other methods applied because this effect could not be detected with our SI test or subjective LEr. The next required step would be to examine if this effect also holds for hearing-impaired listeners, hearing-aid users, other types of signal processing, and how RT relates to perceived listening effort.

Acknowledgments

We would like to thank all participants who took part in this study, and Sabine Engels who was willing to edit this article for language errors.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Alcántara, J. I., Moore, B. C. J., Kühnel, V., & Launer, S. (2003). Evaluation of the noise-reduction system in a commercial digital hearing aid. *International Journal of Audiology*, *42*(1), 34–42.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28.

Bentler, R. A. (2005). Effectiveness of directional microphones and noise reduction schemes in hearing aids: A systematic review of the evidence. *Journal of the American Academy of Audiology*, *16*(7), 473–484.

Bentler, R., & Chiou, L.-K. (2006). Digital noise reduction: An overview. *Trends in Amplification*, *10*(2), 67–82.

Bentler, R., Wu, Y.-H., Kettel, J., & Hurtig, R. (2008). Digital noise reduction: Outcomes from laboratory and field studies. *International Journal of Audiology*, *47*(8), 447–460.

Boymans, M., & Dreschler, W. A. (2000). Field trials using a digital hearing aid with active noise reduction and dual-microphone directionality. *Audiology*, *39*(5), 260–268.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436.

Brons, I., Houben, R., & Dreschler, W. A. (2012). Perceptual effects of noise reduction by time-frequency masking of noisy speech. *Journal of the Acoustical Society of America*, *132*(4), 2690–2699.

Brons, I., Houben, R., & Dreschler, W. A. (2013). Perceptual effects of noise reduction with respect to personal preference, speech intelligibility, and listening effort. *Ear and Hearing*, *34*(1), 29–41.

Brookes, M. (1997). Voicebox: Speech processing toolbox for MATLAB. Retrieved from www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

Desjardins, J. L., & Doherty, K. A. (2014). The effect of hearing aid noise reduction on listening effort in hearing-impaired adults. *Ear and Hearing*, *35*(2), 600–610.

Downs, D. W. (1982). Effects of hearing aid use on speech discrimination and listening effort. *Journal of Speech and Hearing Disorders*, *47*(2), 189–193.

Ephraim, Y., & Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *32*(6), 1109–1121.

Gatehouse, S. & Gordon, J. (1990). Respose times to speech stimuli as measures of benefit from amplification. *British Journal of Audiology*, *24*, 63–68.

Gustafson, S., McCreery, R., Hoover, B., Kopun, J. G., & Stelmachowicz, P. (2014). Listening effort and perceived clarity for normal-hearing children with the use of digital noise reduction. *Ear and Hearing*, *35*(2), 183–194.

Houben, R., Dijkstra, T. M. H., & Dreschler, W. A. (2012). Analysis of individual preferences for tuning noise-reduction algorithms. *Journal of the Audio Engineering Society*, *60*(12), 1024–1037.

Houben, R., van Doorn-Bierman, M., & Dreschler, W. A. (2013). Using response time to speech as a measure for listening effort. *International Journal of Audiology*, (0), 1–9.

Hu, Y., & Loizou, P. C. (2008). *Techniques for estimating the ideal binary mask*. Paper presented at the Proceedings of the 11th International Workshop Acoustic Echo and Noise Control, Seattle, WA.

Huckvale, M., & Frasi, D. (2010). *Measuring the effect of noise reduction on listening effort*. Paper presented at the Audio Engineering Society Conference: 39th International Conference: Audio Forensics: Practices and Challenges, Hillerød, DK.

Huckvale, M., & Leak, J. (2009). *Effect of noise reduction on reaction time to speech in noise*. Paper presented at the Proceedings of the 10th Annual Conference of the International Speech Communication Association (pp. 1–4), Brighton, UK.

Jorgensen, S., & Dau, T. (2011). Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *Journal of the Acoustical Society of America*, *130*(3), 1475–1487.

Li, N., & Loizou, P. C. (2008). Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. *Journal of the Acoustical Society of America*, *123*(3), 1673–1682.

Lunner, T., Rudner, M., & Rönnberg, J. (2009). Cognition and hearing aids. *Scandinavian Journal of Psychology*, *50*(5), 395–403.

Luts, H., Eneman, K., Wouters, J., Schulte, M., Vormann, M., Buechler, M., . . . Spriet, A. (2010). Multicenter evaluation of signal enhancement algorithms for hearing aids. *Journal of the Acoustical Society of America*, *127*(3), 1491–1505.

Marzinzik, M. (2000). *Noise reduction schemes for digital hearing aids and their use for the hearing impaired* (PhD thesis). Universität Oldenburg, DE, Germany.

McGarrigle, R., Munro, K. J., Dawes, P., Steward, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'white paper'. *International Journal of Audiology*, *53*(7), 433–440.

Neher, T., Grimm, G., & Hohmann, V. (2014). Perceptual consequences of different signal changes due to binaural noise reduction: Do hearing loss and working memory capacity play a role? *Ear and Hearing*, *35*(5), e213–e227.

Ng, E. H. N., Rudner, M., Lunner, T., Pedersen, M. S., & Rönnberg, J. (2013). Effects of noise and working memory capacity on memory processing of speech for hearing-aid users. *International Journal of Audiology*, *52*(7), 433–441.

Olsen, W. O. (1998). Average speech levels and spectra in various speaking/listening conditions: A summary of the Pearson, Bennett, & Fidell (1977) report. *American Journal of Audiology*, *7*(2), 21–25.

Pals, C., Sarampalis, A., & Baskent, D. (2013). Listening effort with cochlear implant simulations. *Journal of Speech, Language, and Hearing Research*, *56*(4), 1075–1084.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*(3), 510–532.

Ricketts, T. A., & Hornsby, B. W. (2005). Sound quality measures for speech in noise through a commercial hearing aid implementing "digital noise reduction.". *Journal of the American Academy of Audiology*, *16*(5), 270–277.

Rudner, M., Lunner, T., Behrens, T., Thorén, E. S., & Rönnberg, J. (2012). Working memory capacity may influence perceived effort during aided speech recognition in noise. *Journal of the American Academy of Audiology*, *23*(8), 577–589.

Sarampalis, A., Kalluri, S., Edwards, B., & Hafter, E. (2009). Objective measures of listening effort: Effects of background noise and noise reduction. *Journal of Speech, Language, and Hearing Research*, *52*(5), 1230–1240.

Smits, C., Kapteyn, T.S., & Houtgast, T. (2004). Development and validation of an automatic speech-in-noise screening test by telephone. *International Journal of Audiology*, *43*(1), 15–28.

Smits, C., Goverts, S. T., & Festen, J. M. (2013). The digits-in-noise test: Assessing auditory speech recognition abilities in noise. *The Journal of the Acoustical Society of America*, *133*(3), 1693–1706.

Studebaker, G. A. (1985). A "rationalized" arcsine transform. *Journal of Speech, Language, and Hearing Research*, *28*(3), 455–462.

Wagenaar, W. A. (1969). Note on the construction of diagram-balanced Latin squares. *Psychological Bulletin*, *72*(6), 384–386.

Wang, D. (2005). On ideal binary mask as the computational goal of auditory scene analysis. In P. Divenyi (Ed.), *Speech separation by humans and machines* (pp. 181–197). Boston, MA: Kluwer Academic Publishers.

Wang, D. (2008). Time-frequency masking for speech separation and its potential for hearing aid design. *Trends in Amplification*, *12*(4), 332–353.

Wang, D., Kjems, U., Pedersen, M. S., Boldt, J. B., & Lunner, T. (2009). Speech intelligibility in background noise with ideal binary time-frequency masking. *The Journal of the Acoustical Society of America*, *125*(4), 2336–2347.

Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record*, *58*(3), 475–482.