*Original Article*

Trends in Hearing
Volume 23: 1–21

DOI: 10.1177/2331216519854597
journals.sagepub.com/home/tia
SAGE

# Energetic and Informational Components of Speech-on-Speech Masking in Binaural Speech Intelligibility and Perceived Listening Effort

**Jan Rennies[1,2,3], Virginia Best[1], Elin Roverud[1], and Gerald Kidd Jr.[1]**

**Abstract**

Speech perception in complex sound fields can greatly benefit from different unmasking cues to segregate the target from interfering voices. This study investigated the role of three unmasking cues (spatial separation, gender differences, and masker time reversal) on speech intelligibility and perceived listening effort in normal-hearing listeners. Speech intelligibility and categorically scaled listening effort were measured for a female target talker masked by two competing talkers with no unmasking cues or one to three unmasking cues. In addition to natural stimuli, all measurements were also conducted with *glimpsed* speech—which was created by removing the time–frequency tiles of the speech mixture in which the maskers dominated the mixture—to estimate the relative amounts of informational and energetic masking as well as the effort associated with source segregation. The results showed that all unmasking cues as well as glimpsing improved intelligibility and reduced listening effort and that providing more than one cue was beneficial in overcoming informational masking. The reduction in listening effort due to glimpsing corresponded to increases in signal-to-noise ratio of 8 to 18 dB, indicating that a significant amount of listening effort was devoted to segregating the target from the maskers. Furthermore, the benefit in listening effort for all unmasking cues extended well into the range of positive signal-to-noise ratios at which speech intelligibility was at ceiling, suggesting that listening effort is a useful tool for evaluating speech-on-speech masking conditions at typical conversational levels.

**Keywords**

speech intelligibility, listening effort, categorical scaling, cocktail party problem, informational masking

Date received: 19 September 2018; revised: 7 May 2019; accepted: 13 May 2019

## Introduction

The perception of speech in complex sound fields is influenced by many factors including the particularly adverse effects of environmental noise and interfering voices. Listeners with normal hearing often are able to overcome these adverse effects and may recognize and comprehend the speech of a target voice even at very unfavorable signal-to-noise ratios (SNRs). In a previous study investigating speech in noise perception (Rennies & Kidd, 2018), we found that spatial separation between target and nonspeech noise not only improved speech intelligibility but also reduced listening effort, even in conditions when speech intelligibility was at ceiling, that is, when 100% of the words could be correctly identified. The relationship between improvements in intelligibility and listening effort has important implications for many real-world listening situations, such as those in which the listener uses assistive listening devices (e.g., hearing aids

[1]Department of Speech, Language and Hearing Sciences, Boston University, MA, USA
[2]Fraunhofer Institute for Digital Media Technology IDMT, Project Group Hearing, Speech and Audio Technology, Oldenburg, Germany
[3]Cluster of Excellence Hearing4all, Carl-von-Ossietzky University, Oldenburg, Germany

**Corresponding author:**
Jan Rennies, Fraunhofer Institute for Digital Media Technology IDMT, Project Group, Marie-Curie-Str. 2, Oldenburg 26129, Germany.
Email: jan.rennies@idmt.fraunhofer.de

or cochlear implants) to improve spoken communication. The purpose of the present study was to investigate whether a spatial release from listening effort also may be observed in speech-on-speech (SOS) masking conditions and whether speech intelligibility and listening effort are related in conditions with different spatial and nonspatial sound segregation cues.

The ability to segregate target speech from interfering sound sources is often discussed using the term *cocktail party effect* (e.g., recent series of reviews in Middlebrooks, Simon, Popper, & Fay, 2017) and is commonly attributed at least in part to the benefit of listening with two ears when competing sounds are spatially separated (Cherry, 1953; see also reviews in Bronkhorst, 2000, 2015; Carlile, 2014; Yost, 1997). In SOS masking conditions, intelligibility is greatly enhanced when interfering voices are spatially separated from a target voice. For example, Marrone, Mason, and Kidd (2008) measured differences between speech reception thresholds (SRT, i.e., the SNR at which 50% of the test words can be reported correctly) for conditions where two competing talkers were colocated with the target versus when the masker talkers were spatially separated from the target. They found that spatial release from masking increased with increasing (symmetrical) separation in azimuth from the frontal target reaching a plateau of about 13 dB for separations larger than about 45°. Several other studies also have reported a large spatial release under similar conditions, although the amount of masking release differed between studies (e.g., Andéol, Suied, Scanella, & Dehais, 2017; Best, Marrone, Mason, & Kidd, 2012; Gallun, Kampel, Diedesch, & Jakien, 2013; Kidd et al., 2016; Swaminathan et al., 2015; Zekveld, Rudner, Kramer, Lyzenga, & Rönnberg, 2014).

In addition to spatial separation of sources, other sound segregation cues may also considerably improve target speech intelligibility. For example, gender differences between target and masking talkers (e.g., Kidd et al., 2016; Xia, Noorale, Kalluri, & Edwards, 2015; Zekveld et al., 2014) and rendering the masker speech unintelligible by time reversal (e.g., Best et al., 2012; Freyman, Balakrishnan, & Helfer, 2001; Gallun et al., 2013; Iyer, Brungart, & Simpson, 2010; Marrone et al., 2008; Swaminathan et al., 2015; see also review in Kidd & Colburn, 2017) may produce large reductions in SRTs. These masking release effects can equal—or exceed—the benefits due to spatial separation of sources (e.g., Kidd et al., 2016; Swaminathan et al., 2015). However, not all studies agree with respect to the relative benefit obtained from these various factors (e.g., Freyman et al., 2001; Xia et al., 2015; Zekveld et al., 2014). Similarly, the additional release from masking reported for combinations of cues (e.g., spatial separation and gender differences or time reversal) has differed substantially between studies. Some studies have found that combining multiple source segregation cues improved SRTs somewhat more than each cue individually (e.g., Best et al., 2012; Marrone et al., 2008; Swaminathan et al., 2015), whereas other studies have reported similar release from masking for either one or two unmasking cues (e.g., Xia et al., 2015; Zekveld et al., 2014).

One particular difficulty with comparing across studies is that they often differ in the relative contributions of energetic masking (EM) and informational masking (IM) due to differences in the stimuli and procedures that were used. EM describes the interference caused by maskers exciting the same spectrotemporal regions as the target sound, thereby reducing the target information available in the neural representation of the stimulus. In contrast, IM occurs when interferers share perceptual attributes with the target that can draw attention away from the target, lead to explicit confusion of the masker with the target, or generally cause uncertainty in the observer (e.g., see review in Kidd et al., 2008a). It has been shown that IM can occur even when there is no spectrotemporal overlap between target and maskers (e.g., Broadbent, 1952; also Best et al., 2011, 2013; Kidd et al., 2008b). The importance of assessing the role of IM when investigating the influence of unmasking cues was highlighted by Marrone et al. (2008) and Best et al. (2012), who found that spatial release from masking was larger for high-IM maskers (intelligible talkers) than for low-IM maskers (time-reversed talkers).

One way of disentangling the contributions of IM and EM in speech masking experiments is to use ideal time–frequency segregation (ITFS). This concept, proposed by Brungart, Chang, Simpson, and Wang (2006, 2009) for application to SOS masking, is based on the assumption that some *tiles* (units bounded by a specified frequency range over a specified duration) in the time–frequency (T-F) representation of the summed target and masker signals are dominated by the maskers and are, hence, not available to the listener due to EM. In contrast, other tiles are dominated by the target and can contribute to recognizing the target. If the masker-dominated tiles are eliminated from the mixture, then this does not remove previously accessible/useful target energy. However, it strongly reduces the masker energy and effectively eliminates IM. This reconstructed ''glimpsed speech'' therefore contains the same amount of accessible target energy as the natural mixture but little IM.

One implication of this is that the differences observed in speech intelligibility between natural masked speech and glimpsed speech can be attributed primarily to differences in IM. Another implication is that the glimpsed speech can be used to determine how much EM is present in a given mixture. This concept was proposed by Brungart et al. (2006, 2009) and was recently adopted by Kidd et al. (2016) to disentangle the role of IM and EM in speech intelligibility for different unmasking cues.

Kidd et al. (2016) found that glimpsed speech always produced lower SRTs than natural masked speech but that this difference was much larger in the baseline condition (intelligible, colocated, same-gender maskers) than in conditions where a source segregation cue was made available (spatial separation of sources, gender difference between target and masker talkers, or time reversal of the masker speech). Kidd et al. (2016) argued that the contribution of IM was much larger in the baseline condition and that the different cues provided considerable release from IM (e.g., as much as 20 dB in one case). The present study extended the findings of Kidd et al. (2016) by evaluating the benefit of three sound segregation cues individually and in combination. Moreover, corresponding estimates of listening effort were obtained for all conditions.

Listening effort, which can be defined as "the deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a listening task" (Pichora-Fuller et al., 2016, p. 5S), has received a great deal of attention in the research and clinical literature, although studies differ in the terminology they use in characterizing the task (e.g., *listening effort*, *listening difficulty*, *ease of listening*, *cognitive load*). The main motivation for studying listening effort is that it may provide insights into speech perception in realistic communication situations that extend beyond those gained from speech intelligibility measurements alone. For example, it has been shown that listening effort ratings can be used to discriminate among listening conditions where speech intelligibility is at or close to ceiling (e.g., Houben, van Doorn-Bierman, & Dreschler, 2013; Krueger et al., 2017b; Morimoto, Sato, & Kobayashi, 2004; Rennies & Kidd, 2018; Rennies, Schepker, Holube, & Kollmeier, 2014; Sato, Morimoto, & Wada, 2012; Schepker, Haeder, Rennies, & Holube, 2016). In such circumstances, the listening conditions may still differ considerably in how challenging it is for the listener to follow the target talker. This observation is particularly relevant for many everyday communication situations in which the typical SNRs are relatively high (Smeds, Wolters, & Rung, 2015). However, studies investigating listening effort in SOS conditions differ in their conclusions as to the relative strength of different unmasking cues in releasing listening effort (e.g., Andéol et al., 2017; Xia et al., 2015; Zekveld et al., 2014). This may be in part due to the different degrees of IM present in different studies. It may thus be helpful to estimate IM in SOS conditions to determine the role of IM and release from IM in listening effort.

In summary, the main goals of this study were (a) to examine the role of IM and EM for isolated and combined unmasking cues in speech intelligibility, (b) to assess listening effort in the same conditions and subjects, (c) to determine the relation between speech intelligibility and listening effort in these conditions, and (d) to test whether the various unmasking cues—separately and in combination—also could produce a release from listening effort in conditions for which speech intelligibility is at ceiling.

## Methods

### Subjects

Ten subjects (six female, four male) aged between 19 and 30 years participated in all parts of this study. All had pure tone thresholds not exceeding 20 dB hearing level at audiometric frequencies between 250 Hz and 8 kHz and had English as their native language. All but one subject had experience with speech intelligibility and listening effort measurements. Subjects were paid for their participation and gave informed consent, and all procedures were approved by the Boston University institutional review board (Protocol 2633E).

### Stimuli

The target speech was always uttered by the same female talker with a mean fundamental frequency of 198.1 Hz (standard deviation 37.1 Hz) and an average speaking rate of 3.1 syllables per second. Targets were sentences taken from the American English matrix sentence test (Kollmeier et al., 2015), having the fixed five-word structure *name-verb-numeral-adjective-object*. For each word group, 10 alternatives are available which can be randomly combined to produce syntactically correct, but semantically unpredictable sentences. The test material consists of 90 such sentences, which are combined to lists of 10 sentences, in which each of the 50 words occurs once. These short lists are then combined to lists of 20 or 30 sentences for intelligibility measurements. Due to the lack of semantic predictability, each sentence appears to the subjects as one of any of the $10^5$ possible random combinations, and memorizing any of the 90 sentences is unlikely. The sentence lists have been optimized to produce highly homogenous SRTs (see Kollmeier et al., 2015).

In all conditions, two interfering talkers were used as maskers. These masking talkers were either two different female talkers (in the same-gender condition) or two different male talkers (in the different-gender condition). These pairs were the same throughout the experiments. All masking talkers uttered the same type of matrix sentences as the female target talker and were taken from recordings of Hochmuth, Kollmeier, and Shinn-Cunningham (2018), which comprised the same 90 matrix sentences as the original matrix test. A random sentence was used for each masker in each trial. The masking talkers had similar average speaking rates (3.1

to 3.4 syllables per second). Their fundamental frequencies were $179.0 \pm 33.1$ Hz and $210.0 \pm 38.4$ Hz (female masker talkers) and $95.5 \pm 12.7$ Hz and $115.1 \pm 19.6$ Hz (male masker talkers). None of the talkers had a pronounced regional accent. The masker sentences started together with the target sentences. Due to variations in individual sentence lengths, it was possible that the interfering sentences were shorter than the target sentence in a particular trial. To avoid an unmasked target word at the end of the sentence in these cases, a second randomly selected masker sentence was appended, and then faded out after the end of the target sentence using a 20-ms Hann window.

Target and maskers were convolved with head-related impulse responses (HRIRs) to produce the desired spatial conditions. HRIRs were taken from the database of Kayser et al. (2009) and had been recorded with a head-and-torso simulator in an anechoic chamber with a distance of 80 cm between the source and the center of the head. The target was always presented from the front, while the interfering talkers could be either colocated at the front, or spatially separated at $\pm 90^\circ$ azimuth.

The intelligibility of the maskers was varied by either using intelligible forward sentences or time-reversed sentences. Time reversal of the masker sentences was achieved by reversing the full sentence. All combinations of masker gender, spatial configuration, and masker intelligibility were included in the experiment. This resulted in eight conditions: the reference condition (same-gender, colocated, intelligible masker); three conditions with one additional cue (masker gender, location, or intelligibility); three conditions with two of these cues; and one condition with all three cues. The third-octave spectra of each of the 90 masker sentences from each masker talker were matched to the corresponding frontal target sentence (after convolving with the HRIRs and time reversal if applicable) to minimize spectral differences between target and interferers.

### Ideal Time–Frequency Segregation

All eight conditions were tested both with natural masked stimuli as well as with glimpsed speech. To keep the processing the same in all conditions, the analysis and resynthesis required to produce the glimpsed conditions also was applied to the natural conditions. ITFS processing was identical to that employed in a previous study (Kidd et al., 2016, adopted from Brungart et al., 2006, 2009), consisting of the following steps conducted separately for each ear:

- Convolving the target signal with HRIRs,
- Summing the masker signals, scaling the stimuli to the desired SNRs, and performing the convolution with the HRIRs and the time reversal (if applicable),
- Separating the stimuli into 128 frequency channels (spanning 80 Hz to 8 kHz) and 20-ms time windows with an overlap of 10 ms to obtain a matrix of T-F values representing the energy in each of those tiles, separately for target and (summed) maskers,
- Applying a local SNR criterion of 0 dB, that is, comparing the target energy to the summed masker energy for each T-F unit to create a binary T-F-mask containing 1 for each unit in which the target energy is equal to or exceeds the masker energy, and 0 otherwise,
- Of the summed target and maskers, discarding all tiles with 0 and resynthesizing the waveform from the remaining tiles to obtain the glimpsed speech.

All stimuli were downsampled to 20 kHz before the ITFS processing, and then upsampled again to 44.1 kHz for playback.

### Calibration and Equipment

A large range of SNRs including very low (negative) and very high (positive) SNRs was used in this study (see next section). To avoid strong variations in overall loudness, which would have occurred if either the target or the interfering speech had been fixed in level, the overall presentation level was fixed at 70 dB sound pressure level (SPL; as in Zekveld et al., 2014). For natural masked speech, the desired SNRs were set in the digital domain, and then the mix was scaled to produce this presentation level. Note that the SNRs reported in this study differ from the target-to-masker ratios (TMRs) reported in previous studies (e.g., Kidd et al., 2016) in that the SNR is calculated based on the level of the sum of both interferers, whereas the TMR was calculated as the ratio between the target level and each individual interferer. A TMR of 0 dB hence corresponds to an SNR of about $-3$ dB for the conditions with two interferers in the present study. For glimpsed speech, the presentation level was set after applying the ITFS mask.

All stimuli were generated and controlled using MATLAB. For the intelligibility measurements, the AFC-MATLAB framework of Ewert (2013) was used. Listening effort measurements were conducted using a custom MATLAB framework. The digital output was D/A converted via an RME HDSP 9632 (ASIO) 24-bit sound card and delivered to the subjects via Sennheiser HD280 Pro headphones in sound-attenuated booths. The setup was calibrated to SPL using a Brüel and Kjær (B&K) 4153 artificial ear, a B&K 4947 ½ in. microphone, a B&K ZC-0032 preamplifier, and a B&K 2250 sound level meter. The right ear served as reference point for the calibration, but the level at the two ears was always the same within the limits of the headphones and the recorded HRIRs.

## Procedures

The experiments were conducted in two sessions of about 2 hr each, that is, one session for the speech intelligibility measurements and one for the listening effort measurements. Half of the subjects conducted the speech intelligibility measurements first, and the other half started with the listening effort measurements. Prior to starting the measurements in each session, subjects read brief instructions about the task and were informed that the target talker would always be the same female talker, while interfering voices could be either male or female, intelligible or unintelligible, and at the same or different locations. To familiarize them with the target voice, subjects then listened to an example sound file of about 94 s duration, which consisted of concatenated sentences of the target talker in the presence of all speech maskers of the experiment as well as glimpsed speech. Two SNRs were included for each condition in the example file for natural masked speech. The lower SNR was +6 dB (reference condition) and −6 dB (all other conditions); the higher SNR was 10 dB above these values. The examples of glimpsed speech were generated at an SNR of 0 dB. These example SNRs were selected to make sure the target voice could clearly be heard and was intelligible.

*Speech intelligibility measurements.* The speech intelligibility measurements were conducted using a closed-set procedure, that is, subjects could select the words they had recognized after each sentence on a graphical user interface, which consisted of the entire matrix of 50 words. Subjects confirmed their choices by pressing a button, which triggered the presentation of the next sentence. Two initial training SRTs were measured using an adaptive track with lists of 20 sentences and a stationary speech-shaped noise to familiarize the subjects with the word grid and the task to reduce training effects typical for matrix sentence tests (Wagener, Brand, & Kollmeier, 1999). Each list contained each word of the matrix 2 times, but no sentence occurred twice. Subsequently, the experimental conditions were measured, each with a new random list of 30 sentences (i.e., each word was included 3 times). The initial SNR of the adaptive track was +6 dB for the reference condition, −6 dB for the remaining natural masked speech conditions, and −10 dB for the glimpsed conditions. These starting levels were chosen to ensure both good intelligibility of the target at the beginning of the track and fast convergence of the adaptive procedure. The SNR of the subsequent sentence presentation was varied adaptively using the procedure described by Brand and Kollmeier (2002). The 30 sentences were equally split into two interleaved tracks, one of which converged to an SNR corresponding to 80% correctly understood words ($SRT_{80}$), while the other converged to 20% correctly understood words ($SRT_{20}$). The reference condition was measured twice to increase robustness of the individual baseline. The order of the 16+1 conditions was randomized for each subject, and each condition was measured completely before a new condition was started.

*Listening effort measurements.* Listening effort was measured by categorical listening effort scaling using a constant stimuli procedure. Each presentation consisted of three different concatenated target sentences with randomly chosen interferer sentences or ITFS processing. These three target-masker sentence pairs were played in a loop until the subjects made their assessment of "How much effort do you have to spend to understand the speech?" There was an initial period of 6 s during which the subjects had to listen only and could not make a choice. This ensured that they heard about two entire target sentences before making their choice. The response scale was the same as used in previous studies (Krueger, Schulte, Brand, & Holube, 2017a; Krueger et al., 2017b; Rennies & Kidd, 2018; Rennies, et al., 2014; Schepker et al., 2016) and consisted of 13 listening effort ratings with the seven named categories *no effort* (1 effort scaling categorical unit, ESCU), *very little effort* (3 ESCU), *little effort* (5 ESCU), *moderate effort* (7 ESCU), *considerable effort* (9 ESCU), *very high effort* (11 ESCU), and *extreme effort* (13 ESCU), as well as six unnamed categories in between. In addition, a 14th category labeled *I cannot understand the target talker at all* was provided for conditions in which subjects could not hear any target speech (and hence an assessment of listening effort would not be meaningful, see Krueger et al., 2017a). The label of this 14th category differed from the previous studies, which had all employed unintelligible noise and had labeled this category *only noise*.

Target and maskers were presented at different fixed SNRs that had been selected in pilot experiments to cover a large range of the response scale for each condition. For the reference condition, the SNRs ranged from −10 to 30 dB in steps of 5 dB. For all other natural masked conditions, the SNRs ranged from −30 to +20 dB in steps of 5 dB. All 86 combinations of natural masked conditions and SNRs were combined in one block with a random order. SNRs for all glimpsed conditions ranged from −30 to +10 dB in steps of 5 dB. All 72 combinations of SNR and glimpsed condition were also combined in one block, and the order was again randomized. For each combination, new random target and interferer sentences were used, and the next combination was presented immediately after the subjects made their choice. Each block was repeated 4 times resulting in four listening effort ratings per condition and SNR from each subject. The median of the four ratings was calculated as the estimated listening effort. All $4 \times 2$ blocks were conducted in random order.

### Data Analyses

Model functions for the speech intelligibility data were determined by fitting a sigmoidal psychometric function with two degrees of freedom (Brand & Kollmeier, 2002) to the obtained $SRT_{20}$ and $SRT_{80}$ data for each subject and condition. The chance performance level of the model function was set to 0% because listeners were not forced to respond. The initial data analysis showed that the slopes of the psychometric functions were quite similar in all 16 conditions (between 3.9% and 6.0%/dB, mean 5.0%/dB), that is, the conditions differed mainly in a horizontal shift along the SNR axis. Therefore, the different conditions were analyzed by deriving 50% correct SRTs from the psychometric functions and comparing these across conditions.

The listening effort data were fitted using the same model function as proposed by Krueger et al. (2017a, 2017b). This function consists of two straight lines that intersect at 7 ESCU (i.e., the midpoint of the listening effort scale). The transition between the two lines is smoothed by a Bézier function. The model function has three degrees of freedom: the slopes of the two lines and the SNR of the intersection point. Ratings of the 14th category (i.e., when subjects could not understand the target talker at all) were excluded from the fitting. Despite the relatively broad range of SNRs employed in each condition, the resulting listening effort ratings did not cover a large range of the scale for all subjects and conditions. Specifically, in some cases, no data points at or above 10 ESCU, or at or below 4 ESCU were available. In these cases, a first-order polynomial was used as a simpler model function. Furthermore, the influence of floor and ceiling effects on the fitting procedure was limited by investigating the ratings at the highest and lowest SNRs, respectively. Some subjects tended to rate listening effort consistently lower than others such that several of the highest SNRs in a given condition were rated with 1 ESCU, that is, *no effort*. In these cases, only the lowest SNR for which this rating was obtained was kept for fitting the model function. The same was applied at the upper end of the rating scale. This is equivalent to the data collection procedure in the adaptive version of the categorical listening effort scaling (Krueger et al., 2017a), where SNRs at which the subject rated *no effort* (*extreme effort*) are not repeated or exceeded on an individual basis. As for the speech intelligibility data, SNRs at a fixed listening effort rating (e.g., *medium effort*) were derived from the psychometric functions and then compared across conditions.

For both speech intelligibility and listening effort, the data were tested for normal distribution. If normality could be assumed, repeated-measures analyses of variance (ANOVAs) were conducted with a significance level of .05. Degrees of freedom were Greenhouse-Geisser corrected. Paired comparisons were conducted as post hoc tests if applicable, and Šidák corrections were applied for multiple comparisons unless stated otherwise. If normality could not be assumed, Friedman tests followed by Wilcoxon rank sum comparisons (if applicable) were conducted. The data will be made available upon request.

## Results

### Speech Intelligibility Measurements

*Masking release.* Figure 1 shows mean SRTs for natural masked speech (top) and glimpsed speech (middle) as well as the *additional masking* (bottom), which was calculated as the difference between SRTs for natural masked speech and glimpsed speech. The term *additional masking* is used because—apart from representing the amount of IM—the difference may also include residual effects of masker energy retained in the target-dominated tiles (see Discussion section). Error bars represent inter-individual standard errors. For natural masked speech, the highest SRT of −2.9 dB SNR was measured in the reference condition. SRTs in conditions with a single unmasking cue (dark gray) were considerably lower (−20.5 to −17.3 dB SNR). A further reduction of SRTs was observed when two unmasking cues were combined (light gray, SRTs −24.6 to −22.8 dB SNR). The lowest SRT of −26.3 dB SNR was observed when all three unmasking cues were provided (white). A Friedman test confirmed the significant effect of condition, $\chi^2(7) = 56.629$, $p < .001$. This test was of course considerably influenced by the obvious difference between the reference condition and all other conditions. As a post hoc analysis, the Friedman test was therefore rerun without the reference condition, indicating that the effect of condition was still significant, $\chi^2(6) = 45.345$, $p < .001$. However, because pairwise comparisons of all conditions required 21 Wilcoxon tests, a considerable reduction in significance level was required, and none of the differences could be reported as significant. The SRT differences for natural masked speech with one or more unmasking cues are hence reported as trends here.

For glimpsed speech, all SRTs were lower than in the corresponding natural masked speech condition. The largest difference (additional masking of 19.2 dB) was observed for the reference condition. For the conditions with at least one unmasking cue, SRTs were between −25.0 and −30.9 dB SNR and tended to decrease with increasing number of unmasking cues. A one-way ANOVA confirmed the significant effect of condition, $F(3.279, 29.510) = 24{,}823$, $p < .001$. Post hoc tests indicated that SRTs in the reference condition differed significantly from all other SRTs except for male maskers, and the SRT in the condition with all three unmasking

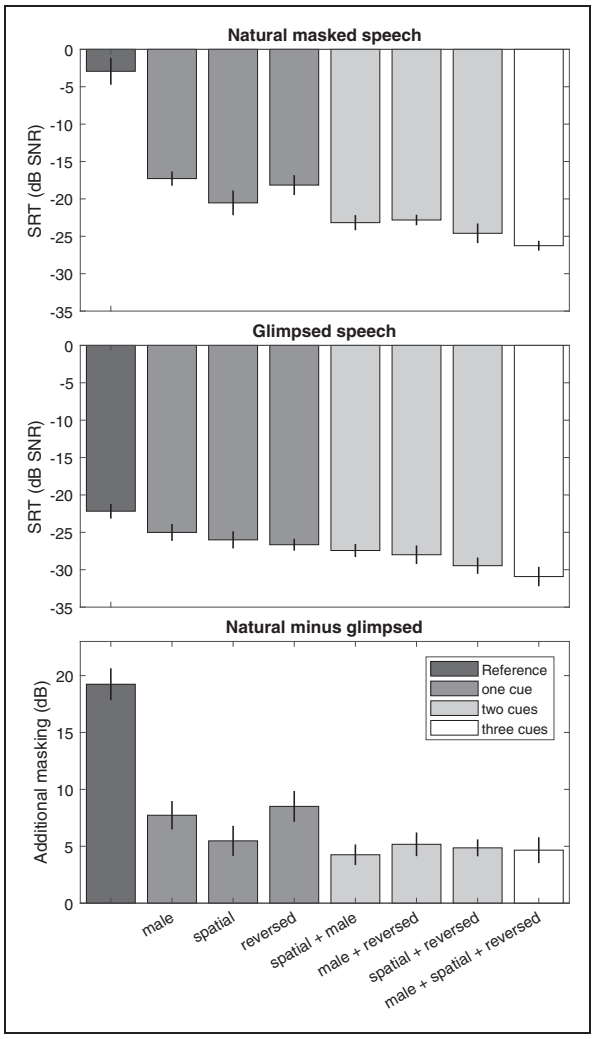


**Figure 1.** Mean SRT and standard errors for natural masked speech (top) and glimpsed speech (middle). The SRT difference (natural minus glimpsed) is termed *additional masking* and indicated in the bottom panel. Gray scales represent the number of provided unmasking cues of the different conditions.
SRT = speech reception threshold; SNR = signal-to-noise ratio.

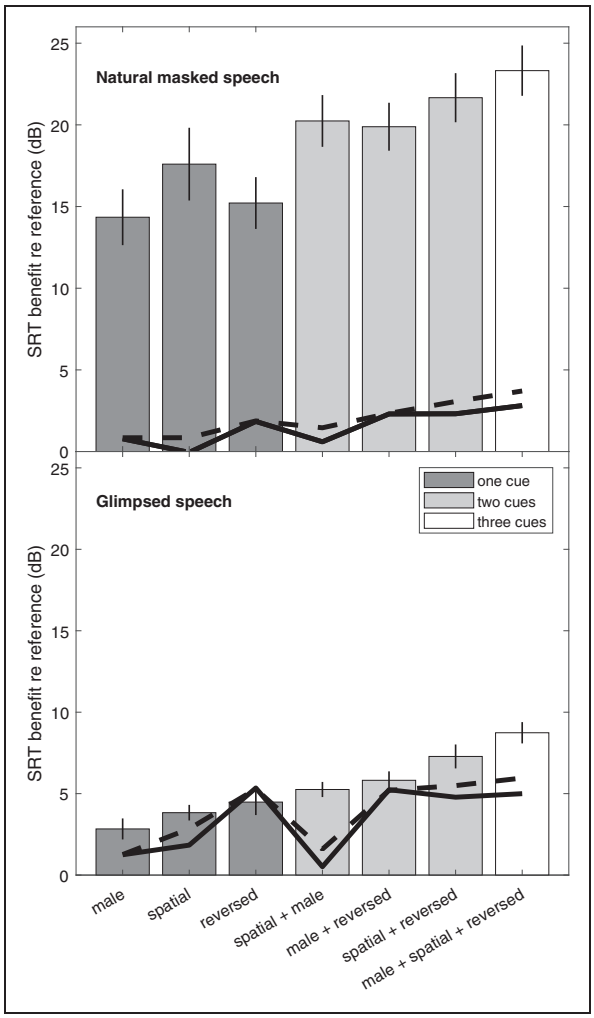


**Figure 2.** Mean benefit in SRT obtained by the different unmasking cues for natural masked speech (top) and glimpsed speech (bottom). Error bars indicate standard errors. Gray scales represent the number of provided unmasking cues of the different conditions. Black solid and dashed lines represent the energetic differences estimated from the retained monaural and better-ear glimpse energy, respectively (see text).
SRT = speech reception threshold.

cues differed significantly from the first three SRTs (reference, male maskers, spatial maskers). In addition, some of the intermediate SRTs also differed significantly from each other so that overall it could be concluded that SRTs measured for glimpsed speech were not uniform. A one-way ANOVA indicated that the effect of condition on additional masking was significant, $F(2.275, 24.538) = 29.041$, $p < .001$. Post hoc tests indicated that, apart from the obvious differences between the reference condition and all other conditions, the additional masking differed significantly between male versus male+reversed maskers and between male versus male+spatial+reversed maskers.

The masking release due to the different unmasking cues was assessed as a shift in SRT relative to the reference condition and is illustrated in Figure 2 for natural masked speech (top) and glimpsed speech (bottom). For natural masked speech, the smallest mean masking release was observed for masker gender difference (14.4 dB) and time reversal (15.3 dB), while the SRT benefit due to spatial separation was slightly larger (17.6 dB). When two combined unmasking cues were provided, the mean masking release increased to between 19.9 and 21.7 dB. The largest masking release of 23.3 dB occurred when all three unmasking cues were provided. The differences in masking release between conditions were significant according to a one-way ANOVA, $F(2.861, 25.571) = 22.722$, $p < .001$. Post hoc comparisons indicated that masking releases did not differ

significantly within the group of conditions with a single unmasking cue (dark gray). In comparison with conditions with more than one unmasking cue, male and reversed maskers always showed significantly lower masking release. In contrast, the masking release due to spatial separation did not differ significantly from the masking release measured with any combination of unmasking cues. Within the group of conditions with two unmasking cues (light gray), masking releases did not differ significantly. The condition with all three unmasking cues differed significantly from all other conditions that did not include spatial separation.

For the masking release measured with glimpsed speech, the main effect of condition was also significant, $F(2.990,\ 26.913) = 12.958$, $p < .001$. The differences between the following pairs of conditions were significant after correction: male versus male+reversed, male versus spatial+reversed, male versus male+spatial+reversed, spatial versus spatial+reversed, and spatial versus male+spatial+reversed.

*Estimating the contribution of EM.* In an attempt to estimate the contribution of EM/unmasking to the observed SRTs, an analysis of the target speech glimpses retained after applying the ITFS was conducted in the same way as done by Kidd et al. (2016). Note that the glimpsing only retained those speech tiles that were assumed to be target-dominated. Thus, with increasing overall SNR, an increasing proportion of tiles (and thus target energy) was retained in the glimpsed speech. This is illustrated in the top panel of Figure 3. The data shown here were derived as the ratio of root-mean-square values of the glimpsed speech to unprocessed target speech (computed with an *all-ones* binary mask), calculated from 50 randomly selected combinations of target and masker sentences in each condition as a function of SNR. The obtained energy ratios were very similar for both ears and are shown for the right ear only. Different symbols and line styles indicate the different conditions. This analysis revealed that conditions including masker time reversal (dashed lines) comprised a consistently higher proportion of retained target energy compared with conditions with forward maskers (solid lines) at the same overall SNR. Differences within the group of reversed maskers and within the group of forward maskers appeared to be small. To explore the role of the differences in retained energy between forward and reversed maskers, the time course of the retained glimpses was analyzed by calculating the proportion of retained glimpses across frequency channels for each 20-ms analysis window. An example of this is illustrated in the inset in the bottom panel of Figure 3. The solid line shows the mean proportion of retained tiles as a function of time for colocated, same-sex forward maskers, calculated across 50 realizations of target sentence and maskers at

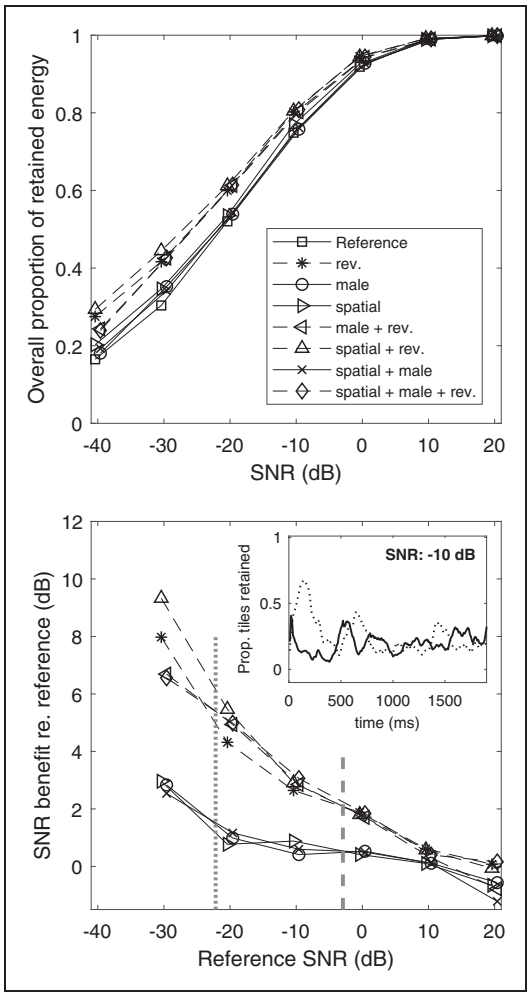


**Figure 3.** Top: proportion of retained energy after glimpsing as a function of SNR. Dashed lines represent reversed maskers, solid lines represent forward maskers. Bottom: corresponding SNR differences at equal retained energy relative to the reference condition as a function of reference SNR. Data are slightly shifted horizontally for readability. Vertical dashed and dotted lines mark the SRTs obtained for natural masked speech and glimpsed speech in the reference condition, respectively. The inset in the bottom panel illustrates the proportion of tiles retained during glimpsing as a function of time (see text) for forward (solid) and reversed maskers (dotted), for colocated same-sex maskers at an SNR of −10 dB.
SNR = signal-to-noise ratio.

an SNR of −10 dB. The dotted line shows the corresponding proportion of retained tiles for colocated, same-sex reversed maskers. This example shows that, while both curves fluctuate somewhat over time, there is a distinct difference in the first about 300 ms in that a considerably larger proportion of tiles are retained for the reversed maskers than for the forward maskers. This pattern of differences between forward and reversed maskers was also observed for spatial maskers, male maskers, and combinations thereof. This is in line with the

observed pattern of retained target energy across conditions (Figure 3).

To estimate what contribution the differences in target energy across conditions might have had on the SRT benefits across conditions, first, the equivalent SNR benefit relative to the reference condition at a given proportion of retained energy was computed as the horizontal distance between the curves for the reference condition (squares, solid line) and all other conditions. The bottom panel of Figure 3 shows the resulting SNR differences as a function of reference SNR. For forward maskers (black solid lines), the increase in retained target energy was no larger than 1 dB except at very low reference SNRs. In contrast, the retained energy for reversed maskers (black dashed lines) was considerably higher than in the reference condition, decreasing from more than 6 dB at very low SNRs to about 2 dB at a reference SNR of 0 dB. At positive reference SNRs, the SNR benefit disappeared. The gray vertical lines indicate the SRTs measured in the reference condition for natural masked speech (dashed) and glimpsed speech (dotted). Next, the SNR benefits at these reference SRTs were interpolated. The resulting values are plotted as solid lines in Figure 2. For natural masked speech (top panel), only a small portion of the observed unmasking effects (about 1 dB for forward maskers and about 2.5 dB for reversed maskers) could be attributed to the increased target energy at each ear. In contrast, the SRT benefits relative to the reference condition observed for glimpsed speech were quite similar to the derived differences in retained target energy (see bottom panel of Figure 2). Notable differences between SRT differences and differences in retained target energy of $>3$ dB occurred for conditions including spatial separation of target and maskers, where the observed SRT benefit was larger than could be expected from the monaural energy advantage (see Discussion section). One factor likely to affect speech perception but not captured by a monaural glimpse analysis is that the auditory system is believed to make use of better-ear glimpses, that is, favorable glimpses occurring in either ear at a given time frame. To estimate how better-ear glimpsing affected the present data, a new set of target signals was constructed from the glimpsed speech signals of the left and right ear as follows: For each T-F tile, the target energy at the left and right ear was compared, and the tile of the ear with the higher energy was kept while that of other ear was discarded. The energy of the resulting *better-ear glimpse stimulus* was computed for 50 random target-masker combinations for each condition, and then compared across conditions in the same way as described earlier for the monaural glimpse analysis. Dashed lines in Figure 2 show the SRT benefit relative to the reference condition that would be expected due to better-ear glimpsing. It can be seen that for all conditions with spatially separated maskers, the energy retained in better-ear glimpses was about 1 dB higher than the monaurally retained energy, while for all other conditions, no advantage occurred.

To assess the degree to which the differences in masking release between conditions could be attributed to differences in monaural or better-ear glimpse energy, the estimated differences in retained target energy (i.e., the values represented by solid and dashed black lines in Figure 2) were subtracted from the measured masking releases, and separate one-way ANOVAs for monaural and better-ear glimpses were conducted. For natural masked speech, these analyses showed that the main effect of condition was still significant—monaural glimpses: $F(2.880,\ 25.924) = 17.081,\ p < .001$; better-ear glimpses: $F(2.871,\ 25.843) = 13.233,\ p < .001$—indicating that not all of the observed differences could be accounted for by differences in glimpsed target energy. Post hoc comparisons indicated that the same significant differences as without the subtraction of the (monaural or better-ear) glimpse target energy were significant. In addition, the difference between spatial and reversed maskers was also significant when subtracting the monaural glimpse energy (which was not the case without subtracting monaural glimpse energy or with subtracting better-ear glimpse energy).

For glimpsed speech, the same analysis also revealed a significant main effect of condition—monaural glimpses: $F(3.011,\ 27.100) = 11.126,\ p < .001$; better-ear glimpses: $F(2.999,\ 26.995) = 6.730,\ p = .002$. Here, however, the pattern of significant pairs was changed considerably compared with the analysis without subtracting the energetic glimpsing advantages. While previously significant differences had only occurred between the two leftmost conditions and the two or three rightmost conditions in Figure 2 (see earlier), these differences were no longer significant when the energetic glimpse advantage was subtracted (both for monaural and better-ear glimpses). Instead, when subtracting the monaural glimpse advantage, the SRT benefit measured with spatial+male maskers now differed significantly from all other conditions except the one with all three unmasking cues. In addition, the difference between reversed maskers and male+spatial+reversed maskers was significant. The emerging difference between spatial+male maskers and the other conditions seems intuitive given that there was only a negligible monaural energetic advantage in the spatial+male condition (see solid black line in the bottom panel of Figure 2) while, for all other conditions, a considerable portion of the observed masking release was subtracted in this analysis. As a post hoc analysis, $t$ tests were conducted for each condition to test if the means (measured release minus the differences in glimpse energy) differed significantly from 0 dB. This was the case for spatial, spatial+male, and male+spatial+reversed maskers. When subtracting the better-ear glimpse

advantage, significant differences were between spatial+male maskers and spatial, reversed, and male+reversed maskers. The same post hoc analysis as for monaural glimpses showed that the means (measured release minus the differences in better-ear glimpse energy) differed significantly from 0 dB for spatial+male maskers and maskers with all three unmasking cues.

*Analysis of error patterns.* When considering error patterns in SOS masking conditions using closed-set target and masker stimuli, three types of errors could occur: subjects could (a) leave response fields blank (they were not forced to respond, referred to as *omissions* in the following), (b) select a word uttered by one of the maskers *(confusions)*, or (c) indicate a wrong word not uttered by any of the maskers *(random error)*. The error analysis conducted for all trials of the adaptive tracks is shown in Figure 4. The top panel shows the proportion of omissions among the overall errors made for natural masked speech (black) and glimpsed speech (white). The data indicated that omissions made up a similar proportion of errors in each condition (between about 0.3 and 0.4). The proportion of omissions was always higher for glimpsed speech than for natural masked speech, but these differences were small (0.06 on average across conditions). A somewhat smaller proportion of omissions was observed for natural masked speech in the reference condition (0.24) than for glimpsed speech in this condition (0.41).

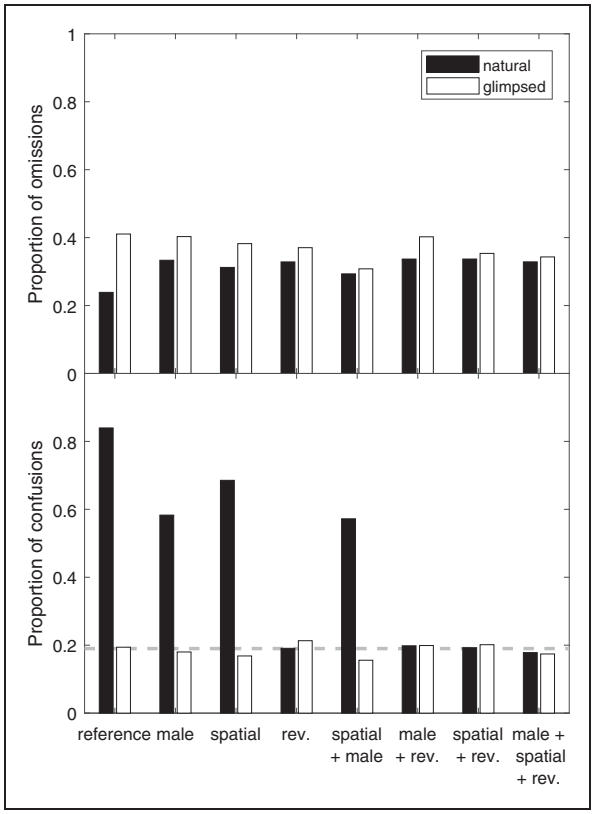


**Figure 4.** Proportion of errors due to omissions (top) and proportion of errors due to masker confusions (bottom) for natural masked speech (black) and glimpsed speech (white). The dashed line in the bottom panel indicates the chance level for masker confusions.

The bottom panel of Figure 4 shows the proportion of masker confusions measured based on the overall cases of wrongly selected words (i.e., the distance to a proportion of 1.0 in this panel is the proportion of random errors). The dashed horizontal line represents the chance level for random errors (which was $1 - 0.9^2 = 0.19$, i.e., the probability of the wrongly selected word being uttered by neither of the two maskers). For glimpsed speech, the proportion of confusions was close to the chance level, which was expected because there was little masker energy remaining. The same was observed for all conditions including reversed maskers, where the masker words were not intelligible. In contrast, the proportion was considerably higher than chance in all conditions with intelligible maskers. The largest proportion of masker confusion was observed in the reference condition (0.84), indicating that the majority of errors occurred because subjects chose a word uttered by one of the maskers. For spatially separated or male maskers, the proportion of confusions was reduced to between about 0.6 and 0.7.

## Listening Effort Measurements

*Psychometric functions and comparison across conditions.* Figure 5 shows the median listening effort ratings across subjects (symbols). Error bars represent interquartile ranges. Psychometric functions were fitted to the data points for each condition and are shown as solid and dashed lines for natural masked and glimpsed speech, respectively. The reference condition is shown in the top left panel. The other panels show data for the different combinations of unmasking cues as indicated. For comparison, the psychometric function obtained in the (natural speech) reference condition is replotted in each panel as gray lines.

The following general trends could be observed in the data: First, listening effort always decreased with increasing SNR. For some conditions, the highest SNRs employed in the experiment did not result in median ratings of very low listening effort (i.e., the data points did not cover the lowest three categories of the employed scale). Second, listening effort of natural masked speech was always reduced at a given SNR when one or more unmasking cues were available (compare black and gray solid lines in each panel). This listening effort reduction depended on condition, and it tended to be larger at lower SNRs, while toward the higher SNRs, the two curves converged. In general, the reduction in listening

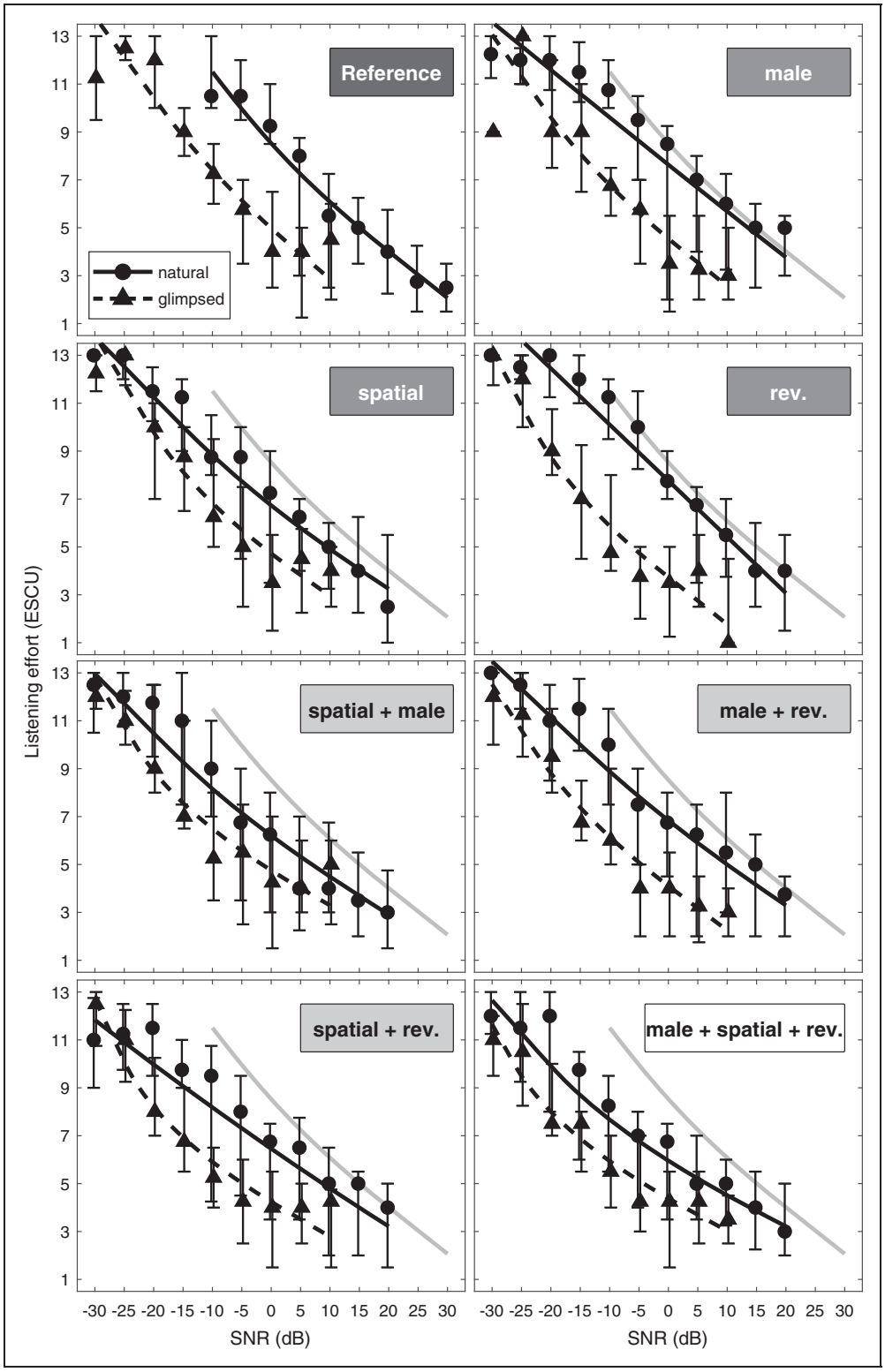


**Figure 5.** Psychometric functions of the listening effort ratings for natural masked speech (solid lines) and glimpsed speech (dashed lines). Symbols and error bars represent median values and interquartile ranges across subjects. Different panels show data of the eight combinations of unmasking cues. For comparison, the function of the natural reference condition is replotted in each panel (gray).
ESCU= effort scaling categorical unit; SNR = signal-to-noise ratio.

effort at lower SNRs appeared to be larger when more than one unmasking cue was available than for isolated cues. The third general observation was that listening effort for glimpsed speech was always lower than listening effort for natural masked speech with the same maskers (compare dashed and solid black lines in each panel). In the reference condition, the psychometric functions of natural masked speech and glimpsed speech were parallel across the entire range of measured SNRs (shifted by about 14 dB SNR). For all other conditions, the function for glimpsed speech showed a steeper slope at lower SNRs and listening effort for natural masked speech and glimpsed speech was more similar (or the same in some conditions) at very low SNRs.

To further explore the observed differences between conditions, the SNRs resulting in fixed listening effort ratings (i.e., the horizontal distances between the psychometric functions) were derived from the individual psychometric functions of each subject and condition. One subject was excluded from this analysis because no data points were available at the SNRs of interest. This was because the psychometric functions of this subject were considerably steeper than those of the other subjects. While the group psychometric function shown in the top left panel of Figure 5 suggests a decrease in listening effort from about 12 ESCU to 2 ESCU over a range of approximately 40 dB SNR, the corresponding ratings of this subject covered only about half this range, and the minimum rating of *no effort* was already reached at an SNR of 0 dB in most conditions. Consequently, no comparison of listening effort across conditions could be made at higher SNRs. Differences in listening effort ratings between subjects were also observed at the lower end of the SNR range. In particular, subjects differed with respect to the SNR at which they gave the highest rating *(extreme effort)* or indicated that they could not understand the target talker at all. This means that, at very low SNRs, data for comparing measurement conditions would not be available for all subjects because they reached ceiling at different SNRs. A closer inspection of the individual data in all 16 conditions showed that, for the remaining nine subjects, the range in which data were available covered SNRs between 0 and +15 dB. This SNR range was of particular interest because, as stated in the introduction and confirmed in the speech intelligibility measurements of the present study, speech intelligibility could be assumed to be very high/at ceiling at these SNRs, but conditions might still differ in listening effort.

To conduct a quantitative comparison across conditions, reference SNRs of 0, 10, and 15 dB were selected. According to the group psychometric functions (Figure 5), these SNRs corresponded to mean listening effort ratings of 8.8, 5.7, and 4.2 ESCU in the reference condition, respectively. For each subject and condition, the listening effort ratings at these SNRs in the reference condition were derived, and the SNR differences required to produce the same listening effort ratings were calculated for all other conditions. The corresponding mean values and interindividual standard errors for natural masked speech conditions are shown in the top panel of Figure 6. These SNR differences can be considered as a measure of the benefit provided by the different unmasking cues. The three bars in each group correspond to reference SNRs of 0, 10, and 15 dB. The unmasking benefit differed considerably between conditions: The smallest benefit was observed for gender difference or time reversal as the only unmasking cue, where the benefit was about 5 and 3 dB at reference SNRs of 0 and 15 dB, respectively. For spatial separation alone (second group of bars), the benefit was slightly larger (between 8 and 5 dB). When unmasking cues were combined, the SNR benefit was generally increased compared with isolated unmasking cues, except for the combination of male and reversed maskers, for which SNR differences were comparable with spatial separation alone. The largest benefit was observed when all three unmasking cues were combined (between about 13.5 and 7.5 dB). For all conditions, the benefit decreased with increasing reference SNR, which corresponded to the converging psychometric functions described earlier. A two-way ANOVA with factors reference SNR and condition confirmed these observations: Both main effects were significant—reference SNR: $F(1.097, 8.776) = 6.526$, $p = .030$; condition: $F(2.854, 22.832) = 10.992$, $p < .001$. The interaction between both factors was also significant, $F(3.652, 29.216) = 5.654$, $p = .002$, indicating that the effect of condition depended on reference SNR. Separate one-way ANOVAs for each reference SNR were conducted as post hoc analyses. In each case, the main effect of condition was significant—0 dB: $F(3.045, 24.363) = 16.230$, $p < .001$; 10 dB: $F(2.720, 21.758) = 7.882$, $p = .001$; 15 dB: $F(3.176, 25.408) = 5.246$, $p = .005$. The significance of the pairwise comparisons between conditions differed somewhat between reference SNRs: For a reference SNR of 0 dB, the masking release was significantly larger when all three unmasking cues were combined than for any of the cues in isolation. The masking release was also significantly larger for spatial+reversed maskers than for male or spatial maskers, larger for spatial+male maskers than for male maskers, and larger for spatial than for male maskers. For a reference SNR of 10 dB, the masking release observed for each individual unmasking cue was significantly smaller than for the condition with spatial+male maskers and for the condition with all three unmasking cues. For a reference SNR of 15 dB, the only significant differences were between spatial+male maskers and both reversed maskers and spatial maskers, and between reversed maskers and male+spatial+reversed maskers. It is worth mentioning that the differences between conditions in

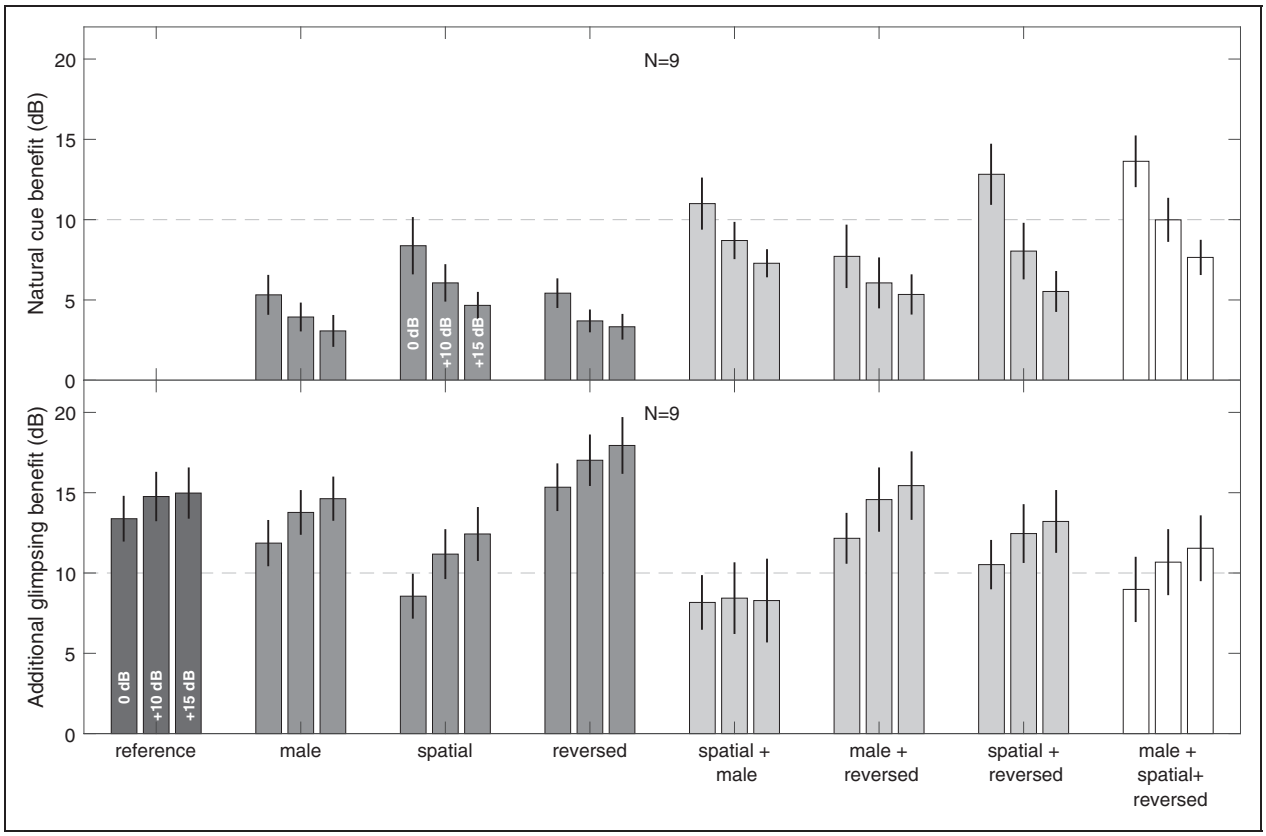


**Figure 6.** Top panel: mean equivalent SNR changes relative to the reference condition to obtain the same listening effort ratings as in the reference condition for natural masked speech at SNRs of 0, 10, and 15 dB (three bars of each group). Error bars indicate standard errors. Bottom panel: mean additional masking calculated as the difference between SNRs in natural listening conditions and glimpsed speech at the same listening effort values. The horizontal dashed line at 10 dB is included for visual guidance.

terms of retained glimpse energy were very small at these SNRs ($\leqslant$1 dB, see Figure 3) such that the observed differences were probably not attributable to differences in available target energy.

The same was true when considering the additional reduction in listening effort that was observed for glimpsed speech relative to natural masked speech, which was quantified using the same approach as described earlier: The *additional* SNR reduction required to produce the same three listening effort ratings was calculated for each subject and condition. This reduction corresponds to the horizontal distance between the solid and dashed black curves in Figure 5 and is the equivalent of the *additional masking* shown for the SRT data in Figure 1. In the context of listening effort, this can be considered as a measure of the listening effort reduction the subjects experienced when the target-from-masker segregation was conducted for them by means of glimpsing, while the amount of speech information was the same. The mean values and standard errors are shown in the bottom panel of Figure 6. In contrast to the benefit of unmasking cues relative to the reference condition, the additional benefit due to glimpsing increased with increasing reference SNR for most conditions, reflecting the convergence of the dashed and solid black curves at lower SNRs mentioned earlier. For the reference condition (dark gray bars), the glimpsing benefit was between 13 and 15 dB. It was similar in magnitude for male maskers and male+reversed maskers. For reversed maskers, the highest mean glimpsing benefit was observed (about 15 to 17 dB). When spatial separation was provided as an unmasking cue (either alone or in combination with other unmasking cues), the glimpsing benefit was reduced and was between about 8 and 13 dB. These data were subjected to a two-way ANOVA, which confirmed a significant main effect of reference SNR, $F(1.059,\ 8.464) = 15.026$, $p = .004$, as well as of condition, $F(2.958,\ 23.661) = 6.544$, $p = .002$, while the interaction between the factors was not significant, $F(3.376,\ 27.007) = 1.343$, $p = .281$. Post hoc comparisons indicated that data for a reference SNR of 0 dB differed significantly from the two higher reference SNRs, while data for the two higher reference SNRs did not differ significantly from each other. The only conditions that differed significantly

from each other after correction for multiple comparisons were those for reversed maskers and spatial+male maskers.

Comparing the top and bottom panel of Figure 6 indicates that the additional benefit due to glimpsing (bottom) was larger than the initial benefit due to the isolated unmasking cues (top, dark gray). This difference was more pronounced for male and reversed maskers than for spatially separated maskers and was also observed for male+reversed maskers. For combined unmasking cues including spatial separation, the difference between the natural cue benefit and the additional glimpsing benefit was smaller.

### *Relation Between Speech Intelligibility and Listening Effort*

One way of comparing speech intelligibility data and listening effort data is to extract speech recognition scores and listening effort ratings from the respective psychometric functions at fixed SNRs (cf. Rennies & Kidd, 2018). This was done for all conditions of the present study by sampling the group psychometric functions from SNRs of −30 to 30 dB in steps of 1 dB. The resulting pairs of intelligibility and listening effort are plotted in Figure 7. Different symbols represent the different conditions (gray: glimpsed speech, black: natural masked speech). For each condition, this representation consisted of two main parts: In the first part, intelligibility was at ceiling and listening effort increased from very low ratings to approximately *medium effort* (7 ESCU). In the second part, speech intelligibility decreased as listening effort increased further. Comparing the different conditions, it can be seen that most data points were aligned within a corridor of ±1 ESCU around a sigmoid function fitted to the data (solid and dashed lines), that is, the relation between intelligibility and listening effort was similar for most conditions. One notable exception was the reference condition, which was systematically below all other conditions in the representation of Figure 7. The differences (as measured to the fitted sigmoid function) were approximately 2 ESCU and 25% to 30% in the horizontal and vertical direction, respectively. Another smaller but seemingly systematic effect was that the data points for glimpsed speech were consistently below the data points of natural masked speech in the region where intelligibility started to decrease.

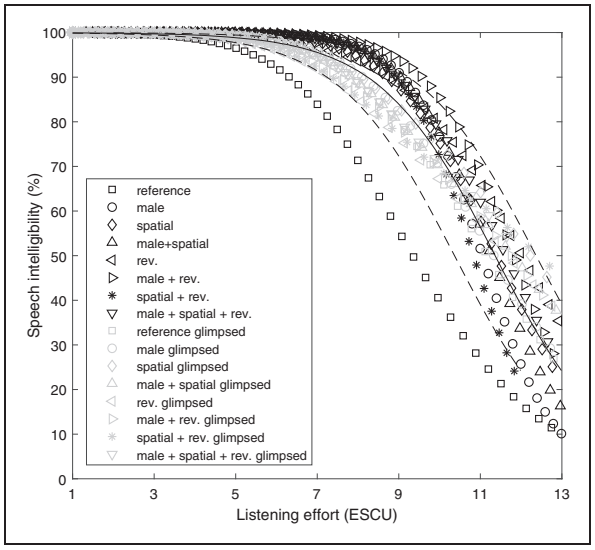


**Figure 7.** Speech intelligibility plotted against listening effort for all conditions. Each data point corresponds to the group values of listening effort and speech intelligibility at a specific SNR between −30 and 30 dB. The solid line is a sigmoid function fitted to the data, while the dashed lines illustrate the same sigmoid shifted horizontally by ±1 ESCU.
ESCU = effort scaling categorical unit.

Another way of comparing speech intelligibility and listening effort measurements is to consider individual data. To test if individual speech recognition performance was correlated with individual listening effort ratings, correlations between SNRs measured at the midpoint of the psychometric functions (i.e., 50% correct and 7 ESCU, respectively) were computed for each experimental condition. It was found that these SNRs were only significantly correlated for spatial+reversed maskers ($R = .67$ for natural masked speech, $R = .72$ for glimpsed speech), but not for any other masking condition. This indicates that there was not a strong systematic relation between individual results in speech intelligibility and listening effort. Similarly, there was no significant relation between the individual SRT benefits and the individual SNR benefits relative to the 7 ESCU point in any of the measurement conditions, indicating that the consistency in the rank order of conditions between speech intelligibility and listening effort observed at a group level was not pronounced at the level of individual data.

## Discussion

### *Informational and Energetic Masking in Speech Intelligibility*

One goal in the design of the experimental conditions of this study was to create a reference condition comprising a high degree of IM, thereby providing a large range over which to compare the efficacy of different isolated or combined unmasking cues. Among the conditions tested here, the reference condition had the highest SRT (−2.9 dB SNR) by a large margin. This value is very similar to the SRT of 0 dB TMR (≈−3 dB SNR) reported by Kidd et al. (2016) for a different group of

subjects and speech materials, but for essentially the same reference condition. Roughly equivalent SRTs for similarly defined reference conditions also have been reported for other kinds of closed-set speech material (e.g., Brungart et al., 2006; Ellinger, Jakine, & Gallun, 2017). The present results support the conclusion of previous studies that under such challenging listening conditions, subjects are able to understand, on average, 50% or more of the target speech only when the target level exceeds the level of the masking talkers. The observation that listeners must rely on a relative level cue to segregate talkers becomes particularly important when considering the effort involved in solving the task because simply attending to the louder talker apparently requires less effort than exploiting other cues (e.g., Figure 7, further discussion later). Furthermore, the reference condition produced the largest amount of *additional masking* (19 dB), that is, the largest difference between natural and glimpsed speech, consistent with the view that it caused the most IM. For all conditions that included at least one unmasking cue, the additional masking that was observed was roughly similar in magnitude (about 4–8 dB).[1]

The present findings are in line with several previous studies in that a considerable SRT reduction was observed for each of the unmasking cues tested in isolation. The unmasking that was observed was rather similar for these three cues (14–17 dB). Some studies also found similar amounts of masking release across cues (e.g., Kidd et al., 2016; Swaminathan et al., 2015), while other studies have reported different amounts of masking for different cues (e.g., Freyman et al., 2001; Zekveld et al., 2014). The reasons for the differences in masking release across studies are not entirely clear but may be due partly to different spatial separation methods, different speech materials, and different subject samples, especially given the relatively large individual differences typically found for conditions high in IM. Consistent with this observation, the ranges of individual SRT benefits (minimum to maximum) found in this study were 4.4 to 27.8 dB, 3.6 to 26.4 dB, and 7.6 to 24.0 dB for spatial, male, and reversed maskers, respectively. These ranges are comparable with those reported by Kidd et al. (2016), and, with such large ranges across individuals, it seems plausible that group mean differences of several decibels would be expected when comparing results from small groups of subjects. Another factor contributing to these differences in group mean thresholds could be that the present study employed the same target talker and the same pairs of masking talkers (as did Zekveld et al., 2014), while Kidd et al. (2016) and Xia et al. (2015) used target and masking talkers randomly selected from the set of talkers on each trial. It is possible that the reduced uncertainty associated with constant target and masker voices decreased the unmasking benefit found here.

The current study extended the work of Kidd et al. (2016) by incorporating all combinations of the three unmasking cues. Our findings suggested that a consistent additional decrease in SRT could occur when combining more than one unmasking cue. The additional benefit clearly was subadditive, that is, the reduction achieved by adding a second cue to a single unmasking cue was much smaller than its primary unmasking effect. This is in line with other studies that combined unmasking cues in similar SOS conditions (e.g., Swaminathan et al., 2015). Again, there were large differences between studies as to the additional benefit of a second unmasking cue, and not all studies have found that adding a second cue provided an additional benefit (e.g., Freyman et al., 2001; Zekveld et al., 2014).

In addition to differences in the methods, subjects, or precise listening conditions, another factor that makes comparisons between this study and previous studies difficult is that the contribution of EM (e.g., the extent to which EM dominates masking) was not always estimated. In this study, this estimation was obtained by including glimpsed stimuli as a control and by considering the differences in target energy retained after glimpsing at each ear. The monaural glimpse energy analysis supported several important observations: First, it was shown that significant energetic differences between stimuli/conditions may still exist (cf. Figure 3) even when attempting to minimize these differences by equalizing the long-term spectra of all maskers and conditions. This was especially pronounced for time-reversed maskers in this study (see later). Second, the variations in SRTs across conditions for the glimpsed speech could be explained largely by differences in the proportion of target energy retained, with the exception being those cases containing a spatial separation cue (see Figure 2). In contrast, for three out of the four conditions containing the spatial separation cue with glimpsed speech, SRTs were significantly lower than in the (glimpsed) reference condition. This residual benefit raises the possibility that the spatial information in the masker energy that was retained in the target-dominated tiles provided an advantage that could be exploited by the auditory system. A third conclusion to be drawn from the glimpse analysis is that, for natural masked speech, the significance of SRT differences in the conditions with at least one unmasking cue remained after subtracting the (monaural) energetic advantages. This suggests that monaural glimpse energy cannot fully explain the differences in unmasking effects across conditions. Fourth, the lower SRTs observed in conditions with spatially separated maskers were hardly affected by energetic advantages in the (monaural) glimpses (see black solid line in Figure 2). Any energetic advantage present in these conditions would hence have to originate from binaural processing, which is not captured in the proportion of

retained energy at either ear. A simplified model process of binaural processing is better-ear glimpsing. An estimate of an assumed optimal better-ear switching for each T-F tile showed that such a process could partly account for the gap between observed SRT differences and energetic differences expected from monaural glimpsing, but only in the order of about 1 dB.

One particular factor affecting intelligibility in the presence of reversed maskers is that explicit masker confusions cannot occur. Previous studies indicated that an analysis of response errors can help to disentangle the effects caused by EM and IM. In speech identification tasks with high IM, subjects tend to report the words uttered by the masker talkers rather than by the target talker when errors are made (Brungart, 2001; Kidd, Mason, Arbogast, & Mason, 2005; Ihlefeld & Shinn-Cunningham, 2008; Kidd et al., 2016; Wightman & Kistler, 2005). In contrast, the errors that occur for tasks dominated by EM tend to be randomly distributed. In the present study, the error patterns observed fit this expectation: For all conditions with unintelligible maskers (reversed maskers or glimpsed speech), the probability of masker confusions was near chance (19%). In all conditions with intelligible maskers, a considerable proportion of confusions occurred, which was highest for the reference condition (84%). The proportion of omissions (i.e., subjects deciding not to indicate a target word) was also reduced somewhat in the reference condition (24%) compared with the other conditions (30%–42%). This suggests that, when in doubt, subjects decided to guess more frequently in the high-IM reference condition than in the other conditions and, consequently, were more likely to report a word uttered by one of the maskers. This should be discussed in the context of the assumed chance level during the fitting of the psychometric functions. Setting this chance level was not straightforward here because subjects were not forced to guess, that is, the true chance level was likely between 0% (no guessing) and 10% (always guessing) and also depended on the individual listener's tendency to guess. To estimate how this may have affected the present data, we computed differences in $SRT_{50}$ between psychometric functions with 0% (as for the data reported earlier) and 10% chance level. The latter produced about 1.6-dB lower $SRT_{50}$ values, and this difference was very similar for all conditions. Consequently, derived $SRT_{50}$ differences between conditions were largely unchanged (mean absolute difference <0.1 dB). However, we cannot rule out that the "true" chance level differed between conditions, especially because the error pattern analysis indicated that the proportion of guesses was somewhat larger in the reference condition than in the other conditions (in which it was similar, see Figure 6). It is hence possible that the amount of unmasking reported here depended slightly on the assumed chance level. Because the number of omissions was similar for all conditions with at least one unmasking cue, the comparison between unmasking conditions would not have been strongly affected.

The analysis of glimpse energy (Figure 3) indicated that an increased number of glimpses in the first 300 ms, approximately, was available for reversed maskers compared with forward. It is likely that the method for time reversing the maskers used in the present study was responsible for the differences between forward and reversed maskers found here. In contrast to the method used by Kidd et al. (2016) in which individual words drawn from a matrix of words were concatenated to form sentences, the speech materials used as target and maskers in this study were recorded as naturally produced sentences. This meant that the decline in level that is characteristic of naturally produced single sentences that occurs toward the end of the sentences was present in these stimuli, although care was taken during the recordings to minimize this effect (Hochmuth et al., 2018). Because the stimuli were scaled to equal root-mean-square during presentation, the relative level of the words at the end of the sentence was slightly lower than at the beginning. When the entire masker sentence was reversed, the lower level words were superimposed on the initial words of the target sentence producing more target-dominated glimpses. In combination with the often salient onsets of the names in the target sentences, it seems reasonable that a decrease in masker level due to time reversal would produce better intelligibility at the beginning of the sentence (i.e., the first word). One practical consequence of the rather strong energetic effect of masker time reversal is that glimpses (or other suitable measures) should be analyzed and reported to facilitate comparison between studies especially those that use masker time reversal as a cue.

### *Informational and Energetic Masking in Listening Effort*

One goal of this study was to extend the SOS paradigm employed in previous studies of speech intelligibility to the measurement of listening effort. To that end, listening effort was investigated over a large range of SNRs including those where intelligibility was at ceiling. To the best of our knowledge, listening effort has not been examined in this way previously for high-IM SOS masking conditions or for the corresponding glimpsed speech conditions.

Listening effort was systematically lower when one or more unmasking cues were provided than in the reference condition at a given SNR. For all combinations of cues, this benefit was largest at low SNRs but was reduced or eliminated at high SNRs. The dependence of unmasking benefit on SNR was also observed for

spatial unmasking of speech in stationary noise (Rennies & Kidd, 2018) and seems intuitive given that, at a high enough SNR, listening effort will be dominated by SNR and become independent of other unmasking cues. However, the release from listening effort relative to the reference condition extended well into the range of positive SNRs (see Figure 6), and it is important to point out that these effects were unlikely to have been affected by energetic advantages as was observed for the speech intelligibility data. As suggested by the analysis of available glimpse energy (see Figure 3), the differences between conditions were very small at positive SNRs. In this range, the significant differences between conditions were presumably related to a reduced effort required to suppress the masker talkers rather than to energetic effects. The amount of release from listening effort depended on the unmasking cue: It was larger for spatial maskers (equivalent to a 5–8 dB SNR change) than for male or reversed maskers (equivalent to a 3–5 dB SNR change) and was generally smaller for isolated than for combined unmasking cues, and it reached an equivalent change in SNR of 10 dB or more for some conditions (see Figure 6). As stated earlier, the release from listening effort was even larger at lower SNRs but was affected more by energetic differences (e.g., a 3-dB advantage would be expected for reversed maskers at a reference SNR of −10 dB, see Figure 3).

It should be noted that previous studies of listening effort in SOS masking conditions in which one or more unmasking cues were provided typically made measurements at lower (usually negative) SNRs. Zekveld et al. (2014) quantified listening effort using pupillometry for SNRs converging to the SRT. They reported that a gender difference reduced listening effort, while spatial separation of the masker did not. This seems at odds with the present findings, where the benefit from spatial separation was 2 to 3 dB larger than (but not statistically different from) the benefit due to gender differences. As pointed out by Xia et al. (2015), the results of Zekveld et al. (2014) likely were affected by differences in the SNR at which listening effort was measured. Xia et al. (2015) found that listening effort, as quantified by a performance cost in a secondary task, was lower for spatially separated maskers than for maskers differing in gender, even though speech recognition scores were similar. They argued that location-based speech segregation is less cognitively demanding than gender-based speech segregation. This seems to be in line with the conclusion of Zekveld et al. (2014) that "one possible interpretation is that spatial separation eases speech understanding at a more peripheral level of processing, perhaps subcortical, whereas voice cues have to be dealt with at the cortical level by using top-down processing" (p. 8). In light of this discussion, it is interesting to consider the findings of the present study from the point of view of listening effort as it relates to the task of segregating the target from the maskers. In the extreme case of glimpsed speech as employed here, the segregation of sources has been performed "for the subject," that is, no effort is expended for segregation itself. Instead, the perceived effort could be due to reassembling the target glimpses into a coherent stream of speech and to interpreting the message the speech conveyed. This could be an effortful listening task as indicated by the high effort ratings at low SNRs when only a few target glimpses are available. Compared with natural masked speech, however, listening effort was always considerably reduced for glimpsed speech at the same SNR. In other words, keeping the accessible target glimpses the same, but removing the surrounding masker significantly reduced listening effort. This effect was equivalent to an improvement in SNR of between 8 and 18 dB (see Figure 6) and parallels the *additional masking* derived from the SRTs. Interestingly, this effect was not larger in the reference condition than in the other conditions as was found in the SRT data. This may indicate that the effort associated with streaming and reassembling the target glimpses was similar across conditions. Along this line, the differences in listening effort observed between conditions of natural masked speech should then be due primarily to differences related to target-from-masker segregation.

One interesting trend observed in the present results was that the natural cue benefit was larger and the additional glimpsing benefit was smaller when spatial masker separation was included than for the other cues in conditions with colocated maskers. Both of these differences were about 4 dB when averaged across conditions and reference SNRs (see Figure 6) and could indicate that spatial separation is a relatively strong segregation cue, which may be rather low level and require fewer cognitive resources (see Xia et al., 2015; Zekveld et al., 2014) and which leaves less room for additional segregation benefit via ITFS. In contrast, for reversed maskers, the additional glimpsing benefit was relatively high (and the unmasking effect for natural masked speech was relatively low), which may indicate that masker time reversal is a rather weak segregation cue. It should be pointed out that the glimpsing benefit did not differ significantly between conditions (except for a single comparison).

With respect to the individual data, large interindividual differences were observed for some conditions when comparing the benefit of a specific unmasking cue to the corresponding additional glimpsing benefit. For example, one subject rated listening effort essentially the same for spatially separated maskers in natural masked speech and in glimpsed speech, indicating that removing the masker-dominated tiles did not produce an

additional benefit or, in other words, that spatial separation produced the "full amount" of target segregation for this subject. In contrast, another subject did not benefit at all from masker gender difference or time reversal, but benefited strongly from glimpsing, indicating very weak segregation benefit from these cues. These large interindividual differences indicate that individual subject effects should be considered and sufficiently large subject groups should be included when measuring listening effort in SOS masking conditions. One way to further explore the specific effort required for segregating target speech from interfering talkers could be to systematically vary the task difficulty. This was tested by Brungart et al. (2013) who found that, when the SNR was adjusted to equalize performance in an easy task (e.g., speech detection) for a speech masker and a noise masker, performance dropped more rapidly in a more demanding task (e.g., speech identification) for speech maskers than for noise maskers. Brungart et al. (2013) argued that, in the simple task, subjects were able to perform equally well in speech and noise maskers because they invested more resources for the segregation (which is much less effortful in noise than in the presence of speech masker). As task difficulty increased, cognitive resources had to be redeployed to solve the task, resulting in a larger performance drop in speech maskers than in noise maskers. Employing similar methods for the conditions of the present study could shed further light on the effort related to segregating target speech from other talkers and the role of the different isolated and combined unmasking cues in such conditions. Furthermore, applying alternative methods for measuring listening effort which do not rely on subjective rating, such as electroencephalography or pupillometry, would be a useful way to validate the present data and further investigate the role of different unmaking cues in SOS conditions.

### *Relation Between Speech Intelligibility and Listening Effort*

The present data allowed us to compare speech intelligibility and listening effort measured in the same subjects for a variety of SOS masking conditions. The functions of speech intelligibility versus listening effort data measured at the same SNRs (Figure 7) showed a similar relation as was observed previously for speech-in-noise conditions (Rennies & Kidd, 2018), that is, most data points were within ±1 ESCU (i.e., one category on the 13-point scale) of a sigmoidal function fitted to the data in the range where both intelligibility and effort were below ceiling. Another similarity to the data of Rennies and Kidd (2018) was the saturated speech intelligibility up to about 5 to 8 ESCU where listening effort varied but intelligibility remained at ceiling. There were, however, two notable differences in the conditions of the present study. The most obvious was the outlying curve for the high-IM reference condition, which showed consistently lower effort ratings for the same performance. One possible reason is that the reference condition has a stronger relative level cue that is available (because SNRs were higher in general at a fixed performance level). When the target level is close to (or higher than) the level of the maskers, less effort may be required in exploiting other cues such as gender difference. An additional interpretation is that, at a given performance level for speech intelligibility, increased listening effort must be expended when exploiting the available unmasking cues to compensate for the lower SNR. Another systematic difference between conditions notable in this figure was that glimpsed speech always required a somewhat lower listening effort than natural masked speech at a fixed level of performance for intelligibility at intermediate listening effort ratings. This supports the view that removing the need to segregate the target from the maskers reduced listening effort at equal, moderate levels of intelligibility.

Another way of comparing intelligibility and listening effort is to consider the rank order of conditions. Based on the group psychometric functions, it was found that conditions were ordered in the same way in both experiments, that is, listening effort decreased at a given SNR in the order reference—single unmasking cue—two unmasking cues—three unmasking cues, while speech intelligibility increased in the same order. One interesting part of the psychometric function for listening effort is the SNR range at which speech intelligibility is at ceiling. To identify this range, the SNR required to achieve 90% speech intelligibility was extracted from the psychometric function of each condition as a measure of performance at ceiling: These were +8.9 dB for the reference condition, between −16.1 and −4.1 dB for the other natural masked speech conditions, and below −11.5 dB for all glimpsed speech conditions. Thus, high/ceiling speech intelligibility could be assumed at the two highest SNRs at which the benefit in listening effort was evaluated (10 and 15 dB). At these SNRs, all natural unmasking cues provided a reduction in listening effort. Interestingly, this benefit was generally larger when unmasking cues were combined than it was for isolated cues. The increased listening effort benefit for conditions with spatially separated maskers is in agreement with data of Rennies and Kidd (2018), who reported a significant spatial release from listening effort for speech masked by stationary noise. This benefit extended well into the range of high SNRs, at which speech intelligibility was at ceiling. For the glimpsed speech measured in this study, the listening effort benefit was even stronger compared with natural masked speech: The tested SNRs were more than 20 to 30 dB above the 90%-correct point of glimpsed speech, so clearly speech intelligibility was

optimal in these conditions. Future work should include different types of speech material, in particular open-set material, for which ceiling performance may be reached at higher SNRs than for closed-set matrix sentences.

Finally, it is interesting to compare speech intelligibility and listening effort on an individual subject basis. In this study, this was done by correlating both the individual SNRs at a fixed performance level with listening effort rating, as well as correlating the individual SNR benefits in intelligibility with listening effort across conditions. In both cases, no significant correlations were found. The fact that this also was true for SNR benefits (i.e., an individual measure minimizing the effect of response bias by subtracting performance measures in the reference conditions) supports the conclusion that the lack of correlation was not an artifact due to subjects using the listening effort scale in different ways. One possible interpretation is that subjects benefitted differently from the different unmasking cues in terms of intelligibility and effort, even though the individual benefit was rather consistent within each of the two measurements. One important practical implication is that the individual benefit from source segregation/unmasking cues in SOS conditions at high SNRs (as measured by listening effort when intelligibility is at ceiling) cannot be predicted by the individual benefit in speech intelligibility. This could give further impetus for using listening effort assessment to gain insights into natural listening situations that goes beyond what may be learned by assessing speech intelligibility alone.

## Conclusions

The following conclusions can be drawn from the present findings:

1. All three unmasking cues that were tested (masker gender, spatial separation, time reversal), as well as their combinations, improved speech intelligibility. Combined unmasking cues produced a larger SRT reduction than any of the unmasking cues in isolation even after correcting for differences in the available glimpses of target energy or (energetic) spatial unmasking. This suggests that SOS segregation can benefit from adding unmasking cues.
2. Similarly, all unmasking cues also reduced listening effort. This reduction was larger when the listeners were provided with more than one unmasking cue. The added benefit of combined cues likely reflects a release from IM because the calculated energetic differences between conditions were minimal. The benefit in listening effort extended well into the range of SNRs at which speech intelligibility was at ceiling. This suggests that listening effort is a useful tool for evaluating SOS masking conditions at typical conversational levels.
3. Glimpsed speech produced significantly lower SRTs than natural masked speech, confirming previous studies that a large amount of IM was present in the measured natural conditions.
4. In listening effort, the benefit due to glimpsed speech versus natural masked speech was even larger than for intelligibility (equivalent to between 8 and 18 dB SNR), indicating that a large amount of listening effort is associated with segregating the target speech from the maskers.
5. The relation between speech intelligibility and listening effort as derived from the group psychometric functions was very similar across all combinations of unmasking cues when neither listening effort nor intelligibility were at ceiling, that is, the same speech recognition performance corresponded to a similar amount of perceived effort. The only exception was the high-IM reference condition, which may be due to a relatively stronger level cue, which may require less effort than exploiting the other unmasking cues.
6. The individual benefit in intelligibility due to any of the tested unmasking cues was not correlated with the individual benefit in listening effort, suggesting that measuring listening effort can also provide insights into individual speech perception which are not captured by speech intelligibility measurements.

### Acknowledgments

The authors thank HörTech gGmbH and Sabine Hochmuth for providing the speech material.

### 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 disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the German Research Foundation (DFG; grant number RE 4160/1-1) and the National Institutes of Health—National Institute on Deafness and Other Communication Disorders (grant number R01 DC04545).

### ORCID iD

Jan Rennies https://orcid.org/0000-0002-0291-7723

### Note

1. It should be noted that the glimpsed speech was presented at a higher level in the present study than in the study of Kidd et al. (2016). This was a result of the rescaling to an overall presentation level of 70 dB SPL, which was done to keep the overall loudness similar across conditions. It is possible that this rescaling increased the audibility of very soft glimpses,

although it seems likely that the glimpses would have been above absolute threshold (even without the scaling) also at the lowest speech levels included in this study (about 40 dB SPL, i.e., at −30 dB SNR).

## References


Andéol, G., Suied, C., Scanella, S., & Dehais, F. (2017). The spatial release of cognitive load in cocktail party is determined by the relative levels of the talkers. *Journal of the Association for Research in Otolaryngology*, *18*, 457–464. doi:10.1007/s10162-016-0611-7

Best, V., Mason, C. R., & Kidd, G., Jr. (2011). Spatial release from masking as a function of the temporal overlap of competing maskers, *J. Acoust. Soc. Am. 129*, 1616–1625.

Best, V., Marrone, N., Mason, C. R., & Kidd, G. Jr. (2012). The influence of non-spatial factors on measures of spatial release from masking. *Journal of the Acoustical Society of America*, *131*, 3103–3110. doi: 10.1121/1.3693656.

Brand, T., & Kollmeier, B. (2002). Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *Journal of the Acoustical Society of America*, *111*, 2801–2810. doi/10.1121/1.1479152.

Broadbent, D. E. (1952). Listening to one of two synchronous messages. *The Journal of Experimental Psychology*, *44*, 51–55. doi 10.1037/h0056491.

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

Bronkhorst, A. W. (2015). The cocktail-party problem revisited: Early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics*, *77*, 1465–1487. doi: 10.3758/s13414-015-0882-9.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *Journal of the Acoustical Society of America*, *109*, 1101–1109. doi: 10.1121/1.1408946.

Brungart, D. S., Chang, P. S., Simpson, B. D., & Wang, D. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *Journal of the Acoustical Society of America*, *120*, 4007–4018. doi: 10.1121/1.2363929.

Brungart, D. S., Chang, P. S., Simpson, B. D., & Wang, D. (2009). Multitalker speech perception with ideal time-frequency segregation: Effects of voice characteristics and number of talkers. *Journal of the Acoustical Society of America*, *125*, 4006–4022. doi: 10.1121/1.3117686.

Brungart, D. S., Iyer, N., Thompson, E. R., Simpson, B. D., Gordon-Salant, S., Schurman, J., . . . Grant, K. (2013). Interactions between listening effort and masker type on the energetic and informational masking of speech stimuli. *Proceedings of Meetings on Acoustics*, *19*, 060146. doi: 10.1121/1.4800033.

Carlile, S. (2014). Active listening: Speech intelligibility in noisy environments. *Acoustics Australia*, *42*, 98–104.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, *25*, 975–979. doi: 10.1121/1.1907229.

Ellinger, R. L., Jakine, K. M., & Gallun, F. J. (2017). The role of interaural differences on speech intelligibility in complex multi-talker environments. *Journal of the Acoustical Society of America*, *141*, EL170–EL176. doi: 10.1121/1.4976113.

Ewert, S. D. (2013). AFC – A modular framework for running psychoacoustic experiments and computational perception models. In German Acoustical Society (DEGA) (Ed.), *Proceedings of the International Conference on Acoustics AIA-DAGA 2013* (pp. 1326–1329). Merano, Italy: German Acoustical Society (DEGA).

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2001). Spatial release from informational masking in speech recognition. *Journal of the Acoustical Society of America*, *109*, 2112–2122. doi: 10.1121/1.1354984.

Gallun, F. J., Kampel, S. D., Diedesch, A. C., & Jakien, K. M. (2013). Independent impacts of age and hearing loss on spatial release in a complex auditory environment. *Frontiers in Neuroscience*, *252*, 1–11. doi: 10.3389/fnins.2013.00252.

Hochmuth, S., Kollmeier, B., & Shinn-Cunningham, B. (2018). The relation between acoustic-phonetic properties and speech intelligibility in noise across languages and talkers. In German Acoustical Society (DEGA) (Ed.), *Proceedings of the German on Acoustics DAGA 2018* (pp. 628–629). Munich, Germany: German Acoustical Society (DEGA).

Houben, R., van Doorn-Bierman, M., & Dreschler, W. (2013). Using response time as a measure of listening effort. *International Journal of Audiology*, *52*, 753–761. doi: 10.3109/14992027.2013.832415.

Ihlefeld, A., & Shinn-Cunningham, B. (2008). Spatial release from energetic and informational masking in a selective speech identification task. *Journal of the Acoustical Society of America*, *123*, 4369–4379. doi: 10.1121/1.2904826.

Iyer, N., Brungart, D. S., & Simpson, B. D. (2010). Effects of target-masker contextual similarity on the multimasker penalty in a three-talker diotic listening task. *Journal of the Acoustical Society of America*, *128*, 2998–3010. doi: 10.1121/1.3479547.

Kayser, H., Ewert, S. D., Anemüller, J., Rohdenburg, T., Hohmann, V., & Kolmeier, B. (2009). Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses. *EURASIP Journal on Advances in Signal Processing*, 2009, (1), Article ID298605. doi: 10.1155/2009/298605.

Kidd, G., Jr., & Colburn, H. S. (2017). Informational masking in speech recognition. In J. C. Middlebrooks et al. (Eds.), *The auditory system at the cocktail party (Springer Handbook of Auditory Research 60)* (pp. 75–109). Springer International Publishing AG. doi:10.1007/978-3-319-51662-2_4, 75-109

Kidd, G. Jr, Mason, C. R., Arbogast, T. L., & Mason, C. R. (2005). The advantage of knowing where to listen. *Journal of the Acoustical Society of America*, *118*, 3804–3815. doi: 10.1121/1.2109187.

Kidd, G., Jr., Mason, C. R., Richards, V. M., Gallun, F. J., and Durlach, N. I. (2008a). Informational masking, in *Auditory Perception of Sound Sources*, edited by W. A. Yost, A. N. Popper, and R. R. Fay (Springer, New York), pp. 143–190.

Kidd, G., Jr., Best, V., and Mason, C. R. (2008b). Listening to every other word: Examining the strength of linkage

variables in forming streams of speech, *J. Acoust. Soc. Am. 124*, 3793–3802.

Kidd, G. Jr, Mason, C. R., Swaminathan, J., Roverud, E., Clayton, K. K., & Best, V. (2016). Determining the energetic and informationa components of speech-on-speech masking. *Journal of the Acoustical Society of America*, *140*, 132–144. doi: 10.1121/1.4954748.

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M. A., Uslar, V., Brand, T., & Wagener, K. C. (2015). The multilingual matrix test: Principles, applications, and comparison across languages: A review. *International Journal of Audiology*, *54*, 3–16. doi: 10.3109/14992027.2015.1020971.

Krueger, M., Schulte, M., Brand, T., & Holube, I. (2017a). Development of an adaptive scaling method for subjective listening effort. *Journal of the Acoustical Society of America*, *141*, 4680–4693. doi: 10.1121/1.4986938.

Krueger, M., Schulte, M., Zokoll, M. A., Wagener, K. C., Meis, M., Brand, T., & Holube, I. (2017b). Relation between listening effort and speech intelligibility in noise. *American Journal of Audiology*, *26*, 378–392. doi: 10.1044/2017_AJA-16-0136.

Marrone, N. L., Mason, C. R., & Kidd, G. Jr. (2008). Tuning in the spatial dimension: Evidence from a masked speech identification task. *Journal of the Acoustical Society of America*, *124*, 1146–1158. doi: 10.1121/1.2945710.

Middlebrooks, J., Simon, J. Z., Popper, A. N., & Fay, R. R. (Eds.). (2017). *The auditory system at the cocktail party*. *Springer Handbook of Auditory Research*. New York, NY: Springer International Publishing.

Morimoto, M., Sato, H., & Kobayashi, M. (2004). Listening difficulty as a subjective measure for evaluation of speech transmission performance in public spaces. *Journal of the Acoustical Society of America*, *116*, 1607–1613. doi: 10.1121/1.1775276.

Neff, D. L., & Dethlefs, T. M. (1995). Individual differences in simultaneous masking with random-frequency, multicomponent maskers. *Journal of the Acoustical Society of America*, *98*, 125–134.

Noble, W., & Perrett, S. (2002). Hearing speech against spatially separate competing speech versus competing noise. *Perception & Psychophysics*, *64*, 1325–1336.

Peissig, J., & Kollmeier, B. (1997). Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners. *Journal of the Acoustical Society of America*, *101*, 1660–1670.

Pichora-Fuller, K. M., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E.,... Wingfield, A. (2016). Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear and Hearing*, *37*, 5S–27S. doi 10.1097/AUD.0000000000000312.

Rennies, J., & Kidd, G. Jr. (2018). Benefit of binaural listening as revealed by speech intelligibility and listening effort. *Journal of the Acoustical Society of America*, *144*, 2147–2159. doi: 10.1121/1.5057114.

Rennies, J., Schepker, H., Holube, I., & Kollmeier, B. (2014). Listening effort and speech intelligibility in listening situations affected by noise and reverberation. *Journal of the Acoustical Society of America*, *136*, 2642–2653. doi: 10.1121/1.4897398.

Sato, H., Morimoto, M., & Wada, M. (2012). Relationship between listening difficulty rating and objective measures in reverberant and noisy sound fields for young adults and elderly persons. *Journal of the Acoustical Society of America*, *131*, 4596–4605. doi: 10.1121/1.4714790.

Schepker, H., Haeder, K., Rennies, J., & Holube, I. (2016). Perceived listening effort and speech intelligibility in reverberation and noise for hearing-impaired listeners. *International Journal of Audiology*, *55*, 738–747. doi: 10.1080/14992027.2016.1219774.

Smeds, K., Wolters, F., & Rung, M. (2015). Estimation of signal-to-noise ratios in realistic sound scenarios. *American Journal of Audiology*, *26*, 183–196. doi: 10.3766/jaaa.26.2.7.

Swaminathan, J., Mason, C. R., Streeter, T. M., Best, V. A., Kidd, G. Jr., & Pate, A. D. (2015). Musical training, individual differences and the cocktail party problem. *Scientific Reports*, *5*, 1–10. doi: 10.1038/srep11628.

Wagener, K., Brand, T., & Kollmeier, B. (1999). Entwicklung und evaluation eines Satztests für die deutsche Sprache III: Evaluation des Oldenburger Satztests [Development and evaluation of a German sentence test III: Evaluation of the Oldenburg sentence test]. *Zeitschrift für Audiologie*, *38*, 86–95.

Wightman, F. L., & Kistler, D. J. (2005). Informational masking of speech in children: Effects of ipsilateral and contralateral distractors. *Journal of the Acoustical Society of America*, *118*, 3164–3176. doi: 10.1121/1.2082567.

Xia, J., Noorale, N., Kalluri, S., & Edwards, B. (2015). Spatial release of cognitive load measured in a dual-task paradigm in normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, *137*, 1888–1898. doi: 10.1121/1.4916599.

Yost, W. A. (1997). The cocktail party problem: Forty years later. In R. H. Gilkey, & T. R. Anderson (Eds.), *Binaural and spatial hearing in real and virtual environments* (pp. 329–347). Hillsdale, NJ: Erlbaum.

Zekveld, A., Rudner, M., Kramer, S. E., Lyzenga, J., & Rönnberg, J. (2014). Cognitive processing load during listening is reduced more by decreasing voice similarity than by increasing spatial separation between target and masker speech. *Frontiers in Neuroscience*, *8*, 88. doi 10.3389/fnins.2014.00088.