Research Article

Binaural Glimpses at the Cocktail Party?

ANDREA LINGNER,¹ BENEDIKT GROTHE,¹ LUTZ WIEGREBE,¹ AND STEPHAN D. EWERT²

¹*Division of Neurobiology, Department Biology II, Ludwig-Maximilians-Universität Munich, Großhaderner Str. 2-4, 82152, Martinsried-Planegg, Germany*

²*Medizinische Physik and Cluster of Excellence Hearing4All, Universität Oldenburg, Carl-von-Ossietzky Str. 9-11, 26129, Oldenburg, Germany*

Received: 27 March 2015; Accepted: 21 June 2016; Online publication: 13 July 2016

ABSTRACT

Humans often have to focus on a single target sound while ignoring competing maskers in everyday situations. In such conditions, speech intelligibility (SI) is improved when a target speaker is spatially separated from a masker (spatial release from masking, SRM) compared to situations where both are co-located. Such asymmetric spatial configurations lead to a ‘better-ear effect’ with improved signal-to-noise ratio (SNR) at one ear. However, maskers often surround the listener leading to more symmetric configurations where better-ear effects are absent in a long-term, wideband sense. Nevertheless, better-ear glimpses distributed across time and frequency persist and were suggested to account for SRM (Brungart and Iyer 2012). Here, speech reception was assessed using symmetric masker configurations while varying the spatio-temporal distribution of potential better-ear glimpses. Listeners were presented with a frontal target and eight single-talker maskers in four different symmetrical spatial configurations. Compared to the reference condition with co-located target and maskers, an SRM of up to 6 dB was observed. The SRM persisted when the frequency range of the maskers above or below 1500 Hz was replaced with stationary speech-shaped noise. Comparison to a recent short-time binaural SI model showed that better-ear glimpses can account for half the observed SRM, while binaural interaction utilizing phase differences is required to explain the other half.

Keywords: better-ear listening, glimpsing, release from masking, speech intelligibility model, speech reception thresholds

INTRODUCTION

Everyday, our auditory system is challenged to analyse relevant sounds in the background of many irrelevant sounds, as often referred to as cocktail party processing (Cherry 1953) where many talkers are simultaneously active while a listener tries to comprehend only one of them. Normal-hearing humans are surprisingly good at this. An important factor affecting speech intelligibility (SI) in such situations is the spatial relation between target and masking signals (speech or noise). Several studies have shown that a separation of target and masker in azimuth leads to an improvement in SI when compared to a co-located target and masker (reviewed by Bronkhorst 2000). If the masker is noise, this improvement can mainly be explained by the so-called better-ear effect. It refers to the fact that the signal-to-noise ratio (SNR) for a frontal target is higher at the ear opposite to a masker from either side, caused by the sound shadowing effect of the head. Better-ear effects can lead to an improvement in SI of 8–10 dB (e.g. Bronkhorst and Plomp 1988; Freyman et al. 1999). This SI improvement due to spatial separation of target and masker is even higher when speech instead of noise is used as masker (Freyman et al. 1999) suggesting that listeners make use of factors other than the better-ear effect to improve SI with speech maskers. Moreover, maskers are often not only present to one side of the listener

Correspondence to: Stephan D. Ewert · Medizinische Physik and Cluster of Excellence Hearing4All · Universität Oldenburg · Carl-von-Ossietzky Str. 9-11, 26129, Oldenburg, Germany. Telephone: +49 441 798-3900; fax: +49 441 798-3902; email: stephan.ewert@uni-oldenburg.de

but are spatially distributed. For a uniform spatial distribution and a frontal target, an equal target-to-masker ratio (TMR) is observed at both ears of the listener and better-ear effects in a classical 'long-term' sense are absent. Accordingly, if SI is improved by an azimuthal separation of target and spatially distributed maskers, this improvement cannot be explained by the classical better-ear effect. Several studies have shown that when two symmetrically placed maskers were spatially separated from the target in azimuth, listeners showed an improvement of SI, i.e. spatial release from masking (SRM), between 5 and 12 dB depending on the spatial separation of target and masker (e.g. Marrone et al. 2008; Noble and Perrett 2002). One potential mechanism underlying SRM in these conditions is the binaural processing of the interaural level difference (ILD) and interaural time difference (ITD) as, e.g. described by the classic equalization-cancellation (EC) model (Durlach 1963). This model can account for binaural unmasking in psychoacoustics and was successfully combined with a speech intelligibility model (eSII; Rhebergen et al. 2006) in the (short-time) binaural speech intelligibility model (BSIM; Beutelmann and Brand 2006; Beutelmann et al. 2010). A conceptually similar approach was suggested by Lavandier and Culling (2010). The BSIM of Beutelmann et al. (2010) can make use of short-time, per-channel ITDs and ILDs to gain a binaurally improved TMR at the output of the EC stage. Recently, Wan et al. (2014) also suggested a short-time version of the EC model (STEC), additionally focusing on the advantages of binaural processing which adjusts to time-varying input waveforms. Common to both approaches is the assumption that the EC processing takes place independently in each of the time-frequency (T-F) segments and that the auditory system can selectively process these segments.

Another hypothesis for the mechanism underlying SRM is motivated from the field of computational auditory scene analysis and assumes that T-F segments in mixtures are dominated by signal power from one source. By means of object segregation (or with prior knowledge of the sources), an (ideal) binary mask can be constructed for two spatially separated sources, assigning the T-F segments to either source (e.g. Brown and Wang 2005). Brungart and Iyer (2012) have applied this scheme as a better-ear binary mask, selecting better-ear glimpses (only considering short-time, per-channel ILDs and disregarding ITDs). Their experimental results showed that SI in symmetric masker configurations can be explained by using an optimal binaural glimpsing strategy, suggesting that humans can determine the ear with the favourable TMR and moreover are able to rapidly switch between the two ears in order to maximize the SI in an optimal manner. This is in line with the assumption of

selective processing of these segments in BSIM and STEC. However, it should be noted that the ability to use the suggested glimpsing strategy strongly depends on the confusability of target and masker, i.e. informational masking. It was shown that the use of better-ear glimpses cannot fully explain SI if the task produces high informational masking (Best et al. 2015; Glyde et al. 2013). Moreover, Kidd et al. (2010) determined SI in a symmetric speech masker configuration similar to Marrone et al. (2008), when all signals were either unfiltered or filtered into low-, mid- or high-frequency regions. The authors showed that SRM was highest for the unfiltered broadband condition, but the filtered conditions lead to a reduced but still considerable SRM (between 5 and 7 dB), indicating that there may be other factors than better-ear glimpses influencing SI in a multi-talker environment.

Most of the studies introduced so far, and also other studies regarding SI in symmetric masker configurations (e.g. Jones and Litovsky 2011), limited the number of masking talkers or used noise maskers. In this case, the spectro-temporal structure of ITD and ILD variations in the masker is not necessarily comparable to more complex, spatially distributed multi-talker situations (e.g. Bronkhorst and Plomp 1992). Thus, it still remains an open question to what extent human listeners can use better-ear glimpses in a complex environment with multiple speech maskers. Moreover, it is unclear how SI depends on the spatial distribution of the sources and the resulting spectro-temporal distribution of ITDs and ILDs.

Here speech reception thresholds (SRTs) were measured in a paradigm mimicking a cocktail party environment with eight spatially distributed masking talkers. Using more than two masking talkers as have previously often been used allows for a specific trading of spatial auditory cues against glimpsing cues. This trading was implemented as follows: listeners were presented with a frontal speech target and eight single-talker speech maskers. The eight maskers were presented in five different spatial configurations. In order to assess the relative contribution of long-term masking effects and better-ear glimpses, maskers were presented in three different conditions: In a first condition, SRTs were determined with natural broadband speech maskers. In the second and third conditions, the low- or high-frequency parts of all maskers were replaced by stationary speech-shaped noise (SSN). These manipulations left long-term energetic masking effects unchanged but precluded the use of better-ear glimpses in the targeted frequency range. SI was compared to predictions of a state-of-the-art binaural SI model (Beutelmann et al. 2010) which allows for better-ear glimpses, and the contributions of individual model stages to overall SI were analysed.

MATERIAL AND METHODS

Psychophysics

Setup. Listeners were seated in a double-walled sound-attenuated and anechoic chamber (Industrial Acoustics Company GmbH, Niederkrüchten, Germany, lined on all surfaces with 20-cm acoustic foam wedges). Target and maskers were presented in the free sound field via a circular array of 36 speakers (CANTON Plus XS.2, CANTON Elektronik GmbH & Co. KG, Weilrod, Germany) mounted in the chamber. The distance between each loudspeaker and the listener's head was approximately 1 m. Every loudspeaker was equalized with an individual finite-impulse-response filter resulting in a flat frequency and linear phase response from 100 Hz to 20 kHz as measured with a 1/2-in. microphone (B&K 4189) positioned in the middle of the subjects' interaural axis. A touch screen connected to the computer was used to display the graphical user interface.

Stimulus Design. Three experimental conditions were tested in this study. All conditions consisted of a male target speaker and eight maskers (interfering speakers). For the target, sentences of the Oldenburger Satztest (OLSA; Wagener et al. 1999) were used. Maskers were continuous single-talker speech sequences derived from eight different audio books (three male and five female speakers) with a duration ranging from approximately 21 to 24 s. Gaps between words were limited to 100 ms. Each of the eight masker signals was presented at 47 dB SPL.

In the first condition of the experiment, the eight maskers were presented without any further modification (referred to as speech maskers). For the second condition, the frequency range above 1500 Hz of each masker was replaced by stationary, speech-shaped noise (LPS/HPSSN maskers). SSN was generated by calculating the average magnitude spectrum across time and all eight maskers and by using this average spectrum, together with eight different realizations of a random phase spectrum, to generate eight independent steady-state SSN maskers. For the third condition (HPS/LPSSN), the frequency range below 1500 Hz was replaced by SSN. The target sentences always remained unchanged.

The target was always presented from the frontal loudspeaker. The eight maskers were presented in five different in Figure 1. In the reference configuration (left-most panel; 0°), all maskers were co-located with the frontal target. In the 180° configuration, all maskers were co-located in the back. In the $\pm 90^\circ$, $\pm 40^\circ/\pm 140^\circ$ and distributed (distr.) configurations 4, 2 and 1 masker(s) were mapped to the 2, 4 and 8 directions, respectively. Thus, in all spatial configurations, all eight maskers were presented. All stimuli were D/A converted ($f_s = 44.1$ kHz, MOTU 24I/O, MOTU, Inc., Cambridge, MA, USA) and

amplified (NAD CI9120, NAD Electronics International, Ontario, Canada).

Experimental Procedure. SRTs for 50 % speech intelligibility were measured using the OLSA with an adaptive procedure to vary the SNR. The target sentences followed the structure name–verb–numeral–adjective–noun. Each of the five words of this sentence was randomly chosen from a list of ten words, resulting in a grammatically correct but semantically not predictable sentence. Listeners were required to indicate the words they understood by using a touch screen showing a matrix of all possible 5×10 words (closed test). No feedback was provided. After the listener confirmed his/her answer, the next sentence presentation was started. Listeners were instructed to always face the loudspeaker in the front (emitting the target). The maskers were switched on 1 s before the onset of the target sentence. A random part of each of the individual maskers was selected for each trial. Five SRT measurements were performed in an interleaved manner with a new spatial configuration occurring in random order for each presented target sentence. Each listener performed three OLSA runs; each run consisted of 20 target sentences per spatial configuration, i.e. an overall of 100 sentences per run. The whole experimental procedure and stimulus presentation was implemented using the Matlab AFC package (Ewert 2013).

Subjects. Ten human listeners (four males and six females, mean age 23 ± 1 years) participated in the experiments with masker condition 1. Six of these ten subjects (two males and four females) also participated in the experiment with the second masker condition. Ten different listeners (four males and six females, mean age 22.9 years) were recruited for the third masker condition. All subjects participated voluntarily in this study and showed normal hearing at audiometric frequencies between 250 and 8000 Hz.

Analysis. Individual SRTs for each listener were calculated by averaging SRTs across three adaptive tracks. Median SRTs for each of the five spatial masker configurations were then calculated across listeners for each of the three masker versions.

In order to determine the SRM, the difference between the SRT in the reference configuration and the SRT in each of the four spatial test configurations was calculated.

For statistical analysis, one-way repeated-measures ANOVAs and post hoc pair-wise comparisons using Bonferroni adjustments were applied (IBM SPSS Statistics 22).

Simulation

To develop an understanding of the underlying factors responsible for the experimentally determined SRM, four different models were used to predict the current

FIG. 1. Spatial configuration of target and maskers. From left to right: The left-most panel depicts the reference condition, i.e. all eight masking signals are presented together with the target signal (T) from 0°. Panels 2 to 5 depict the four test conditions. In these conditions, the

eight masking signals were either all presented from the loudspeaker in the back of the listener or distributed on either two, four or eight loudspeakers, symmetrically placed around the listener.

results. It should be noted that within each model, the same parameters were used for all predictions.

Long-Term SNR at the Ear Drum

The long-term SNR at the left and right ear was calculated independently in order to simulate SRM resulting from the physical properties of the head and pinna. Each of the eight maskers and the target sentences were replaced by steady-state SSN. Head-related transfer functions (HRTFs: 'large pinnae' from the CIPIC HRTF database) were used to position each of the masker SSNs and the target SSN at a specific azimuth for the five spatial configurations. The eight spatialized masker SSNs were then added resulting in a masker-only SSN. The target SSN was added at 11 different SNRs. A very simple decision device was implemented assuming that the target was detected when the level of the target-plus-masker SSN was at least 1 dB higher than the level of the masker-alone SSN. This calculation was performed for each ear and each masker configuration. The SRM relative to the reference configuration was then derived for each masker condition.

Binaural Speech Intelligibility Model

The further SI simulations (see below) were all based on the short-time BSIM by Beutelmann et al. (2010). For a detailed description of the full model, the reader is referred to their publication. Briefly, the model receives left and right ear input signals (target alone, in the form of SSN, and masker alone). The signals are passed through a Gammatone filterbank. The filterbank consists of 30 frequency bands between 140 Hz and 9 kHz, each frequency band having an equivalent rectangular bandwidth according to Glasberg and Moore (1990). Each frequency band is processed individually by an EC stage (Durlach 1963). In order to find the best SNR at the two ears, this stage delays and attenuates the signals of one ear with respect to the other ear (equalization) and then subtracts the two signals (cancellation). The processing is performed in about 20-ms-long consecutive time frames (1024 samples, 44.1 kHz sampling rate)

with Hann windows (resulting in equivalent rectangular duration of 12 ms). For each temporal frame and frequency band (T-F glimpse), the output of the EC stage provides a maximally improved SNR which can be resulting from either equalization-cancellation or from selection of the better ear. These time frame and frequency-dependent SNRs serve as input for the monaural speech intelligibility index (SII). The SII includes a speech importance weighting of frequencies and transforms the SNRs into an estimate of speech intelligibility between 0 and 1. The SII is calculated based on per-band SNR in the time frames and combined across time yielding a single value for the whole target signal. To derive an SRT estimate, the SII value corresponding to the measured SRT for the co-located speech masker configuration was calculated using BSIM and served as reference value. This reference SII value (0.407) was the only free parameter in the model. For all other combinations of masker conditions and spatial configurations, and for all BSIM versions as defined below, the SNR was then adjusted to match the model SII output to that reference SII value. For each combination of masker condition and spatial configuration, all calculations were performed for eight random segments of the maskers with a duration of about 2.1 s (equivalent to the duration of OLSA sentences). The final SRT estimate was the average SRT across the eight masker segments.

Dual-Monaural BSIM. In this version, each ear (including cochlear pre-processing) was modelled individually but without the EC stage and without the possibility to make use of better-ear glimpses. For this, the EC stage in BSIM was deactivated and SI predictions were generated for the left ear and the right ear independently.

BSIM Without EC (Better-Ear Glimpsing). Better-ear glimpsing was enabled in this version; thus, the model can select the better ear in each 20-ms time frame and auditory filter independently to observe

the best possible SNR. However, there is no further improvement of the SNR by using the EC stage and exploiting interaural phase differences.

BSIM. This was the full BSIM model as briefly described above, including both better-ear glimpsing and the EC stage.

RESULTS

Psychophysics

SRTs expressed as TMR against the summed level of all eight maskers are plotted in the three panels of Figure 2 for the three masker conditions, respectively. Each panel shows boxplots of the SRTs (medians and interquartiles across listeners) for the five spatial masker configurations indicated on the x -axis.

For the speech masker condition (left panel of Fig. 2), listeners showed the highest SRT of about -4.5 dB in the (co-located) reference configuration. An SRM was observed for the four other masker configurations. The lowest SRT of about -10 dB was found when the eight masker signals were presented from four different positions ($\pm 40^\circ/\pm 140^\circ$).

The middle and right panels of Figure 2 show the median SRTs for the LPS/HPSSN and HPS/LPSSN masker condition, respectively. Similarly to the speech maskers, SRTs were highest, i.e. SI was worst, in the reference configuration and a clear SRM occurred for the other spatial configurations. Interestingly, the SRTs for both the LPS/HPSSN and the HPS/LPSSN maskers were generally about 3 dB lower than for the speech masker, i.e. SI improved when maskers were partially replaced by speech-shaped noise.

A one-way repeated-measures ANOVA showed a highly significant effect of spatial masker configuration for all three masker conditions [$F(2.26, 20.32) = 155.68$, $p < 0.001$; $F(4, 20) = 123.66$, $p < 0.001$; and $F(1.41, 12.73) = 79.81$, $p < 0.001$ for broadband, LPS/HPSSN and HPS/LPSSN, respectively]. The assumption of sphericity was violated for broadband and HPS/LPSSN (Mauchly's test), and Greenhouse-Geisser correction was used. Significant differences of the post hoc pair-wise comparisons (Bonferroni corrected) between SRTs within each masker condition are represented by black lines in Figure 2. All differences were significant at $p < 0.001$, except for LPS/HPSSN ($0-180$, $\pm 40^\circ/\pm 140-180$, $\pm 40^\circ/\pm 140-90$, $\pm 40^\circ/\pm 140$ -distr.: $p < 0.01$) and for HPS/LPSSN (± 90 -distr.: $p < 0.05$). Overall, listeners clearly benefitted from the spatial separation of the maskers from the target, confirming results of previous studies with a symmetric configuration of two maskers. Second, SRTs improved when the speech maskers were replaced by SSN in their high- or low-frequency part. Considering the quite steep psychometric functions

that are achieved with the OLSA (15–20 % SI per dB), the observed 3–4-dB SRT improvement reflects a considerable improvement of SI at the same TMR.

To quantify the differential effect of the different spatial masker configurations, SRM in decibel relative to the co-located (0°) configuration was calculated. The SRM is shown in Figure 3 in the same format as Figure 2. A one-way repeated-measures ANOVA showed a highly significant effect of spatial masker configuration for all three masker conditions [$F(1.55, 13.94) = 70.98$, $p < 0.001$; $F(3, 15) = 41.41$, $p < 0.001$; and $F(1.26, 11.60) = 12.19$, $p = 0.003$ for broadband, LPS/HPSSN and HPS/LPSSN, respectively]. The assumption of sphericity was violated for broadband and HPS/LPSSN (Mauchly's test), and Greenhouse-Geisser correction was used. Significant differences of the post hoc, pair-wise comparisons (Bonferroni corrected) between SRM within each masker condition are represented by black lines in Figure 3. All differences were significant at $p < 0.001$, except for LPS/HPSSN ($\pm 40^\circ/\pm 140^\circ$ to ± 90 , $\pm 40^\circ/\pm 140$ -distr.: $p < 0.01$) and for HPS/LPSSN (± 90 -distr.: $p < 0.01$). The data clearly show that listeners benefitted (with a masking release between 3 and 4 dB) from the eight maskers being located at 180° , i.e. directly behind them. When four of the maskers were presented 90° to the left and the other four 90° to the right, listeners benefitted further when the maskers were full speech but not when the maskers were partially replaced by SSN. The strongest SRM was observed when the eight maskers were distributed across four positions ($\pm 40^\circ/\pm 140^\circ$) at least for the speech and LPS/HPSSN maskers. Distributing the eight maskers further to eight distinct azimuthal positions around the listeners did not result in a further release from masking. Instead, the SRM was significantly smaller than in the $\pm 40^\circ/\pm 140^\circ$ configuration across all masker conditions.

Simulations

The three panels of Figure 4 show SRTs generated from the BSIM model for the three masker conditions in the same format as the data in Figure 2. Simulated SRTs are represented by black horizontal lines. The experimental data are given in grey for comparison. The goodness of fit between the experimental data and model predictions is quantified in terms of root-mean-squared error (RMSE) and mean linear deviation (bias). The model was adjusted to match the SRT for the co-located (0°) configuration with speech maskers (left-most data point in the left panel). The model showed a similar pattern of results as observed in the data for the speech maskers. SRTs were slightly higher in the 180° , $\pm 90^\circ$ and $\pm 40^\circ/\pm 140^\circ$ configuration, underestimating the empirical SRM. For the

FIG. 2. Speech reception thresholds. SRTs are depicted as boxplots (medians and interquartiles across listeners) as a function of the five spatial configurations for the speech maskers, LPS/HPSSN maskers and HPS/LPSSN maskers, respectively. The number of listeners is

indicated in each panel, and significant differences between SRTs (post hoc pair-wise comparisons with Bonferroni correction, see text) are represented with *black lines*.

LPS/HPSSN maskers (middle panel of Fig. 4), simulated SRTs deteriorated by about 2 dB, indicating that the model requires a higher target level to reach the same intelligibility when compared to the speech masker condition. This is in strong contrast to the experimental data, where SRTs improve by about 3 dB when speech maskers are replaced by LPS/HPSSN maskers. For HPS/LPSSN maskers (right panel of Fig. 4), simulated SRTs decreased by about 1 dB compared to the LPS/HPSSN maskers but were still worse than the data. Thus, the BSIM model was not able to simulate the listeners' improvement in SI (lower SRTs) for the two modified masker conditions. In summary, the BSIM model showed good agreement with the experimental data concerning SRM, even in the different symmetric complex spatial configurations, but it appears unable to predict the

experimentally observed effect of replacing high- or low-frequency parts of the masking speech by SSN.

To assess the contributions of the different auditory processing stages in the model, four model versions were used for the simulations in Figure 5. SRM calculated from the empirical data is given in grey in the same style as in Figure 3. The model simulations are also expressed as SRM. The columns represent the three masker conditions, and the rows represent different model versions with different processing stages. Again, the goodness of fit between the experimental data and model predictions is quantified in terms of RMSE and bias. The most basic simulation (long-term SNR at eardrum) in the top row determined the long-term SNR at both ears, indicated by the sub-bars for the left (black) and right (grey) ear, respectively. The simulation showed that head

FIG. 3. Spatial release from masking. SRM is depicted with respect to the reference configuration for the speech maskers, LPS/HPSSN maskers and the HPS/LPSSN maskers, respectively. The number of listeners is indicated in each panel, and significant differences between the unmasking (post hoc pair-wise comparisons with Bonferroni correction, see text) are represented with *black lines*.

FIG. 4. Speech reception thresholds simulated by the BSIM. The *black horizontal lines* are model simulations for each of the three masker conditions. The data are re-plotted for convenience in *grey*. The model was adjusted to match the reference configuration (0°) for

the speech masker condition. To allow for quantitative comparisons of the goodness of fit, the root-mean-squared error (RMSE) and the mean linear deviation (Bias) are provided in each panel.

shadow and pinna effects have very low predictive power for the observed SRM. For most masker configurations, the simulation predicted negative unmasking (more masking), whereas the listeners showed positive unmasking (see panels A–C). For the 180° condition, a release from masking could be found for the right ear which is caused by an asymmetry of the CIPIC HRTFs for 0° and 180° . This yielded a long-term SNR of about 1.8 dB. Very similar results were found when simulating both ears individually in the modified dual-monaural BSIM shown in the second row of Figure 5D–F. In contrast to the long-term SNR model in the first row, this model included auditory filters, short-time processing and transformation of per-band SNRs to SI. Again, the model only considered the signal at either ear in isolation yielding two independent predictions indicated by the black and grey lines. A slight SNR improvement at the right ear for the 180° configuration was observed, in line with the long-term SNR at the eardrum. Otherwise, no considerable SRM was observed. In the third row (panels G–I), both ears of the BSIM were enabled which allows the model to use better-ear glimpsing, i.e. in each T-F frame, the better-ear SNR is used as input to the SII backend of the model. The binaural processing stage (EC stage in BSIM) was still disabled. In this case, the SRM increased by 1 to 2 dB. However, it was still considerably less than observed in the psychophysical data. In the bottom row (Fig. 5J–L), the simulation for the complete BSIM model (i.e. the SRM calculated from the simulations in Fig. 4) is shown. This model also included the EC stage and thus an interaural processing stage which utilizes both, short-time, channel-wise ITDs and ILDs, to reduce effects of

energetic masking. As already obvious from Figure 4, the BSIM could account for most of the spatial unmasking observed in the psychophysical data with a large deviation occurring for the 180° configuration, particularly for HPS/LPSSN (lower right panel).

DISCUSSION

The present study determined SRTs for a frontal speech target presented with three different masker conditions in five different symmetrical spatial configurations. Compared to a reference condition (all maskers co-located with the target), a considerable SRM (3–6 dB) was observed for the speech masker. When low- or high-frequency parts of the speech maskers were replaced by steady-state SSN, overall SRTs, independent of spatial configuration, improved considerably but the SRM persisted.

Effect of Spatial Masker Configuration on SRM

Similar to other studies with symmetric masker configurations (Bronkhorst and Plomp 1988; Glyde et al. 2013; Marrone et al. 2008), subjects benefitted strongly from a spatial separation of the maskers from the target. Most comparable to the spatial configurations tested in these studies was the current $\pm 90^\circ$ configuration. In a comparable arrangement of target (0°) and maskers ($\pm 90^\circ$), the SRM of ~ 5.5 dB in the current study was smaller than in, e.g. Marrone et al. (2008; ~ 12 dB). However, it should be noted that in their study SRM was quantified with only two maskers (all talkers female) compared to the current eight

FIG. 5. Spatial release from masking simulated by four model versions (rows) for the three masker conditions (columns). The model simulations are represented by the horizontal black bars. The data are given in grey dashed lines for comparison. In the first two model version (panels A–F), left and right ears were modelled separately (black and grey sub-bar, respectively). Panels A–C: long-term SNR at each ear (left and right sub-bar).

Panels D–F: BSIM with only left or right ear enabled (left and right sub-bar). Panels G–I: BSIM with both ears enabled but EC stage disabled. Panels J–K: full BSIM. Again, to allow for quantitative comparisons of the goodness of fit, the root-mean-squared error (RMSE) and the mean linear deviation (Bias) are provided in each panel.

maskers of mixed gender. Glyde et al. (2013) found an SRM of ~7 dB with two different talkers. The fact

that SRM was lower in the current study can be partly attributed to the much higher number of speech

maskers (Cullington and Zeng 2008; Hawley et al. 2004) limiting the availability of temporal troughs in the combined masker. Another factor influencing SRM, even if the same number of maskers was used across studies, is the exact psychophysical procedure to quantify SRTs (the test type) and the related amount of informational masking. SRM measured in tests with high informational masking was greater than in tests with lower informational masking (e.g. Best et al. 2015; Glyde et al. 2013). Furthermore, the distribution of genders across target and masking talkers is known to affect informational masking (e.g. Brungart 2001) and, consequently, also SRM.

When distributing the eight maskers from two positions to four positions ($\pm 40^\circ/\pm 140^\circ$), SRM increased by about 1.5 dB to almost 6 dB. The simulations showed that even with the most basic model (cf. Fig. 5A–C) this improvement can also be observed, although overall SRM values in this simple model were much too small. Likewise, the dual-monaural BSIM accounted for this improvement. Thus, it is obvious that the improvement is related to spectral changes in the HRTFs. Inspection of the HRTFs and resulting ILDs showed that in comparison to $\pm 90^\circ$, (i) the ILD in the 1–2-kHz range is increased (up to 6 dB) for $\pm 40^\circ$, (ii) both ears show an attenuation of about 1–1.5 dB below 1.5 kHz for $\pm 40^\circ$ and $\pm 140^\circ$ and (iii) for $\pm 140^\circ$ degrees considerable attenuation is observed above 4–5 kHz. In combination, these HRTF differences explain the 1.5-dB increased SRM for $\pm 40^\circ/\pm 140^\circ$ in comparison to $\pm 90^\circ$. Results from Marrone et al. (2008) showed virtually no change of SRM when their maskers were moved from $\pm 90^\circ$ to $\pm 45^\circ$, suggesting that part of the 1.5-dB benefit found in the current data might still be caused by moving four of the masker into the subjects' rear hemisphere. When all eight maskers were distributed equally around the listener, SRM did not improve further but deteriorated, relative to the $\pm 40^\circ/\pm 140^\circ$ configuration. Again studying the different model versions, this difference was absent in the simplest version and only emerged with increasing complexity of the model. At the same time, however, overall SRM values improved with increasing model complexity. The overall deterioration of the SRM in the distributed configuration relative to the $\pm 40^\circ/\pm 140^\circ$ configuration is likely to result from the most frontal maskers coming close to the target. In the distributed configuration they were located at $\pm 20^\circ$, showing reduced interaural differences in comparison to $\pm 40^\circ$. Likewise, the respective rear maskers ($\pm 140^\circ$) which also show only small interaural differences contribute. This is in line with the SRM in Marrone et al. (2008) which deteriorated by ~ 4 dB when their two speech maskers were moved from $\pm 45^\circ$ to $\pm 15^\circ$. The fact that this change in SRM was quite well

captured by the current full BSIM model indicates that also the model cannot exploit the time-variant binaural disparities between the maskers and the target, when the closest maskers were located $\pm 20^\circ$ from the target. It appears that due to the fact that the EC stage of the BSIM model operates on the peripherally filtered representations of the signals, the resulting binaural correlation limits the ability of the EC stage to separate target and maskers even within the 20-ms analysis windows. Such a limitation caused by the band-pass representation of the signals also seems plausible for the human auditory system.

The final and quite interesting spatial configuration in the current experiments was all eight maskers positioned at 180° , i.e. directly behind the listener. In this configuration, the listeners showed a significant SRM of about 3 dB. Previous studies on SRM of speech signals, both with noise and speech maskers, also found an SRM between 0.7 and 3 dB (Bronkhorst and Plomp 1988; Freyman et al. 2005; Peissig and Kollmeier 1997; Platte and vom Hövel 1980; Plomp 1967; Plomp and Mimpen 1981). The current simulations with the simplest model, which still included realistic HRTFs, showed a different amount of SRM for the two simulated ears. While the left ear showed virtually no unmasking, the right ear showed a SRM of about 1.5 dB (Fig. 5A). Not surprisingly, when better-ear glimpsing was allowed, the model consequently relied on the right ear and retained this SRM (Fig. 5G). Adding a binaural process in terms of the EC stage provided a small amount of additional SRM (~ 0.5 dB; Fig. 5J). The HRTFs used for these simulations were extracted from the 'large pinna' data, acquired with 5° azimuthal resolution. Inspection of these measured HRTFs revealed an asymmetry: the 0° HRTFs have a residual ILD (~ 2 dB), while the 180° HRTFs have no ILD but an ITD of about 50 μ s. The simulation results showed that the model can exploit these residual binaural cues to create a SRM. Equally, it is conceivable that the residual asymmetries between the 0° and 180° positions in the human subjects may contribute to the experimentally observed SRM of about 3 dB. Moreover, even though subjects were advised to orient their head to the front and to not move, small head movements and errors in the exact head orientation cannot be ruled out. To assess the effect of head orientations, model simulations were performed for head rotations of $+5^\circ$ or -5° relative to the exact frontal direction. The simulation results showed that when the maskers were presented from either 2, 4 or 8 positions around the listener, such small head movements did not affect SRTs. However, in the 180° configuration, i.e. when all maskers were directly behind the listener and only the target was in front, the simulated SRM was 2 or 5 dB for a rotation of $+5^\circ$

or -5° , respectively, in line with the spread of SRM values observed for the 180° configuration (3 to 4 dB). Thus, it is likely that small head movements had some impact on the SRM values measured on the listeners in this experiment that were not captured by the model. However, even if there are no binaural cues (no ITDs or ILDs) available to the subjects, Martin et al. (2012) showed that when separating target and maskers in elevation (providing spectral cues), there was a substantial amount of SRM which might also explain the SRM found for the 180° condition in the present study. Further research is required to more fully explore the dynamics of how better-ear listening and binaural interactions combine to determine listener performance in complex symmetric cocktail party environments.

Brungart and Iyer (2012) hypothesized that an important contribution to SRM with symmetrically arranged masking talkers originates from an optimal usage of better-ear glimpses, i.e. short-term variations in ILDs that, at each point in time and in each frequency channel, provide a better TMR in one of the ears. In contrast to Brungart and Iyer, the current study used a much higher number of interfering talkers (8 versus 2) which most likely results in a strong reduction of better-ear glimpses with a strongly improved SNR assessable by the listener. The relative importance of better-ear glimpsing might thus be lower in situations with a higher number of interfering talkers as the probability of high-SNR glimpses is reduced. Thus, one could expect interaural phase related cues to play a more important role when the number of available better-ear glimpses is reduced. To further assess the role of glimpsing in the current study, the high-frequency part of the speech maskers was replaced with steady-state SSN (LPS/HPSSN). In this condition, better-ear glimpses providing considerable SNR improvements were substantially reduced given that ILDs are considerably larger above than below 1.5 kHz. Following the hypothesis of Brungart and Iyer, it can be expected that this manipulation reduces SRM. Surprisingly, SRM with the LPS/HPSSN was quite similar to the SRM found with speech maskers, indicating that better-ear glimpses cannot entirely explain SRM. The question is how the replacement of the speech maskers by the LPS/HPSSN maskers is reflected in the model simulations: the two simplest versions (Fig. 5A–F) which did not allow for better-ear glimpses showed hardly any effect of the masker manipulation, as expected. Consistent with the hypothesis of Brungart and Iyer, the model with better-ear glimpsing (Fig. 5H) showed an overall inferior performance with LPS/HPSSN maskers than with speech maskers (Fig. 5G). Analysis of the glimpsing contributions in the model below and above 1.5 kHz showed that the SNR improvement

(in dB) above 1.5 kHz is by a factor of about 3–4 larger than below 1.5 kHz for the speech maskers, in line with the expectation based on the frequency dependency of ILDs. Still a considerable improvement of about 1.5 dB could be observed below 1.5 kHz. The summed band importance function used in the conversion of SNRs to SII for the channels below 1.5 kHz is nearly identical to that above 1.5 kHz. Thus, regarding SI both frequency regions contribute equally in the model, explaining the SRM predicted by the glimpsing model in the LPS/HPSSN conditions of about 1 dB.

Similar to the simulations with speech maskers, adding binaural interaction in terms of an EC stage improved the simulations to match the experimental SRM quite well.

The third masker condition consisted of HPS/LPSSN maskers where the low-frequency part of the maskers was replaced by steady-state SSN. This masker manipulation should retain the use of better-ear glimpses but preclude the use of time-variant ITD cues to be exploited for SRM. The experimentally observed SRM was similar to that for the speech maskers or the LPS/HPSSN maskers. The simulations with the glimpsing model (Fig. 5I) showed more SRM than for the LPS/HPSSN maskers but the difference was small. As pointed out above, in the glimpsing model, two effects combine: exploiting better-ear glimpses improves with increasing frequency and large bandwidth (>1.5 kHz) but the band importance function included in the SII applies the same weight to the low-frequency band (<1.5 kHz) in speech signals. The net effect is a small SRM for both the HPS/LPSSN and LPS/HPSSN masker. Again, adding the EC stage improved overall SRM values to match the experimentally observed SRM values quite well. At first sight, this improvement might be surprising because time-variant ITD cues, as they can be exploited by the EC stage, should have been precluded by the HPS/LPSSN maskers. In the model, the improvement may rely on time-variant envelope ITD or residual (carrier) ITD cues, caused by spectrotemporal ITD fluctuations of the uncorrelated SSN, exploited by the EC stage.

Many previous studies on SRM with speech targets were not only performed using speech maskers but also with SSN maskers (e.g. Culling et al. 2004; Freyman et al. 1999; Hawley et al. 2004; Jones and Litovsky 2011). However, such a masker condition was not included in the current study since the focus was on the role of fluctuating interaural cues in a multi-talker environment. Nevertheless, model simulations with uncorrelated broadband SSN maskers were performed to relate the current findings to these previous studies. The simulation results in Figure 6 show that reduced but still prominent SRM values of

FIG. 6. Spatial release from masking for SSN simulated by the four BSIM versions. Each coloured bar represents one of the four model simulations. In the dual-monaural BSIM, only left or right ear is enabled separately (blue and red sub-bar, respectively). In BSIM without EC, both ears are enabled but only better-ear glimpsing is used (grey bars). In contrast to the previous model versions, the full BSIM (black bars) show a considerable SRM.

up to 3 dB are observed for the full BSIM using broadband SSN as maskers (black bars). This reduction in SRM for SSN maskers is in line with experimental data from other studies comparing SRM with speech or speech-like maskers (e.g. Jones and Litovsky 2011; Noble and Perrett 2002). Comparisons with all less complex BSIM versions (coloured bars) demonstrate that the SRM in such spatial masker configurations is dominated by the EC stage. For the broadband SSN, the EC stage exploited the fact that the (uncorrelated) SSN maskers were spatially distributed and caused spectro-temporal ITD fluctuations, while the target speech had zero ITD and ILD. The absence of any SRM for BSIM without EC, i.e. glimpsing only, underlines the important role of (fluctuating) ITDs which can be exploited for speech perception in complex binaural conditions, while ILD glimpses do not contribute.

Effect of the Masker Conditions on SRTs

Data in Figure 2 show that, relatively independent of the spatial masker configuration, the empirical SRTs improved by about 3 dB when either the high- or low-

frequency part of the maskers was replaced by steady-state SSN. A similar improvement of SRTs (by 5 dB) was observed by Culling et al. (2004) when three co-located speech maskers were replaced with SSN. Differences between their findings and the current ones can be attributed to the lower number of maskers in the study of Culling et al. and that not only high- or low-frequency parts but the whole speech maskers were replaced by SSN. Hawley et al. (2004) found that for a single co-located masker, the replacement of a speech masker by steady-state SSN leads to a deterioration of SRTs while, when the same replacement is performed for three co-located maskers, the replacement leads to an improvement of SRTs. Marrone et al. (2008) also found a strong improvement of SRTs (~12 dB) when they time-reversed the signals of two maskers that were co-located with the target. Finally, Best et al. (2015) measured SRTs using two different experimental tests, the 'Modified Rhyme Test' and the 'Coordinate Response Measure corpus', but the same spatial arrangement of target and masker. Strongly improved SRTs (~10 dB) for the co-located target and masker condition were found for the rhyme test, which is assumed to produce lower confusability of target and maskers. These data for spatial masking conditions, together with the current findings, indicate that the more intelligible the maskers are, the larger the masking, a finding that is likely related to informational masking (Best et al. 2015; Culling et al. 2004; Hawley et al. 2004; Kidd et al. 2008). Such involvement of informational masking and associated cognitive effects is also reflected in the current simulation results which are based on spectral masking in T-F segments (including benefits of EC processing): while the full BSIM can predict the SRM for all masker conditions (Fig. 5J-L) fairly well, the model cannot predict the effect of the masker condition on the SRTs. Specifically, compared to the speech maskers, the model showed worse SRTs for the LPS/HPSSN and HPS/LPSSN condition whereas subjects showed improved SRTs (Fig. 4). In the model, SSN will always lead to higher thresholds than modulated maskers given that less 'dip listening' is available. Thus, by adjusting the model to match the SRTs for the modulated speech maskers, any informational masking occurring in humans and which is not assessed by the model is inherently covered by the threshold adjustment. Accordingly, if the model was adjusted to match the SRTs for LPS/HPSSN and HPS/LPSSN, it would predict considerably lower SRTs for the speech masker as observed in the data. It can then be hypothesized that the resulting gap between the model predictions and the data is related to informational masking which is not considered in the model. The fact that the model still correctly predicts the pattern of the SRTs (and thus SRM) indicates that the hypothesized informational masking effect is independent of the

spatial configuration given that the number of masking talkers and their relative level was fixed.

Another difference between SSN and the interfering talkers which should be mentioned is that SSN has independent temporal fluctuations across frequency bands, while the target speech has correlated modulation across frequency bands (comodulation). Hawley et al. (2004) directly compared SRTs for SSN maskers with and without comodulation. They found that listeners benefitted from the comodulation with only one masker but the benefit was absent for three maskers. These findings argue against comodulation affecting the current patterns of results.

SUMMARY AND CONCLUSIONS

The current experiments show that in a complex symmetric spatial masking situation, comparable to a crowded cocktail party, an SRM of up to 6 dB can be observed. When either the low-frequency or high-frequency part of the masking talkers was replaced by stationary SSN, the same SRM was observed as for natural speech maskers but overall SRTs improved significantly. From complementary modelling based on both binaural interactions and/or better-ear glimpses, the current study suggests the following:

- Informational masking plays a considerable role for SRTs (at least 3 dB of additional masking) in symmetric spatial masking conditions with eight masking talkers and does not depend on the spatial distribution of the maskers. It is likely, however, that the amount of informational masking will depend on the recruited speech material and experimental procedure.
- The observed SRM of up to 5 dB cannot be fully explained by better-ear glimpsing alone.
- Binaural speech intelligibility model predictions incorporating both better-ear glimpsing and binaural interaction in the form of an equalization-cancellation process indicate that the use of fast fluctuating interaural time or phase differences plays an important role for speech perception in complex symmetric spatial masking conditions with multiple talkers, comparable to a crowded cocktail party.

ACKNOWLEDGMENTS

This work was supported by the Bernstein Center for Computational Neuroscience, the German Center for Vertigo and Balance Disorders (IFB) and the DFG SFB TRR 31. We thank Lisa Benda and Annika Sander for their support during data acquisition.

REFERENCES

- BEST V, MASON CR, KIDD G JR, IYER N, BRUNGART DS (2015) Better-ear glimpsing in hearing-impaired listeners. *J Acoust Soc Am* 137:EL213–219. doi:10.1121/1.4907737
- BEUTELMANN R, BRAND T (2006) Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. *J Acoust Soc Am* 120:331–342
- BEUTELMANN R, BRAND T, KOLLMEIER B (2010) Revision, extension, and evaluation of a binaural speech intelligibility model. *J Acoust Soc Am* 127:2479–2497. doi:10.1121/1.3295575
- BRONKHORST AW (2000) The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions. *Acustica* 86:117–128
- BRONKHORST AW, PLOMP R (1988) The effect of head-induced interaural time and level differences on speech intelligibility in noise. *J Acoust Soc Am* 83:1508–1516
- BRONKHORST AW, PLOMP R (1992) Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing. *J Acoust Soc Am* 92:3132–3139
- BROWN A, WANG AC (2005) Separation of speech by computational auditory scene analysis. In: Benesty J, Makino S, Chen J (eds) *Speech enhancement*. Springer, New York, pp 371–402
- BRUNGART DS (2001) Informational and energetic masking effects in the perception of two simultaneous talkers. *J Acoust Soc Am* 109:1101–1109
- BRUNGART DS, IYER N (2012) Better-ear glimpsing efficiency with symmetrically-placed interfering talkers. *J Acoust Soc Am* 132:2545–2556. doi:10.1121/1.4747005
- CHERRY EC (1953) Some experiments on the recognition of speech, with one and with two ears. *J Acoust Soc Am* 25:975–979. doi:10.1121/1.1907229
- CULLING JF, HAWLEY ML, LITOVSKY RY (2004) The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources. *J Acoust Soc Am* 116:1057–1065
- CULLINGTON HE, ZENG FG (2008) Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects. *J Acoust Soc Am* 123:450–461. doi:10.1121/1.2805617
- DURLACH NI (1963) Equalization and cancellation theory of binaural masking-level differences. *J Acoust Soc Am* 35:1206–1218
- EWERT SD (2013) AFC—a modular framework for running psycho-acoustic experiments and computational perception models. *Proceedings of the International Conference on Acoustics AIA-DAGA 2013 in Merano, Italy: 1326–1329*
- FREYMAN RL, HELFER KS, BALAKRISHNAN, U (2005) Spatial and spectral factors in release from informational masking in speech recognition. *Acta Acustica united with Acustica* 91:537–545
- FREYMAN RL, HELFER KS, MCCALL DD, CLIFTON RK (1999) The role of perceived spatial separation in the unmasking of speech. *J Acoust Soc Am* 106:3578–3588
- GLASBERG BR, MOORE BC (1990) Derivation of auditory filter shapes from notched-noise data. *Hear Res* 47:103–138
- GLYDE H, BUCHHOLZ J, DILLON H, BEST V, HICKSON L, CAMERON S (2013) The effect of better-ear glimpsing on spatial release from masking. *J Acoust Soc Am* 134:2937–2945. doi:10.1121/1.4817930
- HAWLEY ML, LITOVSKY RY, CULLING JF (2004) The benefit of binaural hearing in a cocktail party: effect of location and type of interferer. *J Acoust Soc Am* 115:833–843
- JONES GL, LITOVSKY RY (2011) A cocktail party model of spatial release from masking by both noise and speech interferers. *J Acoust Soc Am* 130:1463–1474. doi:10.1121/1.3613928

- KIDD GJ, MASON CR, RICHARDS VM, GALLUN FJ, DURLACH NI (2008) Informational masking. In: Yost William A., Popper Arthur N., R. FR (eds) Auditory perception of sound sources, vol 29. Springer US, pp 143–189. doi:[10.1007/978-0-387-71305-2_6](https://doi.org/10.1007/978-0-387-71305-2_6)
- KIDD G JR, MASON CR, BEST V, MARRONE N (2010) Stimulus factors influencing spatial release from speech-on-speech masking. *J Acoust Soc Am* 128:1965–1978. doi:[10.1121/1.3478781](https://doi.org/10.1121/1.3478781)
- LAVANDIER M, CULLING JF (2010) Prediction of binaural speech intelligibility against noise in rooms. *J Acoust Soc Am* 127:387–399. doi:[10.1121/1.3268612](https://doi.org/10.1121/1.3268612)
- MARRONE N, MASON CR, KIDD G (2008) Tuning in the spatial dimension: evidence from a masked speech identification task. *J Acoust Soc Am* 124:1146–1158. doi:[10.1121/1.2945710](https://doi.org/10.1121/1.2945710)
- MARTIN RL, McANALLY KI, BOLIA RS, EBERLE G, BRUNGART DS (2012) Spatial release from speech-on-speech masking in the median sagittal plane. *J Acoust Soc Am* 131:378–385. doi:[10.1121/1.3669994](https://doi.org/10.1121/1.3669994)
- NOBLE W, PERRETT S (2002) Hearing speech against spatially separate competing speech versus competing noise. *Percept Psychophys* 64:1325–1336
- PEISSIG J, KOLLMEIER B (1997) Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners. *J Acoust Soc Am* 101:1660–1670
- PLATTE H-J, VOM HÖVEL H (1980) Zur Deutung der Ergebnisse von Sprachverständlichkeitsmessungen mit Störschall im Freifeld. *Acta Acustica United with Acustica* 45:139–151
- PLOMP R (1967) Pitch of complex tones. *J Acoust Soc Am* 41:1526–1533
- PLOMP R, MIMPEN AM (1981) Effect of the orientation of the speaker's head and the azimuth of a noise source on the speech-reception threshold for sentences. *Acustica* 48:325–329
- RHEBERGEN KS, VERSFELD NJ, DRESCHLER WA (2006) Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. *J Acoust Soc Am* 120:3988–3997
- WAGENER KC, BRAND T, KOLLMEIER B (1999) Entwicklung und Evaluation eines Satztests für die Deutsche Sprache III: Evaluation des Oldenburger Satztests. *Zeitschrift für Audiologie* 28:86–95
- WAN R, DURLACH NI, COLBURN HS (2014) Application of a short-time version of the equalization-cancellation model to speech intelligibility experiments with speech maskers. *J Acoust Soc Am* 136:768–776. doi:[10.1121/1.4884767](https://doi.org/10.1121/1.4884767)