Intelligibility for Binaural Speech with Discarded Low-SNR Speech Components

Esther Schoenmaker and Steven van de Par

Abstract Speech intelligibility in multitalker settings improves when the target speaker is spatially separated from the interfering speakers. A factor that may contribute to this improvement is the improved detectability of target-speech components due to binaural interaction in analogy to the Binaural Masking Level Difference (BMLD). This would allow listeners to hear target speech components within specific time-frequency intervals that have a negative SNR, similar to the improvement in the detectability of a tone in noise when these contain disparate interaural difference cues. To investigate whether these negative-SNR target-speech components indeed contribute to speech intelligibility, a stimulus manipulation was performed where all target components were removed when local SNRs were smaller than a certain criterion value. It can be expected that for sufficiently high criterion values target speech components will be removed that do contribute to speech intelligibility. For spatially separated speakers, assuming that a BMLD-like detection advantage contributes to intelligibility, degradation in intelligibility is expected already at criterion values below 0 dB SNR. However, for collocated speakers it is expected that higher criterion values can be applied without impairing speech intelligibility. Results show that degradation of intelligibility for separated speakers is only seen for criterion values of 0 dB and above, indicating a negligible contribution of a BMLD-like detection advantage in multitalker settings. These results show that the spatial benefit is related to a spatial separation of speech components at positive local SNRs rather than to a BMLD-like detection improvement for speech components at negative local SNRs.

Keywords Speech intelligibility · Speech interferers · Multitalker situation · Binaural masking level differences · Binaural listening · Binaural detection · Masking · Masking release · Spatial unmasking · Speech segregation

E. Schoenmaker (✉) · S. van de Par
Acoustics Group, Cluster of Excellence "Hearing4All", Carl von Ossietzky University, Carl von Ossietzkystraße 9-11, D-26129 Oldenburg, Germany
e-mail: esther.schoenmaker@uni-oldenburg.de

S. van de Par
e-mail: steven.van.de.par@uni-oldenburg.de

P. van Dijk et al. (eds.), *Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing,* Advances in Experimental Medicine and Biology 894,
DOI 10.1007/978-3-319-25474-6_9

1 Introduction

When listening to speech in a noisy background, higher speech intelligibility is measured for spatially separated speech and interfering sources, as compared to collocated sources. The benefit of this spatial separation, known as spatial release from masking, has been attributed to binaural processing enabling a better signal detection, as can be measured by binaural masking level differences (BMLDs). These BMLDs describe the difference in thresholds for tone-in-noise detection when both the target tone and masking noise are presented interaurally in phase, in contrast to a reversed phase of the tone in one ear. In the latter case thresholds are significantly lower. Similar interaural relations of the stimuli can be achieved by presenting the noise from a spatial position directly in front of the listener and the signal from a different spatial location. Therefore a connection was hypothesized between the spatial benefit in speech intelligibility and the tone detection advantage that had been measured as BMLD.

Levitt and Rabiner (1967) predicted the spatial improvement in speech intelligibility using this hypothesis. The underlying assumption of their model is that a BMLD effect acts within each frequency band, leading to a within-band detection advantage. This was modeled as a reduction of the level of the noise floor equal in size to the BMLD for that particular frequency, resulting in a more advantageous signal-to-noise ratio (SNR). This would make previously undetectable speech elements audible and enable them to contribute to speech intelligibility. More recent models of speech intelligibility also make use of the concept of within-band improvement of SNR to model the spatial release from masking (e.g., Beutelmann et al. 2010; Lavandier et al. 2012; Wan et al. 2014).

Whereas Levitt and Rabiner (1967) proposed their model for speech in stationary noise, we are interested in speech interferers. Since speech shows strong modulations both in time and frequency, the level of the interferer will vary. We will therefore consider BMLD effects at the level of spectro-temporal regions rather than frequency channels. Thus we extend the original hypothesis and postulate that a BMLD effect might take place at the level of individual spectro-temporal units and that this detection advantage would be responsible for the spatially improved speech intelligibility.

The aim of this study is to investigate whether a BMLD-like effect, that would lead to improved detection of speech elements in a spatially separated source configuration, indeed is responsible for improved speech intelligibility in the presence of interfering speech.

We will test the contribution of a BMLD-like effect by deleting spectro-temporal regions of the target signal below a specific SNR criterion and measuring speech intelligibility. We assume that, if certain speech elements contribute to speech intelligibility, the performance on a speech intelligibility task will be lower after deleting them. At low values of the SNR criterion this manipulation will lead to deletion of spectro-temporal regions at which either little target energy was present or the target signal was strongly masked. At such values not much impact on speech

intelligibility is expected. Deletion of target speech components at increasing SNRs, however, will start to affect speech intelligibility at some point.

The question of interest is whether the effect of eliminating spectro-temporal regions will follow a different course for collocated and spatially separated source configurations. Should a BMLD-like effect indeed act at the level of small spectro-temporal regions, then deletion of target speech from spectro-temporal regions with local SNR values between −15 and 0 dB (i.e. between dichotic and diotic masked thresholds, Hirsh (1948)), should degrade speech intelligibility for separated, but not for collocated sources, since those speech components would be inaudible in a collocated configuration anyway. This implies that over this same range of SNR criteria the spatial release from masking, as measured by the difference in performance between separated and collocated source configurations, should decrease.

2 Methods

2.1 Stimuli

German sentences taken from the Oldenburg Sentence Test (OLSA, Wagener et al. 1999) and spoken by a male speaker served as target speech. Each sentence consisted of five words with the syntactic structure *name—verb—numeral—adjective—object*. For each word type ten alternatives were available, the random combination of which yields syntactically correct but semantically unpredictable sentences.

In order to have a spatially complex masker, the interfering speech consisted of two streams of ongoing speech spoken by two different female talkers, and was obtained from two audio books in the German language. Any silent intervals exceeding 100 ms were cut from the signals to ensure a natural-sounding continuous speech stream without pauses. All target and masker signals were available at a sampling rate of 44.1 kHz.

A total of 200 masker samples for each interfering talker were created by randomly cutting ongoing speech segments of 3.5-s duration from the preprocessed audio book signals. This length ensured that all target sentences were entirely masked by interfering speech. The onsets and offsets of the segments were shaped by 200-ms raised-cosine ramps.

The target sentences were padded with leading and trailing zeros to match the length of the interfering signals. Some time roving was applied to the target speech by randomly varying the number of leading zeros between 25 and 75 % of the total number of zeros.

The relative levels of the three signals were set based on the mean RMS of the two stereo channels after spatialization. The two interfering signals were equalized to the same level, whereas the target speech (before application of the binary mask, see next section) was presented at a level of −8 dB relative to each single interferer. This resulted in a global SNR of approximately −11 dB. The sound level was set

to 62 dB SPL for a single interferer. After signal manipulation as will be described below, the three signals were digitally added which resulted in a total sound level of approximately 65 dB SPL.

All speech signals were presented from virtual locations in the frontal horizontal plane. The spatial positions were rendered with the help of head-related transfer functions (HRTF), that had been recorded according to Brinkmann et al. (2013), using a Cortex MK2 head-and-torso simulator at a source-to-head distance of 1.7 m in an anechoic chamber.

Two different speaker configurations were used in this experiment, one with spatially separated speakers and one with collocated speakers. The target speech was presented from 0° azimuth, i.e. directly in front of the listener, in both configurations. In the collocated configuration the two sources of masking speech were presented from this same location. In the spatially separated configuration one interfering speaker was presented from the location 60° to the left and the other speaker from 60° to the right of the target.

2.2 Target Signal Manipulation

In order to investigate the contributions of target speech at various levels of local SNR, the target speech needed to be manipulated.

The first step in the stimulus manipulation consisted of the calculation of the local, i.e. spectro-temporal, SNR of the three-speaker stimulus. Spectrograms were calculated from 1024-point fast Fourier transforms on 1024-samples long, 50%-overlapping, square root Hann-windowed segments of the signals. This was performed for each of the two stereo channels of a spatialized speech signal to accommodate interaural level differences. This resulted in a pair of left and right spectrograms for each signal, from which one estimate of the spectro-temporal power distribution needed to be calculated. This was achieved by computing the squared mean of the absolute values of each set of two (i.e. left and right) corresponding Fourier components and repeating this for the complete spectro-temporal plane (Fig. 1, left column).

Subsequently, the resulting spectral power distributions were transformed to the equivalent rectangular bandwidth (ERB) scale. This was achieved by grouping power components within the frequency range of one ERB. The time-frequency (T-F) units thus created covered a fixed time length of 23 ms (i.e. 1024 samples) and a frequency band of 1 ERB width.

In the final step to calculate the spectro-temporal SNR, the spectro-temporal power of each target signal was compared to the summed spectro-temporal power of the two masker signals with which it was combined in a single stimulus. An SNR value was calculated for each T-F unit of the combined three-speaker signal, resulting in a 2-D matrix of local spectro-temporal SNR values (Fig. 1, center).

The spectro-temporal SNR representation was used to decide which components of the target speech signal would be passed on to the final signal. Any local SNR

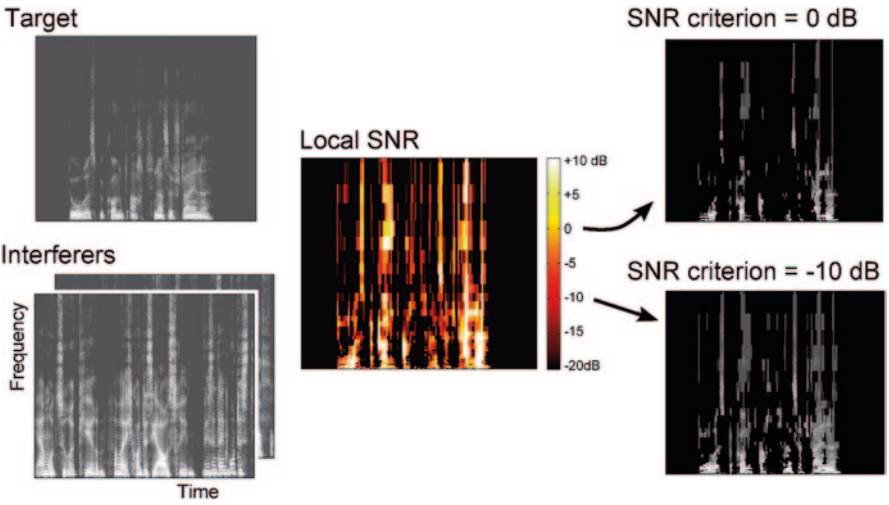

Fig. 1 Overview of the stimulus manipulation. *Left column*: Mean spectrograms of original target and interfering speech. *Center*: Local SNRs of T-F units. *Right column*: Spectrograms of target speech after discarding low-SNR T-F units. Examples are shown for SNR criterion values of 0 and − 10 dB

values exceeding the selection criterion resulted in a value of 1 of a binary mask, and thus in selection of the corresponding spectro-temporal components of the target signal in both ears. Any local SNR values that did not pass the criterion resulted in a value of zero and deselection of the corresponding components of the target (Fig. 1, right column).

A total of 8 different local SNR criteria were included in the experiment based on pilot testing: − 10,000, − 4, 0, 2, 4, 6, 8 and 10 dB. The criterion of − 10,000 dB was chosen to simulate an SNR of $-\infty$, and served as a control condition of unfiltered target speech in which the entire target signal was passed on to the final stimulus.

After addition of the selected spectro-temporal components of the target signal to the complete spectrograms of the two interferer signals, the final signal was reconstructed by calculation of the inverse fast Fourier transform, followed by windowing with square-root Hann windows and overlap-add.

2.3 Procedure

Six normal-hearing participants (4 females, 2 males, aged 22–29 years) participated in the experiment. All were native German speakers.

Stimuli were presented over Sennheiser HD650 headphones in a double-walled soundproof booth. The OLSA speech sentence test was run as a closed test and the 5 × 10 answer alternatives were shown on a PC screen. The participants were instructed to ignore the female speakers, and to mark the perceived words from the

male speaker at the frontal position in a graphical user interface. They were forced to guess upon missed words. The listeners participated in a prior training session in which the target sentences had not been manipulated. The training session comprised 160 trials at a global SNR that started from 0 dB and was gradually decreased to the experimental SNR of −8 dB.

During the two experimental sessions that consisted of 160 trials each, all 16 conditions (8 SNR criteria ×2 spatial configurations) were tested interleaved in random order. One test list of 20 sentences, resulting in 100 test words, was used for each condition. Performance was measured as the percentage correct words.

3 Results

The top rows of Fig. 2 show the percentage of correctly recognized words versus SNR criterion for the six individual listeners. The graphs show a clear spatial release from masking for all listeners, reflected by the consistently higher scores for spatially separated speech. From Fig. 2 it can also be seen that the SNR criterion at which speech intelligibility starts to decrease below the performance in the

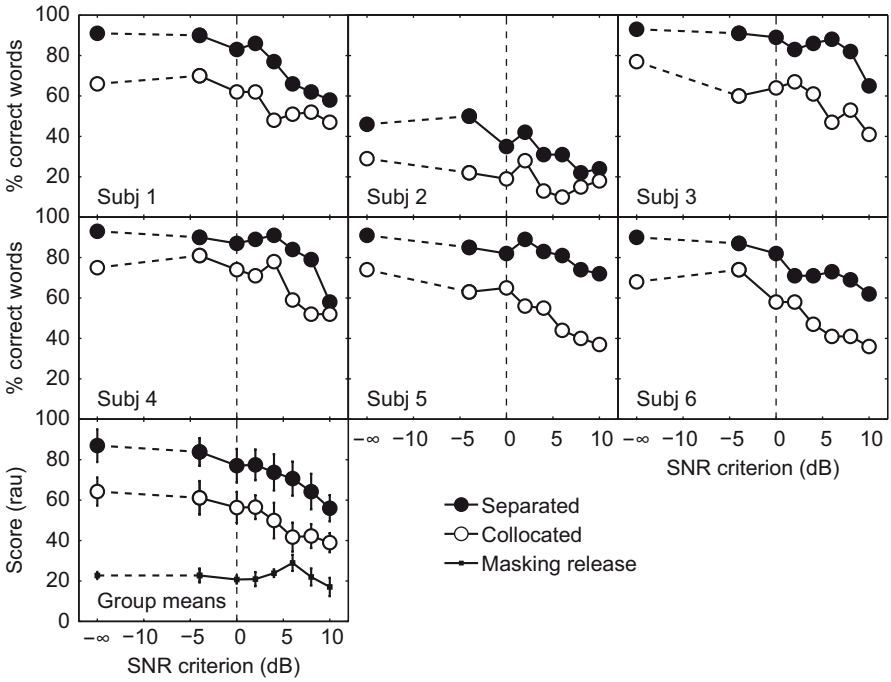

Fig. 2 *Upper two rows*: Individual results for six subjects, expressed as percentage correctly identified words versus SNR criterion. *Bottom row*: Mean results and spatial release from masking expressed in rau. The error bars represent standard errors

unfiltered condition (i.e. at SNR = −∞) differs among listeners between about −4 and 5 dB.

The mean results are shown in Fig. 2 (bottom), together with the spatial release from masking which equals the difference between the spatially separated and collocated data at each SNR criterion. From this graph it becomes clear that the spatial release from masking remains more or less constant over the complete range of SNR criteria tested. The curves for the separated and collocated curves apparently do not converge.

In order to perform statistical analyses, the percentage values were first transformed into rationalized arcsine units (Studebaker 1985). A two-way repeated-measures ANOVA with factors of spatial condition (collocated or separated) and SNR criterion shows a significant effect of spatial condition [$F(1,5) = 125.68$, $p < 0.001$], a significant effect of SNR criterion [$F(7,35) = 34.45$, $p < 0.001$], and no significant interaction of spatial condition and SNR criterion [$F(7,35) = 1.36$, $p = 0.253$]. The absence of interaction between spatial condition and SNR criterion confirms the constant spatial release from masking that could be observed in Fig. 2, bottom.

Simple contrasts comparing scores at each test SNR criterion to the reference criterion at −∞ dB, show that all SNR criteria including and exceeding 0 dB resulted in scores different from the reference criterion at the level of $p < 0.05$ (one-sided Dunnett's test). Only the SNR criterion of −4 dB shows no significant difference from the reference criterion.

4 Discussion

The aim of this study was to gain more insight into the mechanism leading to spatial release from masking in a multispeaker situation. Specifically, we were interested in assessing whether a BMLD-like detection advantage at the spectro-temporal level could be responsible for access to more target speech elements in a configuration with spatially separated sources, as this could potentially contribute to a higher speech intelligibility.

An experiment in which spectro-temporal regions from the target speech signal were removed according to a variable local SNR criterion gave no indication for any BMLD-like detection advantage. The results of the experiment showed a decrease in speech intelligibility when elements with a local SNR of 0 dB and larger were removed from the target speech, suggesting that speech elements at local SNRs above −4 dB did contribute to speech intelligibility. Interestingly, this was the case for both the collocated and spatially separated speaker configurations. The spatial release from masking appeared to be independent of the extent to which lower-SNR components were removed from the target signal.

Our method of applying an SNR-dependent binary mask to the target speech resembles the technique of ideal time-frequency segregation (ITFS) that is known from computational auditory scene analysis studies (e.g., Wang 2005; Brungart 2006; Kjems et al. 2009). Although these studies used diotic signals and applied

the masks to both the targets and interferers, the results are comparable. The ITFS studies found a decrease in target speech intelligibility for SNR criteria exceeding the global SNR of the mixture by 5–10 dB, similar to the local SNR value of about 0 dB (i.e. around 10 dB above global SNR) in our study.

An analysis of the local SNRs of all T-F units in ten unfiltered stimuli (see Fig. 3) was performed to determine the proportion of T-F units exceeding a certain SNR level. The analysis reveals that on average only about 20 % of the total target T-F units exceeded a local SNR value of –4 dB and about 7 % exceeded a value of 10 dB SNR. This corresponds to the proportion of target speech that was retained in the final signal at the SNR criteria mentioned. Note that at the criterion of −4 dB no decrease in speech intelligibility was observed as compared to the condition with intact target speech. This analysis thus demonstrates the sparse character of speech and the fact that only little target speech was needed to achieve a reasonable performance on this experimental task with a restricted speech vocabulary.

The stable spatial release from masking over the range of SNR criteria tested is one of the most important outcomes of this study. It suggests that recovery of subthreshold speech information by binaural cues does not play a role in the spatial release from masking in a multispeaker situation. Instead, masking release appeared to be sustained even when only high-SNR components of target speech were left in the signal. These supra-threshold components should be detectable monaurally in both the collocated and spatially separated conditions and thus their detection should not rely on a binaural unmasking mechanism as is the case for BMLDs.

An explanation for this spatial release from masking can be provided by directional cues of the stimuli that support a better allocation of speech components to the source of interest. The fact that speech sources originate from different spatial positions can eliminate some confusion that arises from the simultaneous presence

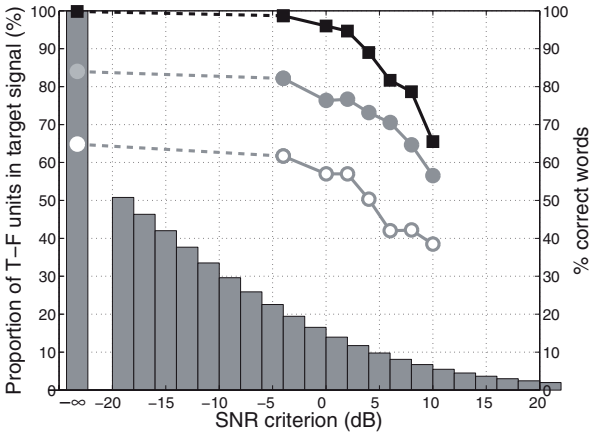

Fig. 3 The average proportion of T-F units that were retained in the manipulated target signal for a given SNR criterion are shown by bars (*left axis*). *Black squares* show the mean intelligibility of this target speech presented in absence of interferers (*right axis*). The mean data from Fig. 2 are shown in *grey* for reference

of multiple speech sources (Durlach et al. 2003). A further improvement can be seen for target speech presented in isolation, but otherwise manipulated identically as before (black squares in Fig. 3). These data show the inherent intelligibility of the sparse target speech, ruling out confusion from interfering speech.

Acknowledgments This work was supported by the DFG (SFB/TRR31 "The Active Auditory System").

References

Beutelmann R, Brand T, Kollmeier B (2010) Revision, extension, and evaluation of a binaural speech intelligibility model. J Acoust Soc Am 127(4):2479–2497

Brinkmann F, Lindau A, Weinzierl S, Geissler G, van de Par S (2013) A high resolution head-related transfer function database including different orientations of head above the torso. Fortschritte der Akustik. AIA-DAGA 2013, Merano, Italy (pp 596–599): DEGA e.V. Berlin

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

Durlach NI, Mason CR, Gerald Kidd J, Arbogast TL, Colburn HS, Shinn-Cunningham BG (2003) Note on informational masking (L). J Acoust Soc Am 113(6):2984–2987

Hirsh IJ (1948) The influence of interaural phase on interaural summation and inhibition. J Acoust Soc Am 20(4):536–544

Kjems U, Boldt JB, Pedersen MS, Lunner T, Wang D (2009) Role of mask pattern in intelligibility of ideal binary-masked noisy speech. J Acoust Soc Am 126(3):1415–1426

Lavandier M, Jelfs S, Culling JF, Watkins AJ, Raimond AP, Makin SJ (2012). Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources. J Acoust Soc Am 131(1):218–231

Levitt H, Rabiner LR (1967) Predicting binaural gain in intelligibility and release from masking for speech. J Acoust Soc Am 42(4):820–829

Studebaker GA (1985) A "rationalized" arcsine transform. J Speech, Lang Hear Res 28(3):455–462

Wagener K, Brand T, Kollmeier B (1999) Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests. Z Audiol 38:4–15

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(2):768–776

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