Influence of acoustic complexity on spatial release from masking and lateralization

Gusztáv Lőcsei, Sébastien Santurette, Torsten Dau, Ewen N. MacDonald
Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark, 2800 Kongens Lyngby, Denmark.

Summary
In realistic listening scenarios, humans are extremely skillful in following one particular talker even in the presence of many others (i.e., the cocktail party effect). One aspect of this is the ability of a listener to make use of the spatial separation between different sound sources. In a complex acoustic scene, as interferers are moved away from the spatial position of the target, speech intelligibility (SI) increases, often referred to as spatial release from masking (SRM). This benefit is largely based on the listeners' ability to make use of interaural level differences (ILD) and interaural time differences (ITD), which vary with the source location. While many studies have explored SRM, few have investigated the effects of overall number and spatial distribution of interferers while controlling for monaural masking effects. In the present study, speech reception thresholds (SRTs) and lateralization thresholds were measured over headphones in babble noise conditions consisting of 2, 4, 8 and 12 talkers. The perceived locations of the signal (female voice) and individual maskers (male, time-reversed voices) were steered separately either to the left or to the right using 680 μsec ITDs. For a fixed number of maskers, the distribution of interfering talkers (i.e., co-located or separated from the target) was varied. Thus, for all conditions with the same number of maskers, the monaural SNR was held constant regardless of the perceived spatial distribution. The performance between the speech and lateralization tasks was highly correlated. No substantial SRM occurred while one or more maskers were co-located with the target. Interestingly, spatially shifting the last 2 co-located interferers resulted in the same SRM, independent of the overall number of maskers. The same finding was found for the last 4 co-located maskers. The results suggest that SRM is independent of the overall number of interfering talkers.

PACS no. 43.71.Gv, 43.71.Es, 43.66.Pn

1. Introduction

The "cocktail party problem", as introduced by Cherry in 1953 [1], refers to the capability of listeners to selectively attend to a single voice in a multitude of interfering talkers. Being an essential phenomenon in everyday communication scenarios, it has received considerable investigation in the past 60 years. It is well established, that one aspect that aids listeners in adverse conditions is their ability to detect interaural level and timing differences (ITDs and ILDs, respectively) between the signals arriving at their ears. These differences vary for sources at different azimuthal angle relative to the listeners', which aids in streaming spatially separated sources [2]. Typically, spatial separation of a masker from the target stream increases speech intelligibility, a phenomenon called spatial release from masking (SRM). Many studies have focused on speech perception in spatially complex virtual environments [2, 3, 4]. Based on these studies, SRM appears to depend on both the number and spatial distribution of interferers, and often an interaction has been observed. However, as varying the spatial position of a sound source alters ILDs, the monaural energetic masking potential of the source will vary with position. As a result, it can be difficult to disentangle the effects of perceived spatial separation from better-ear listening on SRM.

In the present study, speech intelligibility was systematically investigated with regards to both the number of interferers and their distribution in the scene. Importantly, the monaural SNR was controlled such that perceived spatial location and complexity could be directly investigated. Using ITD cues, the perceived location of target and masker streams could be lateralized to either the left or the right. Here, masker streams having the same leading side as the target in the actual trial will be referred to as "co-located", while maskers lateralized towards the lagging side of the target will be called "separated". In addition, participants also performed a lateralization task to determine the relationship between the ability to localize the target and the ability to understand speech.

2. Methods

2.1. Listeners

A total of 9 listeners between the ages of 18-27 (median: 22) participated in this study. All were native Danish speakers with normal audiometric thresholds (i.e., $\leq$ 20 dB HL) and reported no speech or language impairments. The experiments were carried out in accordance to the ethical approval granted by the Science-Ethics Committee for the Capital Region of Denmark (reference H-3-2013-004). All listeners provided informed consent and were paid for their participation.

2.2. Experimental set-up and stimuli

Measurements were carried out in a double-walled sound attenuating listening booth. The listening tests were implemented in Matlab, and subjects provided their responses using a computer interface. The stimuli were presented through Sennheiser HDA200 headphones using an RME DIGI96/8 sound card at a sampling rate of 44.1 kHz. The calibration of the headphones was done using a B&K 4153 artificial ear connected to a B&K 2636 sound pressure level meter, and with a B&K 4230 artificial ear calibrator. In order to achieve a flat frequency response in the headphones, equalization filters were applied to all stimuli before presentation.

Target stimuli. The target sentences used were from the Danish DANTALE II corpus [5]. The five-word sentences are spoken by a female talker and have a fixed syntactical structure (*<name> <verb> <number> <adjective> <noun>*), where each word is a randomly chosen from 10 alternatives. The corpus is organized into 16 lists containing 10 sentences each. To reduce testing time, only the first 1.5 seconds of the sentences were used in the lateralization experiment.

Noise stimuli. To generate the noise stimuli for the conditions having 2, 4, 8 and 12 maskers, the first 2, 4, 8 and 12 male talkers from the GRID corpus [6] were used, respectively. Preprocessing of the stimuli included up-sampling to 44.1 kHz, silent frame removal, and then time-reversal. For silent frame removal, each recording was segmented into 50 ms windows with 25 ms overlap. Silent frames were defined as frames having an energy at least 30 dB lower than the maximum energy present among the frames in the recording. For each interfering talker, a continuous stream of speech was generated by concatenating the preprocessed recordings. From each of the resulting streams, 50 non-overlapping regions with a duration of 5 seconds were chosen randomly. These segments were then windowed (0.1 s cosine onsets and offsets) and stored as masker trials for the speech intelligibility test. A similar procedure was used for the lateralization test, but with masker segments of 4 s duration.

2.3. Procedure

The interfering masker stream was always presented at 50 dB SPL and the target level was adaptively varied to estimate the threshold. This means that the overall masking level increased with the number of masking streams (i.e., for 2, 4, 8 and 12 maskers, the long-term overall masker level was 53, 56, 59 and 60.8 dB SPL, respectively).

In each trial, the onset of the maskers preceded that of the target by at least 2 seconds. The perceived lateralized position of each stream (i.e., both maskers and target) was steered independently to the left or to the right using 680 μsec ITDs. The lateralized position of the target sentence was randomized from trial to trial. Thus, subjects had no *a priori* knowledge about which side to attend to in each trial. A total of 16 conditions varying in the total number and the number of co-located (i.e., with the same lateralized position as the target) maskers were tested.

To denote the different conditions, we will use the following notation throughout the paper: C_y^x. Here, y and x indicate the overall number of maskers and the number of co-located maskers, respectively. For example, C_8^2 denotes a condition with 8 maskers, from which 2 have the same lateralized position as the target. Using the notation above, the following 16 conditions were tested: C_2^2, C_2^1, C_2^0, C_4^4, C_4^2, C_4^0, C_8^8, C_8^6, C_8^4, C_8^2, C_8^0, C_{12}^{12}, C_{12}^8, C_{12}^6, C_{12}^4 and C_{12}^0. In the remainder of the paper, condition groups having 2, 4, 8 and 12 interferers will be denoted as C_2^x, C_4^x, C_8^x and C_{12}^x, respectively.

Threshold estimates for conditions involving the same number of interferers were run in parallel (e.g., estimation of thresholds for C_2^2, C_2^1, and C_2^0 were conducted simultaneously). If these estimates had not been collected simultaneously, the listeners would have known which side to attend to, based on the distribution of maskers. For example, if only C_2^0 trials were presented, listeners would know that the target was lateralized to the side opposite that of the two maskers.

The order in which condition groups were tested was based on the overall number of maskers. Thus, estimates were collected for C_2^x first and C_{12}^x last.

Speech intelligibility tests. Thresholds for 50% correct response were estimated using the standard procedure of the DANTALE II tests. Thresholds were estimated using 3 lists per condition. Subjects provided their response using a computer interface, which contained a 10 × 5 matrix, where columns corresponded to the word categories for the sentences and each row contained one of the alternatives per category. Subjects were requested to mark the words occurring in the target sentence. The presentation level of the next target sentence was adjusted based on the number of correctly identified words. The initial presentation level was set to the estimated level of the interferers (i.e., 53, 56, 59 and 61 dB SPL for C_2^x, C_4^x, C_8^x and C_{12}^x, respectively). The order of sentence presentation within lists was randomized. Lists for each condition were chosen in a semi-randomized way. Each of the 16 lists was presented exactly 3 times during the speech intelligibility tests and none of the conditions contained the same list twice. Thresholds were calculated from the average of the last 20 presentation levels. Before the test session, a training session was performed with 5 random sentences in each condition (80 sentences overall).

Lateralization tests. Thresholds for 79.4% correct response were estimated using a 1-up 3-down procedure in a 2 AFC task. Subjects had to respond using a computer interface, and were requested to indicate which side they heard the target sentence coming from (i.e., "left" or "right"). Each condition consisted of at least 10 reversals. The first two had a step size of 5 and 2 dB respectively, while the last 8 had a step size of 1 dB. Thresholds were calculated as the arithmetic average of the last 8 reversals. Since it was necessary to test all conditions within a group simultaneously, the threshold tracking did not stop until 10 reversals had occurred for all conditions within the group.

All listeners, except one, performed all the tests in two sessions, the first being the speech intelligibility test and the second being the lateralization test. One listener performed the speech intelligibility test in two sessions.

3. Results

SRT and lateralization thresholds. To account for the effect of increasing masker energy with increasing number of interferers, SNR ratios for criterion performance were calculated as target threshold levels minus the average presentation levels of the maskers (i.e., 53, 56, 59 and 60.8 dB SPL). Fig. 1 shows the speech reception and lateralization thresholds presented as box plots, grouped by the number of interferers in the scene (i.e., condition groups; shaded areas).

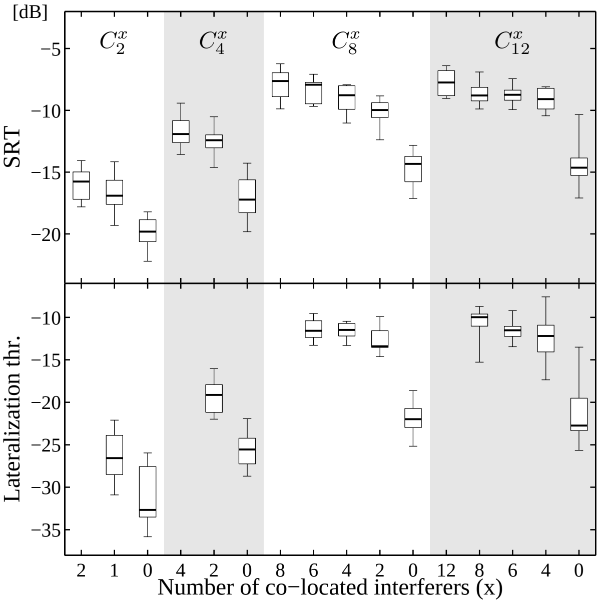

Figure 1. Box plots for the speech intelligibility (top) and lateralization (bottom) experiments. The whiskers extend to the most extreme data points. Condition groups using the same amount of interferers are defined at the top of the shaded areas. The number of interferers co-located with the target side (x) are denoted on the abscissa.

Results of the speech intelligibility experiment generally show very low SRT values, ranging from an average of approximately -19.9 to -7.8 dB across the conditions. The average SRT value for conditions having the same number of interferers (i.e., condition groups) increases gradually from approximately -17.5 to -10.3 dB as the number of maskers increases from 2 to 8, where it plateaus, yielding no further increment for 12 maskers. Within condition groups, SRT decreases (i.e. intelligibility increases) as more and more maskers are lateralized away from the target stimulus. These observations were supported by repeated-measures two-way ANOVA. Comparing the fully co-located (C_x^x), half co-located half-separated ($C_x^{x/2}$) and fully separated interferer distribution (C_x^0) conditions across all condition groups revealed a significant main effect of both the number of interferers $[F(1.6, 12.7)=315.6, p<0.001]$ and the interferer distribution $[F(2, 16)=415.6, p<0.001]$, and a significant interaction $[F(6, 48)=13.3, p<0.001]$. Post-hoc analysis using paired t-tests with Bonferroni correction on each condition group showed, that the C_x^0 conditions were significantly different from all the other conditions $[p<0.001]$. For 8 and 12 interferers, a repeated-measures ANOVA revealed no significant main effect for the number of interferers $[F(1, 8)=3.76, p=0.09]$, but a significant main effect for the distribution $[F(4, 32)=222.1, p<0.001]$ and a significant interaction $[F(1.7, 13.4)=5.1, p=0.03]$.

With the exception of the fully co-located conditions (C_x^x), the trends in the lateralization experiments follow a similar pattern as in the speech intelligibility experiments. For the fully co-located conditions, listeners could give the correct answer without actually hearing the target stimulus, but instead, noting its absence on the side opposite all the maskers. While two strategies were available for determining target lateralization in the fully co-located condition, this trend was only observed for conditions having 4

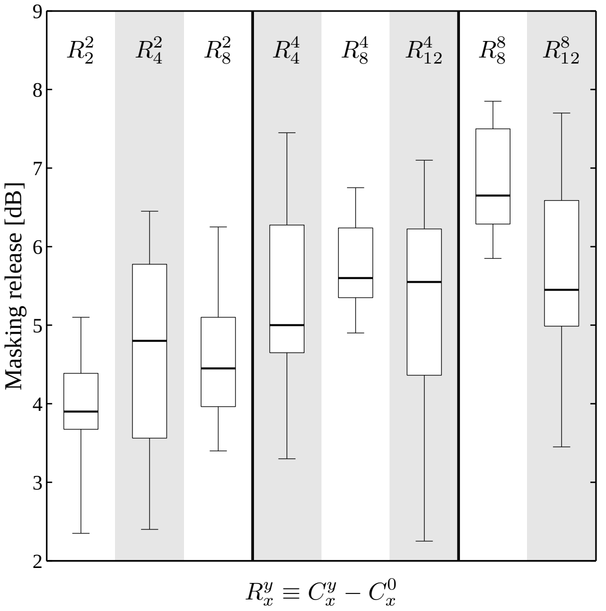

Figure 2. Box plots for the masking release by removing the last two or four co-located interferers from the target side. R_x^y stands for masking release between condition C_x^y and C_x^0.

or more interferers. It is possible that this pattern is a results of an order effect as each of the participants completed the test procedure starting with C_2^x.

Similar to the analysis of the speech intelligibility results, a repeated-measures two-way ANOVA was performed, comparing conditions with half co-located half-separated vs. fully separated interferer distributions across all condition groups. Results showed significant main effects for both the number of interferers [$F(3,24)$=92.4, p<0.001], and distribution [$F(1,8)$=318.6, p<0.001], as well as a significant interaction [$F(3,24)$=13.1, p<0.001]. Post-hoc analysis using paired t-tests (with Bonferroni correction) on each condition group showed that the C_x^0 conditions were significantly different from all the other conditions [p<0.001]. For 8 and 12 interferers, a significant main effect was found for the distribution [$F(3,24)$=131.3, p<0.001], but not for the number of interferers [$F(1,8)$=0.49, p=0.5] and the interaction was not significant [$F(3,24)$=0.12, p=0.95].

All of the reported ANOVA measures were corrected with Greenhouse-Geisser estimates, if the assumption of sphericity was violated.

Release from masking. SRM due to binaural unmasking was calculated to express the SRT benefit obtained when the distribution of the interferers is changed from a difficult layout to a more favorable one. In this case, this meant finding the SRT benefit when a fixed number of co-located masker streams were steered towards the other side of the head. More precisely, "SRM" will refer to the mean of the individual SRM values between the corresponding conditions.

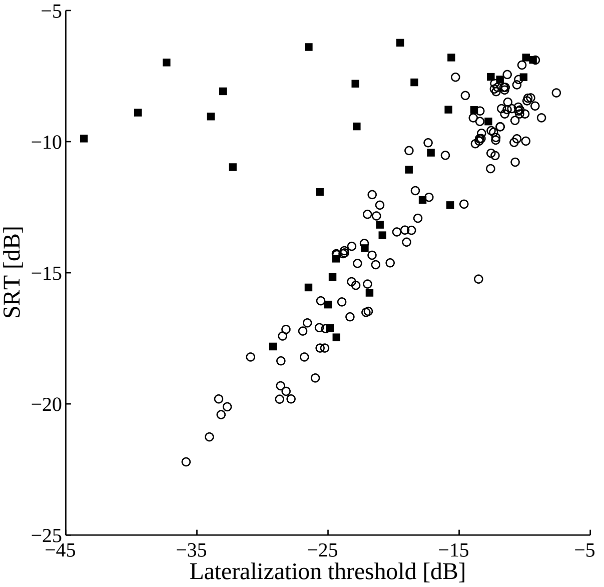

Figure 3. Correlation between SRTs and lateralization thresholds for each condition of each subject. Filled squares represent the conditions where all the interferers were on the target side. Open circles stand for all the other conditions.

In order to assess the overall SRM, level differences between the fully separated and fully co-located conditions were calculated within each condition group. SRM increased from 4 to 6.6 dB as the number of interferers increased from 2 to 12, reaching a maximum of 6.9 dB with 8 talkers. Individual values range from a minimum of 2.4 dB with 2 interferers to a maximum of 8 dB with 12 interferers. However, as interferers were moved away from the target side, no gradual SRM can be observed. Instead, SRT values decrease modestly as interferers are shifted away from the target side, and drop significantly, once all the maskers are separated from the target.

In Fig. 2, SRM values are denoted as R_y^x, and refer to the SRT benefit obtained by removing the last x co-located interferers, calculated as follows:

$$R_y^x = \mathbf{SRT}(C_y^x) - \mathbf{SRT}(C_y^0). \qquad (1)$$

As we can see, there appears to be no effect of overall number of interferers on the SRM from removing the last 2 or 4 co-located maskers. This was confirmed by paired t-tests with Bonferroni correction.

Comparison of the SI and lateralization results. Investigation of the SI and lateralization tests was done by correlation analysis. As listeners used a different paradigm for the detection task in the fully co-located conditions, these results were excluded from the analysis. Nevertheless, these are also plotted on Fig. 3 (filled squares), which presents a scatterplot of all individuals' lateralization vs. SI results in all tested conditions. For each listener, SRT and lateralization thresholds were significantly correlated for all conditions where at least one of the maskers was co-located with the target. Across listeners, the correlation had an average r^2=0.93, ranging from a minimum of r^2=0.803 to a maximum of r^2=0.976. Across individual listeners, the slopes of the regression lines varied between 0.46 – 0.60.

4. Discussion

In this study, speech reception thresholds and lateralization thresholds were measured for normal hearing subjects in time-reversed babble noise, to investigate binaural unmasking in complex auditory settings. Both speech and masker stimuli were lateralized using fixed ITDs. Results showed that while increasing the number of interferers increased the SRT, it did not affect the masking release per masker stream.

The specific pattern of the SRT values can be attributed to two main effects. One is a purely monaural effect which can be observed between condition groups, directly related to the number of interferers. The other can be considered as a binaural effect arising from the differences in the spatial distribution of the maskers, appearing within condition groups.

In general, results for the SI tests show very low SRTs. For the most difficult condition (C_{12}^{12}) an SRT of -7.7 dB was observed, which is comparable to the reference SRT value of -8.3 dB measured in speech shaped noise for the DANTALE II corpus [5]. In all the other conditions, lower threshold levels were expected, either because of fewer interfering talkers in the scene or because of a more favorable perceived spatial distribution of maskers.

Previous studies in speech-on-speech masking tasks have shown that the identification performance of VCV tokens and words varies as a non-monotonic function of the number of interferers in the scene [7, 8]. It is argued that with a large number of interfering talkers, the individual maskers are merged into an unintelligible mixture of speech streams. This babble represents a masker with low informational content, but with high energetic masking potential, offering little or no fluctuations in the spectro-temporal domain. On the other hand, with a few number of interferers, listeners are subjected to a large amount of informational masking; however, they are able to listen to the target in the spectro-temporal dips of the noise. The informational- and energetic components have an opposite effect on performance with increasing number of interferers in the masker, eventually having the most deteriorating effect with an intermediate number of talkers. Thus, our findings of a monotonic increase of SRT values with the overall number of interferers for the fully co-located (C_x^x) and fully separated (C_x^0) conditions are inline with the studies mentioned above. The informational masking content of the masking streams was removed, or at least substantially reduced, by using time reversed speech produced by talkers of the opposite gender to that of the target. Thus, the observed pattern can be attributed mostly to dip listening: as the number of maskers increases, envelope fluctuations in the masking noise gradually disappear, providing less opportunity to the listener to benefit from glimpsing at the target in the spectro-temporal gaps of the masker. As a consequence, an increased SNR is required for the same criterion performance. These gaps in the masker stream are already reduced significantly with 8 maskers, thus no apparent disadvantage arises from adding 4 more maskers (i.e., for a total of 12).

The observed trends within condition groups can be attributed to binaural unmasking. As summarized by Bronkhorst, masking release due to binaural unmasking typically ranges from 1 to 7 dB [9], depending on the number and spatial configuration of the talkers, and on the speech material itself. In this study, the SRM values are found to be within the range mentioned above. The benefit due to binaural unmasking is greatest, once all the interferers are separated from the target.

Interestingly, the benefit from spatially separating the final 2 co-located maskers was the same, regardless of the total number of maskers. The same pattern of results (i.e., no effect of total number of masker streams) was observed for the benefit of spatially separating the final 4 co-located maskers. Thus, the influence of perceived spatial separation and overall complexity of the scene on speech intelligibility are independent.

Based on these observations, we suggest revisiting the utility of spatial release from masking. Traditionally, the fully co-located condition is used as a reference. Thus, the magnitude of *release from masking*, achieved by moving away a certain number of interferers from the target side, depends on the overall number of interferers in the scene. In contrast, if the fully separated condition is used as a reference, differences in performance can be measured as *addition of masking*. Importantly, our observations suggest that addition of masking (e.g., increasing from n to $n+1$ co-located maskers) is not dependent on the overall number of maskers in the scene. By decoupling overall complexity and spatial distribution, measuring addition of masking rather than release from masking could make interpretation of intelligibility differences across conditions easier.

While only normal-hearing listeners were tested in the present study, the results suggest that a similar paradigm might be useful as a clinical measure. For each individual, high correlations were observed between lateralization and speech intelligibility thresholds. Thus, differences in lateralization detection might be suitable as a proxy measure to predict the benefit of binaural unmasking on speech intelligibility. In this context, there are two potential advantages of using a lateralization detection task over a direct speech intelligibility task. First, the larger range of lateralization thresholds across conditions suggests that it may be a more sensitive clinical measure. Second, the lateralization task is language independent. Thus, there is no need to develop a native speech corpus for each language. As the spatial cues in the present study varied only in ITD, a similar paradigm, with only a few conditions, could be used to evaluate whether a hearing-impaired listener could make use of such cues. This could be useful in determining if a patient would gain any advantage with binaural hearing aids vs. two monaural devices. However, further studies involving hearing-impaired listeners are needed to test this.

5. Conclusion

In the present study, the effect of perceived spatial separation and the total number of masker streams was systematically investigated. In general, little spatial release from masking was observed when at least one masker remained co-located with the target. The benefit of removing the last 2 or 4 co-located maskers was found to be independent of the overall number of maskers. This suggests that considering addition of masking rather than release from masking may be a more useful approach. Finally, for each individual, lateralization and speech intelligibility thresholds were highly correlated. Thus, for situations where measuring direct speech intelligibility may be less practical, lateralization tasks could be used to predict changes in speech intelligibility.

Acknowledgement

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no FP7-PEOPLE-2011-290000.

References

[1] E. C. Cherry: Some experiments on the recognition of speech, with one and with two ears. J. Acoust. Soc. Am. **25** (1953) 975-979.

[2] J. Peissig, B. Kollmeier: Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners. J. Acoust. Soc. Am. **101**(3) (1997) 1660-1670.

[3] A. W. Bronkhorst, R. Plomp: Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing. J. Acoust. Soc. Am. **92**(6) (1992) 3132-3139.

[4] M. L. Hawley, R. Litovsky, J. F. Culling: The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. J. Acoust. Soc. Am. **115**(2) (2004) 833-843.

[5] K. Wagener, J. L. Josvassen, R. Ardenkjær: Design, optimization and evaluation of a Danish sentence test in noise. International Journal of Audiology **42**(1) (2003) 10-17.

[6] M. Cooke, J. Barker, S. Cunningham, X. Shao: An audio-visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Am. **120**(5) (2006) 2421-2424.

[7] S. A. Simpson, M. Cooke: Consonant identification in N-talker babble is a nonmonotonic function of N. J. Acoust. Soc. Am. **118**(5) (2005) 2775-2778.

[8] S. Rosen, P. Souza, C. Ekelund, A. A. Majeed: Listening to speech in a background of other talkers: Effects of talker number and noise vocoding. J. Acoust. Soc. Am. **133**(4) (2013) 2431-2443.

[9] A. W. Bronkhorst: The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. Acta Acustica united with Acustica **86**(1) (2000) 117-128.