PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Improved binaural speech intelligibility by adding reverberation to the target speaker

Julian Grosse, Steven van de Par

Acoustics Group, Cluster of Excellence "Hearing4All", University of Oldenburg, Germany

ABSTRACT

In anechoic, multi-source listening situations, the auditory system is remarkably capable of understanding speech. An improvement in speech intelligibility is observable when interfering and target speakers are spatially separated compared to being co-located. The present study investigates the influence of reverberation on speech intelligibility of a reverberated target placed in front of the listener in the presence of symmetrically placed anechoic interferers. A simulated binaural room impulse response (BRIR) is systematically truncated at different durations to investigate the effect of reflections. Measurement conditions consisted of co-located and separated speakers that were presented, using BRIRs for the target speaker and HRTFs for the interferers. It was found that reflections can have an opposite effect on speech intelligibility. Whereas for spatially separated speakers, added reverberant components impair speech intelligibility, for co-located conditions speech intelligibility improves. This suggests that the auditory system may utilize the spatial diffuseness of the target source in co-located conditions to perceptually segregate it from the interfering anechoic sources. In spatially separated conditions, the difference in perceived location of the sources may be the dominant cue for segregation such that diffuseness caused by adding reverberation does not further improve intelligibility.

Keywords: Binaural speech intelligibility, room-acoustics, reverberation

1. INTRODUCTION

It is well known that in multi-talker conditions, the speech intelligibility improves when the target speaker is spatially separated from the interfering speakers [1]. Various contributing factors have been mentioned such as better ear listening, binaural unmasking, and the contribution of binaural cues to auditory stream segregation [1,2].

Modelling approaches that can predict the improvement in speech intelligibility are often based on Equalization and Cancellation processing [3, 4, 5] which was originally proposed in the context of basic psychoacoustic experiments of tone in noise masking [6]. The essential idea is that the left and right input signals that occurs at a certain moment in time within a specific frequency band is first temporally aligned and equalized in level such that after subtracting of the modified signals the interferer is mostly cancelled thus improving the target speech to interfering speech ratio.

The presence of reverberation can impair speech intelligibility. The precise reasons are not yet well understood. In the context of an Equalization and Cancellation model, the reverberation will lower the correlation of the interfering signals at the two ears which will make cancellation less effective. Alternatively, spatial cues will be modified due to the presence of reverberation which could make their contribution to auditory stream segregation less effective. The Equalization and Cancellation model has been extended to include the effect of reverberation by partitioning the reverberated target speech in components that contribute to speech intelligibility (direct sound and early reflections) and components detrimental for speech intelligibility (late reverberation) [7,8]. These latter components are considered as additional masking components [8].

In this contribution, we were interested to see how well a reverberated target speaker could be understood in the presence of two anechoic interfering speakers. Both co-located and spatially separated configurations will be investigated as well as diotic presentations. More specifically, the Binaural Room Impulse Response (BRIR) will be truncated after various cut-off times. In this way, the effect of early and late reflections on speech intelligibility can be investigated in order to determine the extent to which they are detrimental to speech intelligibility. As will be shown, the expectation that more reverberation should impair speech intelligibility does not always hold and depends on the spatial configuration of the target and interfering speakers.

2. Method

The following section describes the experimental setup, i.e. the room-acoustical conditions, the spatial orientation of the interfering and target speakers and the manipulation of the BRIR to investigate the role of single reflections and reverberation on binaural speech intelligibility.

2.1 Room-acoustical conditions and spatial locations of the speech sources

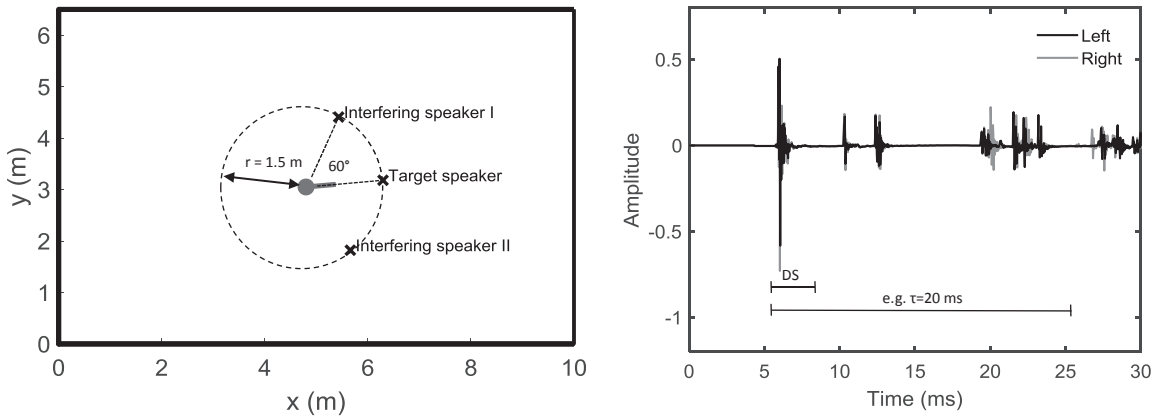

Figure 1 – Room-acoustical conditions that were used in the subjective evaluation. Panel A (left) shows the room dimensions, the orientation of the listener relative to the location of the interfering and target speakers, respectively. Panel B (right) shows the first 30 ms of the BRIR at the target speaker location with an example truncation time τ of only direct sound (DS) or e.g. 20 ms.

In order to investigate the role of reflections and reverberation on speech intelligibility in rooms, the room acoustic simulation software RAZR [9] has been used to simulate binaural room impulse responses (BRIR). It combines an image source model to simulate early reflections and a feedback delay network for the late reverberant part.

Figure 1 shows the dimensions of the room and spatial orientation of the speakers that were used in this study. Table 1 shows additional room properties. The position of the listener is somewhat off-centered and the facing direction is slightly rotated to avoid a symmetrical setup. The target speaker is located in front of the listener at a distance of 1.5 m. The interfering speakers are located at -60° and +60° to the left and right relative to the target speaker. The distance between listener and interfering speakers is kept constant to 1.5 m to have an equal direct sound level independent of the targets' location.

The reverberation time (RT) is controlled via the RAZR simulation software by adjusting the absorption coefficients that are equally distributed across the walls to obtain a RT of 0.8 s and 1.6 s. This adjustment allows to preserve the temporal structure (in terms of the acoustical path lengths) of the early reflections. Consequently, only the strength of the late reflections changes.

Table 1 – Shown are the room properties, i.e. room dimensions, reverberation time, direct to reverberant ratio and the distance between listener and speech source.

	Dimensions (x,y,z)	Reverberation time	DRR	Distance
Room A	10 m x 6.5 m x 3 m	0.8 s	-0.4 dB	1.5 m
Room B	10 m x 6.5 m x 3 m	1.6 s	-2.9 dB	1.5 m

2.2 Stimuli & subjective procedure

The target speech stimuli consisted of ongoing speech tokens (spoken by a male) from the German version of the OLSA-speech corpus [10]. The sentences consist of 5 words (name, verb, numeral, adjective, subject) randomly chosen and without any semantical meaning with a duration of approx. 2.5 s. As interfering speech stimuli, two different female speakers were taken from two audiobooks spoken in German language. The two independent interfering speech tokens were truncated to a total interval-duration of 3.5 s. The target speech was temporally centered within the interferers resulting in an onset-delay between interferer and target speaker.

The spatially separated condition were created by convolving the tokens with the BRIR stemming from one of the three locations. The target speech originated always from 0° and the two interferers either from -60° or +60° allowing the auditory system to make use of binaural cues to segregate the target and interfering speakers or to use better-ear listening to improve speech intelligibility.

The co-located condition consisted of the target and the two interferers originated from 0° which does not allow the auditory system to make advantage from binaural cues but still allows better-ear listening

The diotic condition consisted of the left ear signal of the co-located condition but was presented to both ears. The left ear was arbitrarily chosen. This monaural signal contained no binaural information and does not allow the auditory system to make use of neither binaural cues nor better-ear listening.

In order to investigate the role of early and late reflections on binaural speech intelligibility, only the target speakers' BRIR was manipulated and the interfering speakers' BRIR only consisted of the direct sound (DS) to convey only undistorted directional information to the listener without any reflections. The target speaker BRIR was truncated to:

a) The direct sound (DS) to simulate anechoic/free-field condition,
b) A duration of 10 ms, in which only ground and ceiling reflections occurred (i.e. monaural information of the reflections)
c) 20 ms which includes ground and ceiling (monaural) and left and right-wall (binaural) reflections
d) 125 ms to cover all early reflections but without the late reverberation
e) The full impulse response (full IR) that is technically defined by the reverberation time illustrated in Table 1.

The symmetrically placed interferers were presented **anechoic** by convolving the tokens with the direct sound of the BRIR to obtain a binaural signal with the interaural cues stemming from the proper location of the interferers with respect to the spatial condition.

In order to investigate the effects of additional reflections on speech intelligibility and to avoid an additional loudness cue caused by the reflections, the level was normalized independent of the duration of the truncated BRIR.

The signal to noise ratio (SNR) was controlled by taking the mean-value across both ears of the root-mean-squared value of the binaural signals after spatialization.

In order to measure speech reception thresholds (SRTs), the interfering talkers were calibrated to a constant level of 65 dB-SPL. To reach the 50% speech intelligibility point, the SNR between the target and interfering talkers is adaptively adjusted depending on the answer of the subject. To obtain the SRT50, a total set of 15 sentences per condition were used.

Eight subjects in total participated in the subjective evaluation. Subjects were normal-hearing (pure tone audiogram<20 dB). They participated in two sparate blocks at which one block consisted of 15 threshold measurements (3 spatial coditions x 5 truncations times) and each had a duration of approximately 90 minutes, but subjects were instructed to have a rest whenever it was necessary. Before a daily block, subjects had an initial training of six conditions which spanned a variety of the three spatial conditions but were not part of the main experiment. All 15 conditions were randomized across subjects whereas a room-acoustical condition was not divided across the two daily blocks because literature shows that prior listening in rooms can improve speech intelligibility [11].

3. Results

Figure 2 shows the results of the subjective evaluation for both room-acoustical scenarios. Illustrated are speech reception thresholds (SRT) at 50 % speech intelligibility displayed by mean values and standard error across subjects for the spatially separated, co-located and diotic condition. Panel (A) shows the results for the room with 0.8 s and (B) with 1.6 s.

When both, the target and the interfering talkers are presented in anechoic condition, i.e. only the direct sound but with the binaural cues stemming from the particular location, the spatial separated talker configuration shows a binaural benefit of about 5 to 6 dB compared to co-located and diotic condition. This result is independent of the RT, because the level of the direct sound was kept constant across room-acoustical condition. This behavior can also be seen for a truncation time of τ=10 ms. The small differences can be attributed to the two additional (purely monaural) reflections stemming from the ground and ceiling and cause a comb-filter like effect in the spectra but does not add additional binaural information to the target stimulus.

For the separated condition, the SRTs monotonically decrease from approx. -13 dB to -11 dB for room A and to -9 dB for room B with increasing truncation time. Main differences between the rooms are observable when the full BRIR is presented and results in SRTs of approx. 2 dB across spatial and room-acoustical conditions. This difference can be mainly attributed to the change of the strength of the late reflections.

The co-located condition shows with increasing truncation time an increase in SRTs of about 2 dB for room A and approx. 4 dB for room B compared to the anechoic condition. This behavior seems to be independent of the strength of the late reflections and converges to the separated condition after a truncation time of 125 ms.

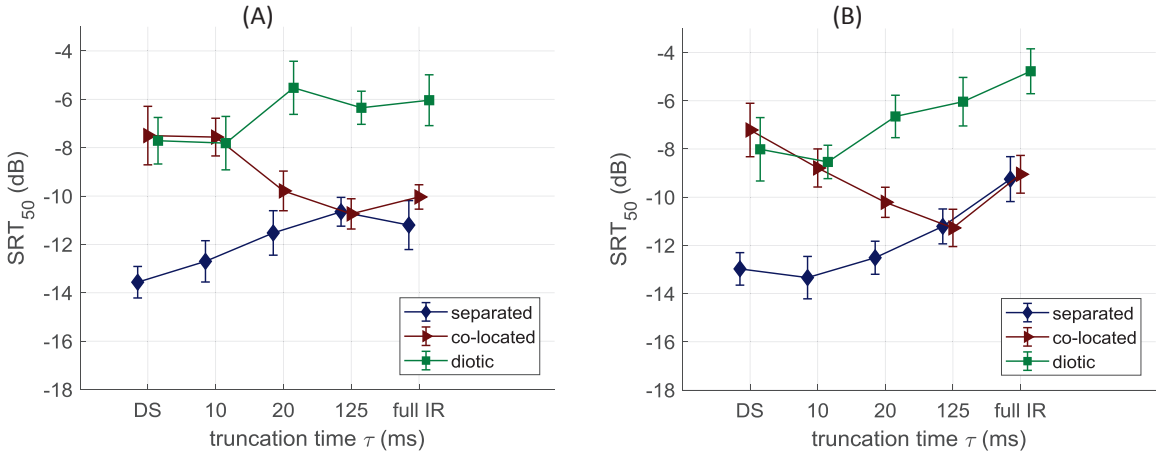

Figure 2 – Speech reception thresholds (SRT50) measured in the presence of two anechoic interferers for room 1 (A) with a reverberation time (RT) of 0.8s and room 2 (B) with RT=1.6s. The x-axis represents the truncation time τ (ms) of the BRIR of the target speaker location and the y-axis represents the SRTs. Illustrated are the mean across subjects and the standard error of the mean for the spatially separated (blue diamonds), the co-located (red triangles) and the diotic condition (green squares).

The diotic condition reflects a constant behavior till a truncation time of 10 ms for room A, increases to a truncation time of 20 ms and is saturated for the remaining part of the BRIR. Room B shows a similar behavior of increasing SRTs after 20 ms even though the increase for room B is more prominent and no saturation in SRTs are visible.

Comparing the separated and co-located with the diotic condition measured with the full BRIR within and across both rooms, the binaural benefit in SRTs is approx. 5 dB for room A and 4 dB for room B when a reverberant target speaker is masked with anechoic interfering speakers.

4. Discussion

In the speech intelligibility experiments we presented in this contribution using anechoic interfering speakers and a target speaker that was convolved with a Binaural Room Impulse Response (BRIR) truncated at various truncation times, we found that for the spatially separated condition, an increase in truncation time leads to reduces speech intelligibility (increased SRTs). This finding is in line with the common understanding that speech intelligibility is impaired by added reverberation. Note that the target signal was always level normalized and thus was independent of the truncation time. This implies that with increasing truncation time, the direct sound component and early reflections decrease in level. Already at 20ms truncation time, we see an impairment in speech intelligibility indicating that for speech intelligibility it is better to concentrate all target signal energy within the direct sound and at most first 10 ms early reflections.

Rather contrary to the findings with separated speakers, the co-located condition showed that with longer BRIRs applied to the target speaker, speech intelligibility improves (i.e. SRTs decrease) up to truncation times of 125 ms. Thus, this leads us to conclude that reverberant components up to 125 ms contribute to speech intelligibility.

An explanation for this finding may be that for the co-located condition, the anechoic interfering speakers and the target in case it is rendered only with a direct sound component create a spatial image that provides no spatial cues that would allow to segregate the target speaker from the interfering speakers. When, however, early reflections are added to the target speaker, the spatial diffuseness of the target increases, and now the spatial image of the target speaker extends beyond the position of the interfering speakers. This would allow to spatially segregate the target from the interferers. A further explanation for this improvement in SRTs might be due to the binaural information conveyed by the first order reflections (sidewalls) that can be used to improve speech understanding and which are in our set-up is co-located with the interfering speakers in the spatially separated configuration.

It is interesting that the BRIR components in the range from 20 ms until 125 ms can be both detrimental, in case of separated speech sources, and useful, in case of co-located speech sources.

To support the contribution of spatial diffuseness cues, results can be compared to the diotic conditions. In this case neither spatial localization cues, nor spatial diffuseness cues are available. In this case, a pattern of SRTs is obtained that is similar to the separated condition, only for the diotic condition SRTs are shifted up by about 4 to 6 dB.

Possibly the pattern of results for the co-located conditions could be predicted by an Equalization Cancellation model for which the Equalization stage attempts to cancel the interfering speakers. When spatial diffuseness increases, more target speech components will remain after cancellation. Such an EC model would, however, in order to model the diotic conditions, require some stage that allocates some of the late BRIR parts of the target speech to be detrimental components that mask the early target speech components. Thus, this leads us to conjecture that the determination of what are detrimental signal components will depend on spatial configuration in line with a similar conclusion reached in [7].

Note also that typically Equalization Cancellation models do not include an explicit stage modelling auditory stream segregation contributions to speech intelligibility although this has been shown to be an important contributor to speech intelligibility [12].

In the experiments shown here, two different T60 time conditions of 0.8 and 1.6 s were used. Despite the large difference in these T60 times, no strong difference in speech intelligibility was found. This may be related to the relatively small source-receiver distance of 1.5 m which creates a relatively strong direct sound component in the BRIRs and thus a relatively similar direct reverberant ratio (see Table 1).

To summarize these results show that target signal components between 20 ms and 125 ms can both be detrimental and useful for speech intelligibility.

5. Summary & Conclusion

We found that reverberation reduces speech intelligibility in case target and interferers are spatially separated. In contrast to this, when adding reverberation only to a target speaker that is co-located with anechoic interfering speakers, the reverberation leads to an improvement of speech intelligibility. It was hypothesized that the spatial diffuseness of the target speaker helps to better segregate this speaker from the interferers.

ACKNOWLEDGEMENTS

The authors acknowledge the Deutsche Forschungsgemeinschaft for supporting this work that is part of the grant titled "Contributions of Auditory Stream Formation to Speech Perception."

REFERENCES

1. Bronkhorst, W.A., The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions, *Acustica*, 86, 2000:117–128.
2. Schoenmaker, E., Brand, T., van de Par, S. The multiple contributions of interaural differences to improved speech intelligibility in multitalker scenarios, *The Journal of the Acoustical Society of America* 128, 2016:2589-2603.
3. Wan, R., Durlach, N. I., and Colburn, H. S. Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers, *The Journal of the Acoustical Society of America* 128, 2010:3678–3690
4. Lavandier, M., and Culling, J. F. Prediction of binaural speech intelligibility against noise in rooms," *The Journal of the Acoustical Society of America* 127, 2010:387–399
5. Beutelmann, R., Brand, T., and Kollmeier, B. Revision, extension, and evaluation of a binaural speech intelligibility model, *The Journal of the Acoustical Society of America* 127, 2479–2497
6. Durlach, N. I. Equalization and cancellation theory of binaural masking-level differences, *The Journal of the Acoustical Society of America* 35, 1963:1206–1218
7. Rennies, J., Warzybok, A., Brand, T. and Kollmeier, B. Modeling the effects of a single reflection on binaural speech intelligibility, *The Journal of the Acoustical Society of America* 135, 2014:1556-1567
8. Leclere, T., Lavandier, M., Culling, J.F. Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking, and binaural de-reverberation, *The Journal of the Acoustical Society of America* 137, 2015:3335-3345
9. Wendt T, van de Par S, Ewert S. A computationally-efficient and perceptually-plausible algorithm for binaural room impulse response simulation. Journal of the Audio Engineering Society 62, 2014: 748-766.
10. Wagener K, Brand T. Kollmeier B. Entwicklung und Evaluation eines Satztests für die deutsche Sprache I-III: Design, Optimierung und Evaluation des Oldenburger Satztests. Zeitschrift fuer Audiologie, *38*(1-3), 1999 4-15.
11. Brandewie E, Zahorik P. Prior listening in rooms improves speech intelligibility. *The Journal of the Acoustical Society of America* 128, 2010:291-299.
12. Schoenmaker, E., van de Par, S. Intelligibility for Binaural Speech with Discarded Low-SNR, *In P. van Dijk et al. (eds.), Speech Components, Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing, Advances in Experimental Medicine and Biology*, 894, 2016:73-81