Prediction of binaural intelligibility level differences in reverberation

Jan Rennies[1], *Thomas Brand*[2], *Birger Kollmeier*[1,2]

[1]Fraunhofer IDMT Hearing, Speech and Audio Technology, Oldenburg, Germany
[2]Medical Physics, University of Oldenburg, Oldenburg, Germany
jan.rennies@idmt.fraunhofer.de, (thomas.brand, birger.kollmeier)@uni-oldenburg.de

Abstract

Speech intelligibility can be substantially improved when speech and interfering noise are spatially separated. This spatial unmasking is commonly attributed to effects of head shadow and binaural auditory processing. In reverberant rooms spatial unmasking is generally reduced. In this study spatial unmasking is systematically measured in reverberant conditions for several configurations of binaural, diotic and monaural speech signals. The data are compared to predictions of a recently developed binaural speech intelligibility model. The high prediction accuracy ($R^2 > 0.97$) indicates that the model is applicable in real rooms and may serve as a tool in room acoustical design.

Index Terms: speech intelligibility, binaural hearing, reverberation

1. Introduction

In real communication environments like e.g. working or public places, speech intelligibility is often reduced by reverberation and interfering noise. In practical applications speech intelligibility models are widely used, e.g. for the room acoustical design of such places, as an evaluation tool of methods to enhance intelligibility, or in speech diagnostics. In this study a recently developed binaural speech intelligibility model is evaluated for a number of conditions involving several aspects of binaural hearing in different degrees of reverberation. The goal is to investigate the mechanisms underlying effects of binaural hearing using a current intelligibility model and to test the model's general applicability in reverberant rooms.

The major advantage of using models to calculate intelligibility from sound signals in comparison to subjective measurements with real listeners is the great reduction of effort. The obvious prerequisite for the application of models is that they accurately account for the influence of the main factors affecting intelligiblity. Many years of research have eventually lead to standardized methods to calculate intelligibility, e.g. the speech intelligibility index (SII) [1] and the speech transmission index (STI) [2]. In practical applications, simpler technical measures are also readily used to describe the effects of reverberation on speech perception. For example, measures such as clarity, definition (also called "Deutlichkeit"), or useful-to-detrimental ratios describe the fact that reflections at the room boundaries do not disturb intelligibility as long as they arrive within a certain time after the direct sound.

All of the above mentioned methods to estimate intelligibility are essentially monaural models, which means that factors related to binaural hearing are not accounted for. This may lead to wrong predictions in certain conditions since many studies have shown that intelligibility can be largely enhanced when speech and noise sources are spatially separated [3, 4]. This spatial unmasking is commonly attributed to head shadow and binaural processing and can be quantified for example by comparing speech reception thresholds (SRTs), i.e. the signal-to-noise ratios (SNRs) at which 50% of the words can be understood. One effect of head shadow is better-ear listening, i.e. lower SRTs at the ear with the favorable SNR. To quantify binaural auditory processing the binaural intelligibility level difference (BILD) can be measured. It is calculated as the difference between a dichotic SRT measured with frontal speech and a noise source at the side and the monaural SRT measured with the ear closer to the noise source blocked. A further effect observed in binaural listening conditions is the so-called binaural squelch, which was measured e.g. in [5], where significantly improved SRTs were measured for binaural compared to diotic speech signals in a reverberant room.

To also account for such effects recent studies have proposed binaural speech intelligiblity models. Van Wijngaarden and Drullman [6] extended the STI by computing interaural modulation transfer functions and showed good prediction accuracy for different conditions including noise, bandwidth limitation, nonlinear distortion and reverberation. Lavandier and Culling [7] presented an efficient model combining an estimation of binaural processing based on interaural correlation with effects of better-ear listening in SII-weighted frequency bands. Their model was successful for a number of conditions comprising different noise azimuths, distances and rooms. While these models are very successful in the tested conditions, they still have their limitations. The STI as well as its binaural version [6] are inherently applicable only for stationary maskers, since the general concept assumes that temporal modulations are only caused by the speech portions of the signal. In many practical cases this assumption is not valid leading to overestimated intelligibility. The model of Lavandier and Culling [7] does not include the effects of reverberation on the speech signal and is therefore only applicable for anechoic or near-field speech. Rennies et al. [8] presented a model based on the work of Beutelmann et al. [9, 10]. Their model quantitatively accounted for effects of spatial unmasking in conditions with varying degrees of reverberation for different distances between speaker and listener. So far, however, the model was only tested in conditions with diotic speech signals [8].

The goal of this study was to investigate the above-mentioned effects of binaural speech intelligibility and their dependence on reverberation both experimentally and by means of model predictions. Experimental data were collected for several types of speech presentation (such as diotic, dichotic, and monaural speech) and degrees of reverberation. The data were compared to predictions of the model of Rennies et al. [8]. The underlying reasoning was that if the model was able to predict SRTs for the different conditions, further indication would be gathered for its general applicability, while possible discrepancies could be used to further refine the model.

2. Experiment

2.1. Subjects

Eight subjects with hearing thresholds $\leq$ 15 dB HL at audiometric frequencies between 125 and 8000 Hz participated in the experiments (seven volunteers and the first author of this paper). All had experience in speech intelligibility measurements.

2.2. Stimuli and acoustic conditions

The stimuli consisted of the German sentences and the test-specific noise signal of the Oldenburg sentence test [11]. The sentences of the test have the fixed syntactic structure *name verb numeral adjective noun*. For each word ten alternatives are available which can be randomly combined to produce unpredictable sentences. The noise was generated by randomly superimposing the sentences resulting in a stationary noise with a long-term spectrum very similar to that of the speech material. The signals were digitally manipulated by convolution with binaural room impulse responses (RIRs) to produce the desired spatial configurations. The RIRs were simulated using the CATT Acoustics software v8.0a. The simulated room (width: 10 m, length: 15 m, height: 3 m) as well as the positions of the listener, speech and noise source were always the same and are indicated in Fig. 1. The speech source was located in front of the listener, the noise source was placed about 40° to the right. The height of the listener and both sources was 1.2 m. Speech and noise source were simulated as omnidirectional sources, the listener was simulated as a KEMAR head and torso simulator. The absorption coefficients α of the walls were set to 0.05 (highly reverberant), 0.20, or 1.00 (anechoic). The value of 0.20 was chosen to facilitate the comparison to the data of [5].

For each absorption coefficient five different presentation modes were tested (see inner gray box in Figure 1): (i) unmanipulated speech using the original binaural RIRs (*binaural*), (ii) diotic speech generated by copying the left ear channel to the right channel (*diotic*$_\ell$), (iii) monaural speech containing only the left ear signal (*monaural*$_\ell$), (iv) monaural speech containing only the right ear signal (*monaural*$_r$), and (v) a condition in which the right ear channel was set to zero for both speech and noise (blocked$_r$). In conditions (i) to (iv), only the speech was manipulated while the noise was kept the same (i.e. convolved with the original binaural RIRs).

The noise was calibrated to 65 dB SPL at the right ear of the listener, the noise level at the left ear was generally lower due to head shadow and depended on the room absorption. In condition *blocked*$_r$ the level of the noise at the left ear was the same as in the other conditions. The speech levels were also adjusted at the right ear of the listener, except in condition *monaural*$_\ell$ where the calibration was the same as in condition *binaural*.

All signals had a sampling rate of 44.1 kHz and were presented to the subjects via free-field equalized Sennheiser HDA200 headphones in a sound-attenuating booth.

2.3. Procedure

For each condition a different test list was used to measure the speech reception threshold (SRT), i.e. the SNR at which 50% of the words could be understood. The subjects listened to a list of 20 sentences. After each sentence, they repeated the words they had understood and an experimenter marked the correctly repeated words. SRTs were determined using the adaptive procedure described in [12]. Briefly, the level of the sentence was increased when less than half of the words of the previous sentence had been understood, otherwise it was decreased. The step size of the level increments and decrements depended on the percentage of correct words and was reduced after each reversal following an exponential rule [12]. Using this procedure the speech level adaptively converges to the level at 50% intelligibility. All 15 combinations of absorption coefficients and presentation modes were presented randomly for each subject using a different test list for each condition. Prior to the actual measurements the subjects received a training of at least two lists of the standard Oldenburg sentence test and one list of this study (*binaural*, α=0.20) to familiarize the subjects with the stimuli and procedure and to reduce the effects of training [11].

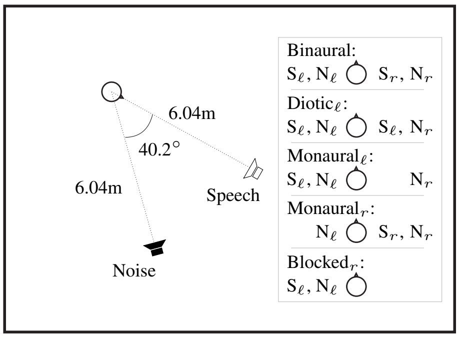

Figure 1: *Spatial configuration of receiver, speech and noise source in the simulated room (outer black line). In the inner box, S_ℓ, S_r, N_ℓ, and N_r indicate which speech (S) and noise (N) parts of the original binaural signal (ℓ: left, r: right) are presented to which ear in the five presentation modes.*

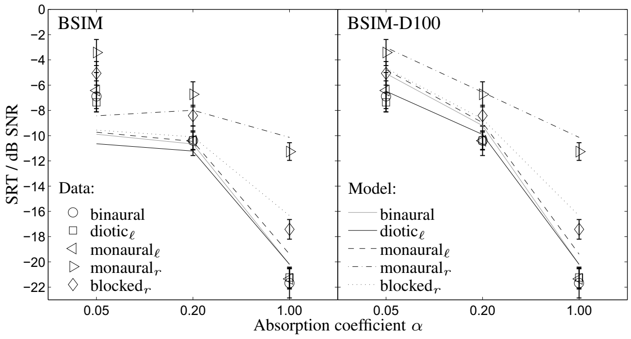

Figure 2: *Mean SRTs (symbols) and standard deviations across subjects for the five presentation modes. Lines represent the corresponding predictions of BSIM (left panel) [10] and its extended version BSIM-D100 (right panel) [8].*

2.4. Results

The mean SRTs and standard deviations across subjects are shown in Figure 2 (symbols) as a function of absorption coefficient. In general, SRTs decreased with increasing absorption coefficient indicating that reverberation impaired speech intelligibility. The decrease was larger between 0.20 and 1.00 than between 0.05 and 0.20. Standard deviations were in the order of 1 dB which shows that subjects responded consistently. The presentation modes also affected SRTs, which were highest for *monaural*$_r$, followed by *blocked*$_r$. For the remaining three presentation modes (*binaural*, *diotic*$_\ell$, and *monaural*$_\ell$), thresholds were the same for absorption coefficients of 0.20 and 1.00 and differed only slightly for 0.05. These observations were supported by a two-way analysis of variance (ANOVA) for repeated measures. Both factors absorption coefficient (F(2,14)=3079.29, p<0.001) and presentation mode (F(4,28)=544.33, p<0.001) were significant. The significant interaction between the two factors indicated that the effect of absorption coefficient depended on presentation mode. (F(8,56)=60.06, p<0.001). Post-hoc tests with Bonferroni correction showed that SRTs for presentation mode *blocked*$_r$ and *monaural*$_r$ differed significantly from all other modes, while *monaural*$_\ell$, *diotic*$_\ell$, and *binaural* did not differ significantly.

3. Modeling

3.1. Model structure

The model of Rennies et al. [8] is based on the binaural speech intelligibility model (BSIM) of Beutelmann et al. [9, 10]. BSIM processes speech and noise signals using an equalization-cancelation (EC) mechanism combined with the SII. The EC processing is performed independently in 30 auditory channels of a gammatone filter bank. The signals at the left and right ears are delayed and amplified relative to each other and subsequently subtracted. Gain and delay are chosen to optimize the SNR after subtraction. In conditions in which interaural level and/or phase differences are different for speech and noise, the SNR can be improved relative to the monaural SNRs. The improved SNR is used to calculate the SII as in the monaural standard [1]. The resulting SII is transformed into an intelligibility value using a nonlinear transform which depends on the speech material and measurement method for which intelligibility is to be predicted [9]. The SRT for a given acoustic condition is derived by selecting a fixed reference SII value and varying the SNR until the SII equals this reference value. Hearing thresholds are included in the model as uncorrelated noise at the two ears which cannot be canceled by the EC mechanism. This allows predictions of speech intelligibility also for hearing-impaired listeners.

Rennies et al. [8] extended BSIM by including room acoustical parameters into the model. They used the quantity definition ("Deutlichkeit") as a post-hoc correction for the binaurally enhanced SNRs after the EC-mechanism. Definition is defined as the ratio between the energy of the early reflections and the entire energy of the RIR $h(t)$ [14]

$$D_{te} = \frac{\int_0^{te} h(t)^2 dt}{\int_0^{\infty} h(t)^2 dt}, \tag{1}$$

where te is the time separating the early from the late reflections. Using te=100 ms, the model (called BSIM-D100 in the following) quantitatively predicted SRTs for different conditions ranging from quasi free-field to highly reverberant. In this study, all parameters of the model were kept the same. Predictions were made for all experimental conditions both with the original BSIM and the extended model to investigate the role of the extension.

3.2. Results

Lines in Figure 2 represent the predictions of BSIM (left panel) and its extended version BSIM-D100 (right panel) for the five presentation modes as functions of absorption coefficient. As found in the experiment, predicted SRTs decreased with increasing absorption coefficient and were highest for *monaural*$_r$, followed by *blocked*$_r$. Predictions for the other three presentation modes were similar, although they differed slightly more than the experimental data. The predictions of the two models were the same for an absorption coefficient of 1.00. This was expected since - without any reverberation - the correction introduced in BSIM-D100 reduces to a factor of 1 (see Eq. 1). Decreasing the absorption coefficient lead to deviations of the models' predictions. While both models predicted SRTs for an absorption coefficient of 0.20 reasonably accurate, SRTs predicted by BSIM hardly increased further for the lowest absorption coefficient. In contrast, BSIM-D100 predicted increasing SRTs which was in line with the observed data. In general, BSIM slightly overestimated SRTs when reverberation was absent and underestimated SRTs in strong reverberation. In contrast, the predictions of BSIM-D100 more consistently represented the course of the measured SRT, although most SRTs were slightly overestimated.

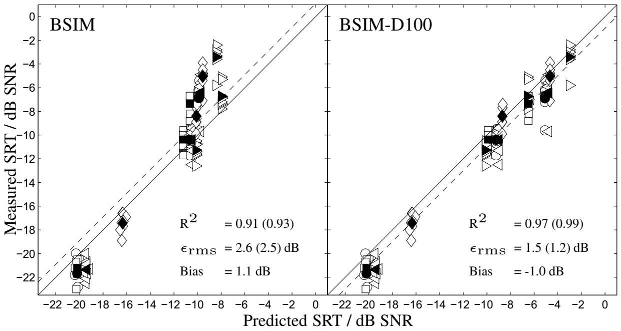

Figure 3: *Measured versus predicted SRTs for all experimental conditions for BSIM (left) and BSIM-D100 (right). The symbols are the same as in Figure 2. Open and filled symbols denote individual and mean data, respectively. The coefficient of determination* R^2*, rms-error* ϵ_{rms}*, and bias are indicated for the individual data and the mean data (in brackets).*

To further evaluate the model predictions a correlation analysis was conducted. Figure 3 shows measured SRTs for all presentation modes and absorption coefficients and the corresponding predictions of the two models. The symbols are the same as in Figure 2. Open symbols represent individual SRTs, filled symbold denote mean data. Three measures of prediction accuracy were computed: the coefficient of determination calculated as the square of the correlation coefficient, the linear offset calculated as the horizontal or vertical distance between the ideal mapping (solid line) and a best fit of unity slope (dashed line), and the root-mean-square (rms-)error ϵ_{rms} between data and predictions. The coefficient of determination can be interpreted as the fraction of variance in the data which can be explained by the model, while the offset quantifies the general bias of the predictions. In general, both models showed a reasonable prediction accuracy with $\mathrm{R}^2 > 0.91$ based on individual data and $\mathrm{R}^2 > 0.93$ based on mean data. The magnitude of the bias was about 1 dB for both models, although, on average, BSIM predicted slightly too low SRTs while SRTs predicted by BSIM-D100 were slightly too high. Correlation and rms-error were better for BSIM-D100 than for BSIM, which was especially due to the better predictions in strongly reverberant conditions.

4. Discussion

The data of this study are in line with previous studies showing that reverberation impairs speech intelligibility. The magnitude of threshold shifts depends on the degree of reverberation and on the presentation mode of the speech signal. Highest SRTs were obtained when speech was only presented to the ear closer to the noise source (right ear). When speech was presented to the left ear only, significantly lower SRTs were observed. This effect is due to head shadow effects resulting in a favorable SNR at the left ear. SRTs measured for diotic and binaural speech did not differ from monaural SRTs (left ear). This indicates that better-ear listening was a major factor affecting SRTs in this experiment and no further benefit could be achieved by additionally providing the same speech signal (*diotic*$_\ell$) or the original speech signal of the simulated RIRs (*binaural*) at the right ear. Evidence for binaural processing could be observed by comparing the SRTs for the conditions *blocked*$_r$ and *binaural*. SRTs were always lower when speech was presented as *binaural* compared to the *blocked*$_r$ condition, although the additional information came from the ear with an unfavorable SNR. This effect is commonly referred to as binaural intelligibility level difference (BILD). The data of this study showed that the BILD depended on reverberation, amounting to about 4 dB for α=1.00 and about 2 dB for lower absorption coefficients. The BILD was predicted by both models for α=1.00, but slightly underestimated for the reverberant conditions (about 0.5 dB).

It is interesting to observe that SRTs obtained with *diotic*$_\ell$ and *binaural* speech did not differ for any absorption coefficient. This was expected for α=1.00 since without reverberation diotic and binaural speech are the same for frontal presentation. For an absorption coefficient of 0.20, however, a previous study [5] found a small but significant benefit of binaural over diotic speech presentation. One possible reason for this discrepancy is a difference in stimuli. In [5], the impulse responses for the noise source were simulated with a different absorption coefficient (0.50) than the speech signal. In addition, the simulated room and the distances between sources and listener were larger in the present study and the listener was placed further away from the center of the room. The fact that speech presented as *monaural*$_\ell$ lead to the same SRTs as *diotic*$_\ell$ and *binaural* speech suggests that the spatial configuration used here lead to SRTs dominated by monaural listening leaving not much room for benefit due to diotic and/or binaural presentation. For α=0.05 SRTs differed marginally (and not significantly), being highest for *monaural*$_\ell$ and about 0.5 and 1 dB lower for *binaural* and *diotic*$_\ell$ speech, respectively. While the present data certainly do not allow to draw conclusions from these small differences, it is interesting to see that the models predict the same trends. Future measurements are necessary to explore the potential squelching effect in reverberation in more detail.

Altogether, the extended binaural speech intelligibility model has been shown to accurately predict SRTs for different noise directions and distances [8] as well as for different speech presentation modes and degrees of reverberation (present study). It therefore seems applicable for technical purposes such as e.g. room acoustical planning or evaluation of binaural speech enhancement algorithms.

5. Conclusions

The following conclusions can be drawn from this study:

- For each of the tested speech presentation modes involving better-ear listening and binaural processing, reverberation affects SRTs of normal-hearing listeners.
- No benefit of binaural over diotic speech (i.e. no binaural squelch) was observed in this study, but has been found in previous studies [5].
- An extended binaural speech intelligibility model [8] accurately predicts the observed SRTs for the different presentation modes. The comparison to the original model [10] shows that, especially in strongly reverberant conditions, the deterioration of the speech signal has to be accounted for.
- The accuracy of the extended model makes it in principle applicable in real environments with different degrees of reverberation.

6. Acknowledgements

This work was supported by the Universitätsgesellschaft Oldenburg and the Deutsche Forschungsgemeinschaft (SFB TRR 31).

7. References

[1] American National Standards Institute , ”Methods for calculation of the speech intelligibility index”, American National Standard S3.5-1997, 1997.

[2] International Electrotechnical Commission, ”Sound System Equipment - Part 16: Objective rating of speech intelligibility by speech transmission index”, International Standard IEC 60268-16, 2003.

[3] Cherry, E. C., ”Some experiments on the recognition of speech, with one and with two ears”, J. Acoust. Soc. Am. 25: 975–979, 1953.

[4] Bronkhorst, A., ”The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions”, Acustica 86: 117–128, 2000.

[5] Lavandier, M. and Culling, J.F., ”Speech segregation in rooms: Monaural, binaural, and interacting effects of reverberation on target and interferer”, J. Acoust. Soc. Am. 123: 2237–2248, 2008.

[6] van Wijngaarden, S.J. and Drulman, R., ”Binaural intelligibility prediction based on the speech transmission index”, J. Acoust. Soc. Am. 123: 4514–4523, 2008.

[7] Lavandier, M. and Culling, J.F., ”Prediction of binaural speech intelligibility against noise in rooms”, J. Acoust. Soc. Am. 127: 387–399, 2008.

[8] Rennies, J., Brand, T., and Kollmeier, B., ”Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet”, J. Acoust. Soc. Am.: *under review*.

[9] Beutelmann, R. and Brand, T., ”Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners”, J. Acoust. Soc. Am 120: 331–342, 2006.

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

[11] Wagener, K., Kühnel, V., and Kollmeier, B., ”Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests (Development and evaluation of a German sentence test I: design of the Oldenburg sentence test)”, Z. Audiol. 38: 4–15, 1999.

[12] Brand, T. and Kollmeier, B., ”Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests”, J. Acoust. Soc. Am. 111: 2801–2810, 2002.

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

[14] International Organization for Standardization, ”Acoustics - Measurement of room acoustic parameters - Part 1: Performance spaces”, International Standard ISO 3382-1:2009, 2009.