Acta Acustica 2022, **6**, 21

https://doi.org/10.1051/aacus/2022009

Available online at:
https://acta-acustica.edpsciences.org

Topical Issue - Auditory models: from binaural processing to multimodal cognition

SCIENTIFIC ARTICLE OPEN ACCESS

Using a blind EC mechanism for modelling the interaction between binaural and temporal speech processing

Saskia Röttges[1,*], Christopher F. Hauth[1], Jan Rennies[2], and Thomas Brand[1]

[1] Medizinische Physik and Cluster of Excellence Hearing4All, Carl-von-Ossietzky Universität Oldenburg, 26129 Oldenburg, Germany
[2] Fraunhofer Institute for Digital Media Technology IDMT and Cluster of Excellence Hearing4All, Marie-Curie-Straße 2, 26129 Oldenburg, Germany

Received 1 April 2021, Accepted 28 February 2022

Abstract – We reanalyzed a study that investigated binaural and temporal integration of speech reflections with different amplitudes, delays, and interaural phase differences. We used a blind binaural speech intelligibility model (bBSIM), applying an equalization-cancellation process for modeling binaural release from masking. bBSIM is blind, as it requires only the mixed binaural speech and noise signals and no auxiliary information about the listening conditions. bBSIM was combined with two non-blind back-ends: The speech intelligibility index (SII) and the speech transmission index (STI) resulting in hybrid-models. Furthermore, bBSIM was combined with the non-intrusive short-time objective intelligibility (NI-STOI) resulting in a fully blind model. The fully non-blind reference model used in the previous study achieved the best prediction accuracy ($R^2 = 0.91$ and RMSE = 1 dB). The fully blind model yielded a coefficient of determination ($R^2 = 0.87$) similar to that of the reference model but also the highest root mean square error of the models tested in this study (RMSE = 4.4 dB). By adjusting the binaural processing errors of bBSIM as done in the reference model, the RMSE could be decreased to 1.9 dB. Furthermore, in this study, the dynamic range of the SII had to be adjusted to predict the low SRTs of the speech material used.

Keywords: Speech intelligibility prediction, Temporal processing, Binaural processing, Auditory model

1 Introduction

Human listeners are able to understand speech in noisy backgrounds thanks to their ability to segregate the target speech signal from interfering signals [1]. This ability is more efficient when speech and interfering signals differ in their interaural level differences (ILD), interaural time differences (ITD), and/or interaural phase differences (IPD) [2]. Binaural unmasking (BU) and better ear listening (BEL) are two mechanisms applied by the human auditory system to benefit from spatially separated signals. These mechanisms rely on the distance between the ears and the head shadow effect. According to Lord Rayleigh's Duplex theory [3], BU is most effective for frequencies below 1500 Hz and BEL is most effective for frequencies above 1500 Hz. In this study, we investigate the BU mechanism for speech by analyzing the binaural release from masking (BRM), which is defined as the decrease in the speech reception threshold (SRT) due to BU. The SRT is defined as the signal-to-noise ratio (SNR) where 50% of the words of the test sentence are perceived correctly. One model predicting BRM in human listeners is the equalization-cancellation (EC) mechanism [4], which assumes that the internal representations of the left and right ears are temporally aligned, adjusted with respect to their amplitude, and subtracted from each other. This can lead to an improvement in SNR, because the noise is attenuated due to destructive interference. The EC mechanism is used as an effective model of binaural processing in speech intelligibility models [5–12].

When listening to speech in reverberant conditions, intelligibility is generally reduced [13–16] because temporal and binaural processing can be hampered due to decorrelation of the signals arriving at the left and right ears. However, different studies found that listeners can also benefit from early speech reflections (typically up to a delay time of 50–100 ms), because these reflections can be integrated with the direct sound of the target signal [16–20].

In order to investigate the complex interaction between binaural and temporal processing, a previous study [20] measured SRTs of reverberant speech in noise with different numbers of speech reflections at different delays and amplitudes. BU was evaluated by imposing an IPD of 0 or π on different components of the binaural room impulse responses (BRIRs), i.e., on the direct sound of the speech and/or the speech reflections and/or the noise. The SRT data were modelled in [20] using the binaural speech intelligibility model (BSIM) [7] after separating the BRIR into a useful part and a detrimental part. The target speech

*Corresponding author: saskia.roettges@uol.de

convolved with the useful part of the BRIR was integrated to the effective clean target signal, and the target speech convolved with the detrimental part was added to the interfering signals [15]. The best predictions were obtained by separating useful and detrimental parts using a temporal window with a length of about 200 ms. Speech reflections within this window were considered useful and reflections outside were considered detrimental. The optimal temporal position of this window was found by optimizing the intelligibility, as predicted by BSIM [7]. Interestingly, this "useful" window was not always an "early" window, as it did not require inclusion of the direct sound of the target speech; instead, it was more important that the window contained the maximum number of useful speech reflections. Later reflections can be beneficial for understanding speech in noisy and reverberant conditions, when the energy or IPD of the later speech reflections are more dominant cues for speech understanding. In such cases, the auditory system seems to focus on these late components rather than on the direct sound or previous components of the signal [20].

The model applied in [20] used auxiliary knowledge that is not explicitly available to listeners: First, the model knows the BRIRs for the target speech. Second, it knows the clean speech and the noise alone for optimizing the binaural parameters in the EC processing stage of the model. And third, the speech intelligibility index (SII) [21], which is used as the model's back-end for predicting the SRT, requires knowledge about the SNR (in other words, which signal components are target speech and which are not). Such auxiliary knowledge is also required by other similar models [15, 22, 23]. The necessity of this auxiliary knowledge, however, reduces the applicability of these models.

A modified version of BSIM (called BSIM20) was proposed in [8]. The front-end of BSIM20 employs a modified EC mechanism, which uses only the mixture of target speech and interferers as input, such that the EC processing and the selection of the better ear at high frequencies work blindly. Therefore, this blind front-end of BSIM20 is called bBSIM in the following (see Sect. 2.3 for details). bBSIM in combination with the SII [21] very successfully predicted SRT data measured in an anechoic situation at negative and positive SNRs [8]. Note that bBSIM works independently of the back-end, i.e., the back-end is not involved in the optimization of the EC parameters, such that bBSIM can be combined with arbitrary speech intelligibility back-ends.

Similar approaches to blind models were used in other studies to predict speech intelligibility in binaural conditions with noise [24, 25]. Cosentino et al. [24] used a localization model [26], which assumes that the target signal is located in front, to differentiate between the target signal and interferers. Because of this assumption, this approach is only applicable if the target signal is in front of the listener's head, which is not the case in this study. Another blind approach [25] combines a model of the auditory periphery [27], an EC mechanism, and a dynamic time warp (DTW) speech recognizer [28]. This approach is similar to the approach used in this study. However, it is only intended for negative SNRs, and it requires a transcription of the test sentences in order to decide whether the DTW algorithm has recognized them correctly. bBSIM works at arbitrary SNRs and in this study we combine it with a blind back-end that does not require a transcription of the test sentences.

In this study, we evaluate the capability of bBSIM [8] to predict the interactions between speech reflections and BU without using the auxiliary information about the BRIRs used in [20]. To achieve this, we combine bBSIM with three alternative back-ends, which use different signal parameters for predicting speech intelligibility:

For the first back-end, we use the SII [21], which uses the frequency-weighted SNR. However, in conditions with reverberation or reflections, the SII can only be expected to predict SRTs when the useful and the detrimental parts of BRIR are separated, as described in [20]. bBSIM does not perform this type of separation. Nevertheless, the SII reveals how bBSIM influences the SNR, which gives worthwhile insight into the model.

An alternative concept to the SII is the speech transmission index (STI) [29], which is more suitable for reverberant conditions, because it analyzes the modulation spectrum, which is influenced by both noise and reverberation. Therefore, for the second back-end, we use the STI in a speech-based version, which is based on the normalized covariance ($\mathrm{STI_{NCV}}$) between the clean target speech and the degraded speech [30]; note that the $\mathrm{STI_{NCV}}$ is very similar to the short-time objective intelligibility (STOI) measure [31].

Finally, for the third back-end we use the non-intrusive short-time objective intelligibility (NI-STOI) measure [32], which estimates the clean speech signal from the degraded signal and correlates the envelope of this estimate with the envelope of the degraded signal. Note that the combination of bBSIM and NI-STOI [32] can be regarded as completely blind, while the other model versions can be considered as hybrid models with a blind front-end and a non-blind back-end still requiring auxiliary information. It can be hypothesized that the fully blind model will outperform the hybrid models because the auxiliary information used by the hybrid models is not optimal, as the whole target speech signal with all reflections is assumed as useful. In contrast, the blind back-end has to extract the useful information of the mixed (target speech and interferers) output of bBSIM, which might be possible using an adequate back-end that fits to the front-end. Ideally, we are aiming for a binaural front-end that can be applied without further adaptation to arbitrary speech materials and that can be combined with different back-ends. Consequently, the main questions of this study are:

- Is the bBSIM able to predict the amount of BRM with an accuracy similar to the non-blind front-end used in [20]?
- Is the NI-STOI able to estimate the useful information based on the output of bBSIM well enough to predict SRTs in complex listening conditions with an accuracy similar to the non-blind baseline model used in [20]?

If the latter is the case, that would be very promising for further attempts of blind prediction of speech intelligibility in complex binaural listening conditions.

2 Method

2.1 Data basis

A data-set [20] from two groups of eight normal-hearing listeners aged between 18 and 31 years was used to predict SRTs in stationary noise. Twenty experiments were performed with a total of 86 conditions with different numbers of speech reflections at different delay times. The delay times of the reflections were varied from 0 to 200 ms in order to investigate how speech reflections are temporally integrated with the target speech. The number of reflections ranged from 0 to 9. The reflections of the target speech were realized by convolving the target speech with artificially created BRIRs. These BRIRs were generated by copying a BRIR with direct sound only and adding this copy to the original BRIR with delay times Δt of 10, 25, 50, 75, 100, 125, 150, 175, and/or 200 ms. With this procedure, all of the conditions were created, ranging from direct sound in noise to direct sound with up to nine reflections in noise. Additionally, the IPDs of direct sound, reflections, and/or noise were set either to 0 or to π in order to investigate the interaction of energetic and temporal processing with binaural processing. In two experiments, the level difference between the direct sound and a single reflection with a delay time of 200 ms was manipulated by multiplying the amplitude of the reflection by a factor of α, ranging from 0 to 2.5, in order to investigate the interaction between temporal integration and energetic effects. Note that the levels of speech (including all reflections) and noise at both ears were always equal, in order to ensure that there was no better ear. All experimental conditions used in the present study are shown in Table 1, where the direct sound is marked as D, the reflection(s) as R, and the noise as N. The IPD of these three components (D, R, N) is indicated by either 0 or π.

The speech material of the American English (AE) matrix sentence test (female talker) and a closed-set procedure were used [33]. All sentences had the same five-word structure: name–verb–numeral–adjective–objective, for example, *Allen ordered 19 white houses*. The sentences were syntactically correct but semantically unpredictable. The signals were presented via headphones (Sennheiser HD280 pro) in a sound-attenuated booth. The target speech was masked with a speech-shaped noise with the same long-term frequency spectrum as the speech material. The masker level was fixed at 65 dB SPL and SRTs were determined by adaptively varying the speech level based on the listeners' responses and by fitting a logistic model function to the data according to [34]. Note that the overall speech level was calculated after including all speech reflections. Before starting the actual measurements, each listener was allowed to get familiar with the sentences by carrying out two SRT measurements using lists of 20 sentences.

Table 1. Experiment conditions of the data-set [20]: This table shows the IPD of the direct sound (D), the reflection(s) (R), and the noise (N) for each experiment. Note that in experiments VI and VII, the amplification factor (α) for the reflection was varied between 0 and 2.5. For all other conditions α was set to one.

Exp no.	D-IPD	R-IPD	N-IPD
Varying a single R:			
Exp. I	0	0	0
Exp. II	0	π	0
Exp. III	π	0	0
Exp. IV	0	π	π
Exp. V	π	0	π
Varying α:			
Exp. VI	0	0	0
Exp. VII	0	π	0
Increasing number of R from early R:			
Exp. VIII	0	0	0
Exp. IX	0	π	0
Exp. X	π	0	0
Exp. XI	π	π	0
Exp. XII	0	0	π
Increasing number of R from late R:			
Exp. XIII	0	0	0
Exp. XIV	0	π	0
Exp. XV	π	0	0
Exp. XVI	π	π	0

2.2 Non-blind baseline model

The SRTs of this data-set were predicted in [20] using the non-blind BSIM [7, 15] with a modification of the input signals. This model served as a baseline model in the present study and consists of three processing stages: In the first stage, the input signals are modified by separating the BRIR into useful and detrimental parts. In the second stage, the non-blind EC processing is applied to predict BRM, and in the third stage the SRT is predicted using the SII [21]. In the following, these stages are described in more detail.

2.2.1 Useful/detrimental separation

A movable temporal window is used to integrate the useful part of the BRIR and to separate it from the detrimental part. The target speech convolved with the part of the BRIR falling inside this integration window is processed in the following stages of the model as the effective target signal. The target speech convolved with the part of the BRIR falling outside the integration window is processed like masking noise in the following stages of the model. The integration window has linear, symmetric ramps in a triangular shape with a ramp duration of 100 ms. Note that the integration window does not require inclusion of the direct sound, as the window is moved along the BRIR until the maximum target speech energy is integrated. This resulted in better predictions of the data in [20] than an early/late separation (where the useful part always included the direct target sound) in conditions where the later part provided more speech energy or better SRM compared to the earlier part.

2.2.2 Non-blind EC processing

The EC processing as proposed in [7] receives the left and right ear signals of target speech and interferers separately. To simulate the frequency selectivity of the auditory system, a gammatone filterbank [35] is used, which splits the input signals into 30 equivalent rectangular bandwidth-(ERB-) spaced [36] frequency bands between 150 and 8500 Hz. In each filter, the noise is equalized between the left and right ear channels by applying an interaural delay and an interaural gain so that the subsequent cancellation (that is, the subtraction of left and right ear signals) leads to maximum improvement of the SNR. Note that the equalization process is limited by jittering the interaural delay and gain according to [37].

2.2.3 SII

The SII [21] is used to predict the SRT based on the SNR calculated by the EC processing described above. The SII analyzes the SNR in each frequency channel. Note that the dynamic range of the SII is limited to −15 to 15 dB in each frequency band; bands with an SNR below −15 dB do not contribute, and the maximum SNR in each band is limited to 15 dB. These SNRs are normalized and weighted according to their importance for human speech perception and integrated across the 30 gammatone bands. This results in an index value between 0 (representing no intelligibility) and 1 (representing maximum intelligibility). However, this index does not represent speech intelligibility directly, but has to be mapped to empirical data (see Sect. 2.5). The SII is a non-blind measure, as it requires separate input signals for clean target speech and interferers.

2.2.4 Reference SII and fitting of binaural processing errors

In order to calculate SRTs based on SII predictions, a reference SII value is required. This reference SII was determined as the SII value at the observed SRT of the first condition of experiment I (D_0N_0, no reflection). This was determined by varying the SII value until the predicted SRT matched the mean measured SRT for this condition. For all other conditions, the SRT prediction was determined by varying the input SNR until the SII matched the reference SII. Rennies et al. [20] used the reference SII values 0.098 and 0.083 for the two groups of listeners. Note that these reference SII values are very low compared to other studies [6, 7, 15, 23] because the speech material used for the measurements was very intelligible, and thus had a very low SRT.

Furthermore, [20] increased the ITD processing errors in the EC stage by a factor of 1.6 and 1.8 for the two groups of listeners in order to fit SRT predictions to the data. These factors were determined by varying them until the predicted SRTs for the first condition of experiment III ($D_\pi N_0$) matched the mean measured SRT for this condition. Without this adjustment, the binaural benefit was overestimated by 2 dB and 2.5 dB for the two listener groups.

2.3 Blind front-end of BSIM (bBSIM)

The binaural processing used in [20] requires auxiliary knowledge in order to separate the BRIR into useful and detrimental parts, and requires auxiliary knowledge about the clean target and interferer signals for the subsequent calculation of the SII.

BSIM20 [8] uses a blind front-end, which is called bBSIM in this article. bBSIM performs blind EC processing below 1500 Hz and blindly selects the better ear above 1500 Hz and does not require any auxiliary knowledge. Instead, bBSIM receives the mixed speech and noise signals for the left and right ears. In each gammatone filterbank channel, the following processing takes place: The two ear channels are equalized in level and phase. Subsequently, in the blind cancellation two concurrent strategies are applied: First, the two ear channels are subtracted from each other, which minimizes the output of the EC stage and is the best strategy for negative SNRs. Second, the two ear channels are added to each other, which maximizes the output of the EC stage and is the best strategy for positive SNRs. Subsequently, the strategy (subtracting or adding ear channels) that produces the higher speech-to-reverberant modulation energy ratio (SRMR) [38] is selected. The SRMR describes how speech-like a signal is by calculating the ratio between the energy in the modulation frequency channels below 16 Hz (speech-like) and the energy in the modulation frequency channels above 16 Hz (not speech-like). bBSIM models BEL above 1500 Hz by selecting the ear channel leading to the higher SRMR. In bBSIM, the uncertainties of human binaural processing are realized by jittering the delays and gain factors that are applied in the EC stage. These jitters are realized by using delay and gain factors that slightly deviate from their optimum values found by minimizing (or maximizing) the EC output (see above). The deviations are drawn from normal distributions with standard deviations according to vom Hövel [37] using Monte-Carlo simulations (MCSs).

For each SRT, 100 MCSs are performed by using 10 MCSs for each sentence for 10 sentences total. Subsequently, the resulting 10 back-end values of each sentence are averaged to get the output of bBSIM. This output is calculated in 1 dB steps for an SNR range of −30 to −5 dB. This range always includes the measured SRT. From this range of SNRs, the different back-ends (see Sect. 2.4) select the SRT, as described in Section 2.5.

2.4 Back-ends

For non-blind back-ends that require auxiliary knowledge about clean target and/or interferer signals, bBSIM can also calculate the enhanced target and interferer signals separately. Note that the EC parameters and the better ear selection are still calculated blindly, and that clean speech and interferer signals are processed in the same way as in the mixture. This is possible because the processing of the signals in bBSIM is linear. In this study, the SII [21] and the $\mathrm{STI_{NCV}}$ [30] are used as non-blind back-ends. Blind

back-ends, which require only the mixture of target and interferer signals, can use the enhanced front-end output directly. In this study, NI-STOI [32] is used as a blind back-end. In the following, the three back-ends evaluated in combination with bBSIM are described. Note that in any case, bBSIM is purely signal-driven and not influenced by the final back-end. Instead, it is controlled by the SRMR, which acts as a preliminary back-end that is applied independently in each frequency channel.

2.4.1 SII

The SII [21] is applied to the output of bBSIM as described in Section 2.2.3.

2.4.2 STI_{NCV}

Since bBSIM does not perform a useful/detrimental separation, as done in the baseline model, the SNR (which is the basis of the SII) is not well suited as a predictor of intelligibility. The STI [29] analyzes the modulation transfer function, which is affected not only by the SNR but also by reverberation and speech reflections; we expected it to be better suited than the SII for the output of bBSIM. The STI compares the envelopes of the separated input signals (for example, clean speech and degraded speech) to determine the modulation transmission index (TI) for each frequency band. The final STI results from the weighted average of these TI values across frequency bands [29].

This study uses the speech-based $\mathrm{STI_{NCV}}$ [30]. The standard STI and the $\mathrm{STI_{NCV}}$ differ in the calculations of the envelope signals and the calculations of the TI values. In the $\mathrm{STI_{NCV}}$ method, the covariance between the clean speech and the degraded speech envelopes is calculated in each frequency band and normalized by the individual variances in the clean and degraded speech. From this normalized covariance, the SNR is calculated in each band. The subsequent calculations of an index value are similar to the SII. Like the SII, the STI is an index from 0 to 1, with 1 representing maximum intelligibility. The mapping from STI to intelligibility is described in Section 2.5.

2.4.3 NI-STOI

In contrast to SII and STI, the NI-STOI [32] does not require separate target speech and interferers, but receives only the mixture of them. The NI-STOI predicts intelligibility based on the correlation between the envelopes of clean and degraded target speech calculated in 1/3-octave frequency bands, which is very similar to the $\mathrm{STI_{NCV}}$. However, for estimating the clean speech envelopes from the degraded speech envelopes, NI-STOI uses a statistical model, which requires training with clean speech. Training with particular interferers is not necessary. The software for calculating NI-STOI was taken from [39]. This software provided a trained model, that is, NI-STOI was not trained to the speech material of this study but to clean Danish matrix sentences [40] and to female and male talkers from the TIMIT database [41]. Unlike [32], NI-STOI was calculated here without using the voice activity detection (VAD), as in this study the target speech was present during the whole input signal.

2.5 SRT prediction

In order to predict the SRT based on an index value (SII, $\mathrm{STI_{NCV}}$, or NI-STOI) a reference index that corresponds to the SRT is required for each index. In this study, this reference is determined as the index value corresponding to the SRT in the D_0N_0 condition (no phase inversion and no reflection). This reference condition was also used by [20] for the baseline model. However, [20] overestimated the BRM, this was compensated for by increasing the BSIM binaural processing errors (see Sect. 2.2.4). In the present study, bBSIM also led to an overestimation of the BRM. As in [20], we used the $D_\pi N_0$ condition (no reflection) of experiment III as the representative condition for estimating the BRM. In this condition the overestimation was 3 dB for the SII, 2 dB for the $\mathrm{STI_{NCV}}$, and 5.5 dB for the NI-STOI (see first condition in Fig. S2 in the Supplemental Material for SII and Fig. 3 for $\mathrm{STI_{NCV}}$ and NI-STOI).

2.5.1 Increasing bBSIM's binaural processing errors

One possibility to compensate for overestimates of BRM is to increase bBSIM's binaural processing errors, as done in [20]. We increased the processing errors by a factor of 1.6 for the $\mathrm{STI_{NCV}}$ and by a factor of 2.1 for the NI-STOI (see results in Fig. S4 in the Supplemental Material and Sect. 4.1) in order to compensate for the underestimation of the SRT of the $D_\pi N_0$ condition (no reflection) of experiment III. However, increasing bBSIM's binaural processing errors is inconsistent with other studies where the original binaural processing errors led to accurate predictions for bBSIM [8, 42, 43]. Furthermore, adapting the binaural processing inaccuracies to the specific listening conditions of different studies contradicts the idea of a blind model front-end that does not require any auxiliary information about the listening condition. For these reasons, we evaluated a different method to compensate for the overestimation of BRM that leaves bBSIM unchanged, as described in the following.

2.5.2 Adjusting the reference SRT

An explanation for the different results shown in previous studies ([20] vs. [8, 42, 43]) may be found in the different SRT values of the matrix sentence tests used in these studies [33]: The reference SRT of the German matrix sentence test is −8 dB [8, 42, 43] is, while the reference SRT of the AE matrix sentence test is −12 dB [20]. This low SRT value of −12 dB is very close to the lower limit of the dynamic range that is considered by the SII (−15 dB). As a consequence, the number of frequency bands that are excluded from the SII calculation is larger for the AE matrix test than for the German matrix test, as the SNR in these bands falls below −15 dB. This may cause inaccurate SII

predictions, because the weighted sum is calculated across an insufficient number of frequency bands. In order to solve this problem, we increased the reference value of the SII so that the calculation takes place within the dynamic range of the SII and the measured BRM is predicted accurately. This was achieved as follows:

At first, the relation between the reference SII and the predicted BRM was analyzed: Figure 1 shows bBSIM-SII predictions of BRM (defined as the difference between the SRT in the D_0N_0 condition without delay and the $D_\pi N_0$ condition without delay) for various reference values (Fig. S1 in the Supplemental Material shows the same relation for all back-ends). Then the reference index was determined at which the measured BRM was best predicted. In [20] the measured BRM for this situation was 7.8 dB, leading here to an SII value of 0.28. Using this optimal reference index, the model is perfectly calibrated for predicting the BRM for this condition, even without adjusting bBSIM's binaural processing errors.

The difference between SRT predictions using the original and optimal references is shown in Figure 2. The black dashed line shows the SII for SNRs from −30 to 5 dB for the D_0N_0 condition (experiment I) and the black solid line shows the SII for the $D_\pi N_0$ condition (experiment III). The red line indicates the measured SRT of the $D_\pi N_0$ experimental condition. The reference SRT of −12 dB in experiment I (D_0N_0, no reflection) corresponds to the original SII reference of 0.11 (gray line). Using this reference, the SRT prediction for the first condition of experiment III ($D_\pi N_0$, no reflection) is −23 dB (gray circle) and thus the measured SRT (red line) is underestimated by 3 dB and consequently the BRM is overestimated by 3 dB. To avoid this overestimation, the transformation from SII to SRT is done using the optimal SII reference of 0.28, determined above (see Fig. 1). This optimal SII reference corresponds to an SRT of −6.4 dB in experiment I (yellow line). In other words, by using the optimal reference, we shifted the reference SRT for the SII calculation by 5.6 dB from −12 dB to −6.4 dB. For the first condition of experiment III, the optimal SII reference corresponds to an SRT of −14.2 dB (yellow square) and the BRM (−6.4 − (−14.2) = 7.8 dB) is predicted exactly. In order to compensate the shift of the reference SRT introduced above, we now subtract 5.6 dB from the predicted SRT giving the final SRT prediction of −19.8 dB.

The situation is different for the $\mathrm{STI_{NCV}}$, even though its dynamic range is also limited to a minimum value of −15 dB. However, the original $\mathrm{STI_{NCV}}$ reference is not in a flat area of the input–output function (see Fig. S3, middle panel in the Supplemental Material) and a shift would not move the reference $\mathrm{STI_{NCV}}$ towards a steeper area. In other words, the low SRT of the AE matrix test seems not to be the underlying reason for the overprediction of the SRT. Furthermore, a shift of the reference SRT, as done for the SII would cause an increase in the reference $\mathrm{STI_{NCV}}$ value from 0.1 to 0.6 (corresponding to a shift of the reference SRT by 11.4 dB) which we regard as kind of appropriate because such a large shift would move the reference SRT beyond any value observed in the matrix sentence tests of various languages and talkers [33]. So we decided not to perform this shift.

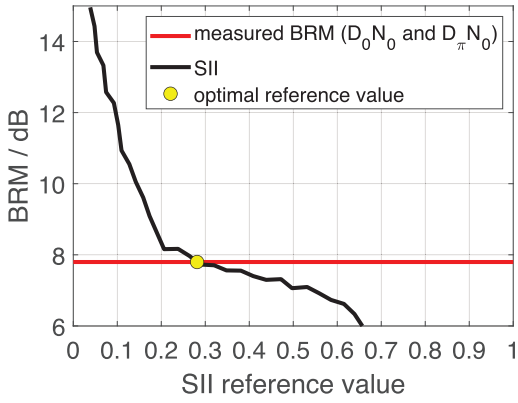

Figure 1. Predicted binaural release from masking (BRM, shown on the y-axis) defined as the difference between SRTs of the first condition of experiment I (D_0N_0) and the first condition of experiment III ($D_\pi N_0$) using bBSIM-SII. The SII reference values are shown on the x-axis. The red line shows the measured BRM of 7.8 dB. The yellow circle shows the optimal reference SII of 0.28, which allows accurate prediction of the measured BRM and was selected for the following predictions.

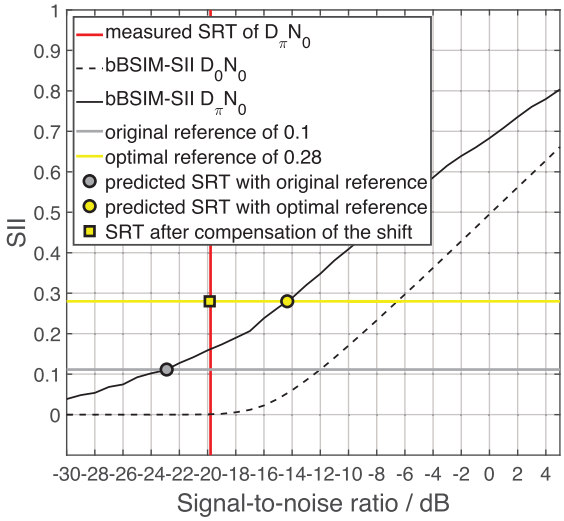

Figure 2. Illustration of the calibration of the reference SRT for accurate prediction of the BRM and example how to estimate SRT in the $D_\pi N_0$ condition. See Section 2.5 for details.

The situation is similar for the NI-STOI as its input-output function shows at no point a steeper slope than at the original reference. This leads to an overestimation of the BRM using any shift of the reference value (see Fig. S3 in the Supplemental Material). As a result of this analysis, we only adjusted the SRT reference of the SII for calculating the results presented in this study. However, we evaluate and discuss the effect of increasing bBSIM's binaural processing errors and adjusting SRT references for all back-ends in Section 4.1.

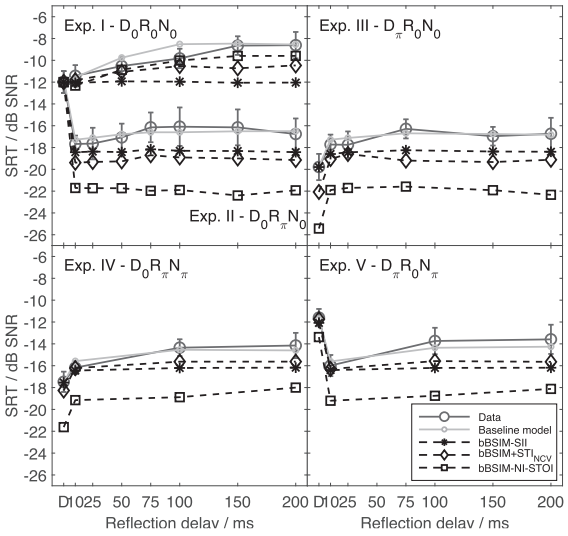

Figure 3. Measured (big, dark gray circles) and predicted SRTs for experiments I to V with the baseline model (small, light gray circles), bBSIM-SII (black stars connected with black dashed line), bBSIM-STI_{NCV} (black diamonds connected with dashed lines), and bBSIM-NI-STOI (black squares connected with dashed lines). This figure shows the influence of a single reflection of the target speech at different delays. Note that the first condition is always direct sound only.

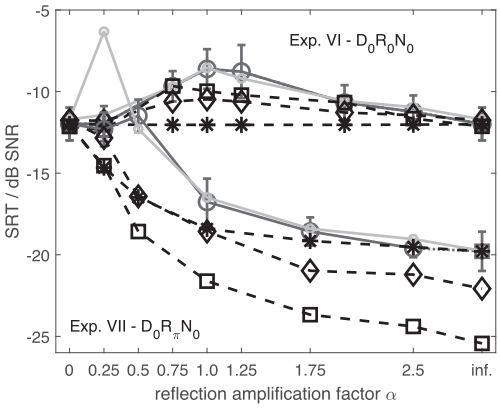

Figure 4. Measured and predicted SRTs for experiments VI and VII in the same display as in Figure 3. This figure shows the influence of a single reflection of the target speech with a delay time of 200 ms, which was varied in its amplitude from 0 to 2.5.

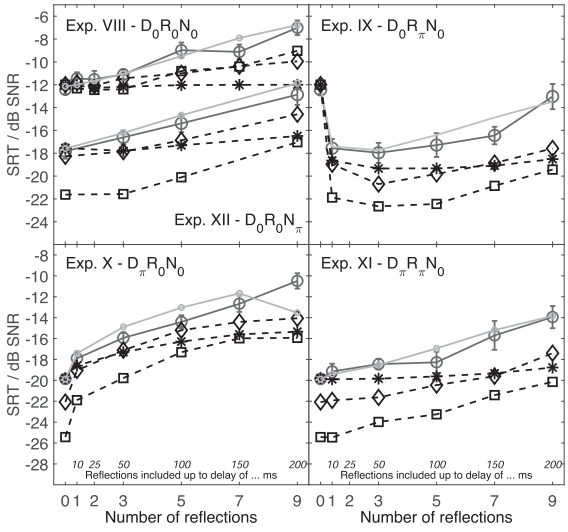

Figure 5. Measured and predicted SRTs for experiments VIII to XII displayed as in Figure 3. This figure shows the influence of multiple reflections of the target speech, which were varied by increasing the number of speech reflection at different delays starting from early delays. IPDs were set as indicated. Note that the first condition is always direct sound only.

3 Results

Figures 3–6 show measured and predicted SRTs for the different experiments. The dark gray circles show the measured SRTs [20] and are connected with gray solid lines. The error bars represent the standard error. The lighter gray circles show the predicted SRTs of the baseline model [20] and are connected with lighter gray solid lines. The black symbols show the predictions of bBSIM and the three different back-ends.

3.1 Delay time for a single reflection

Figure 3 shows experiments I to V, which evaluated the temporal integration of a single reflection with varying delay time. In experiments II to V, an IPD of π was imposed on the direct sound (D), the reflection (R), and/or the noise (N), as indicated in Table 1. Note that in all experiments, the first condition was direct sound only.

In experiment I ($D_0R_0N_0$, top left panel, upper graphs), the measured SRTs increased from -12.0 to -8.6 dB with increasing delay time of the reflection. This increase was predicted by bBSIM-NI-STOI and the baseline model only. The hybrid models did not clearly show the increasing tendency of the measured SRTs.

In experiment II ($D_0R_\pi N_0$, top left panel, lower graphs), the IPD of the reflection was set to π, which enabled BU, making the reflection the dominant cue. Consequently, the measured SRTs showed an improvement compared to experiment I. Furthermore, SRTs did not increase with increasing delay time, but were almost constant. These two main effects were predicted by all back-ends. However, BRM was overestimated by the bBSIM-STI_{NCV} by 1.5 dB and the bBSIM-NI-STOI by 4 dB. This overestimation of BRM occurred in all experiments. In the following, the results will be described without indicating this finding for each experiment, but it will be discussed in Section 4.1.

In experiments III ($D_\pi R_0 N_0$, top right panel) and IV ($D_0R_\pi N_\pi$, bottom left panel), noise and reflection had the same IPD. In these experiments, the direct sound enabled BU, resulting in lower SRTs compared to the D condition of experiments I and II. All models predicted this effect. In experiment IV, all models predicted nearly constant SRTs over all delay times.

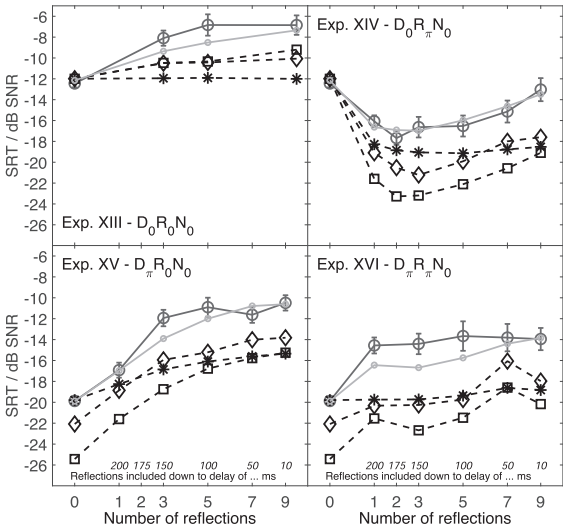

Figure 6. Measured and predicted SRTs for experiments XIII to XVI displayed as in Figure 3. This figure shows the influence of multiple reflections of the target speech, which were varied in their increasing number of reflections at different delays, starting from late reflections. IPDs were set as indicated. Note that the first condition is always direct sound only.

In experiment V ($D_\pi R_0 N_\pi$, bottom right panel) the IPDs of the target speech and noise were set to π. This makes experiment V comparable to experiment II because in both experiments, the reflection had an IPD different from the direct sound and the noise. As in experiment II, the SRTs decreased when the reflection was introduced, but to a lesser extent than in experiment II. All models predicted this smaller effect, with no overestimation for the SII and the STI_{NCV} and less overestimation (compared to experiment II) for the NI-STOI. Only the baseline model and the NI-STOI followed the increase in SRTs with increasing delay time, whereas the predictions of the other models were independent of delay time.

3.2 Amplitude of late single reflection

Figure 4 shows experiments VI (upper graphs) and VII (lower graphs), where the amplitude of a single reflection with a delay time of 200 ms was varied by a factor α ranging from 0.0 to 2.5. The factor inf. at the x-axis defines a condition consisting of reflection only, and should serve as a visual guide. The SRTs of the inf. conditions were copied from other conditions: In experiment VI, the SRTs of the inf. condition were copied from the SRTs for α equals 0, as in both cases there is only one speech signal (the reflection acts like the direct sound). In experiment VII, the SRTs of the inf. condition were taken from the D condition of experiment III.

In experiment VI ($D_0 R_0 N_0$, upper graphs) direct sound, reflection, and noise had an IPD of zero. The measured SRTs increased with increasing amplitude of the reflection until α equals 1; this was caused by the detrimental effect of the late reflection. For α higher than 1, the SRTs decreased again, showing that the auditory system is able to switch attention to the late reflection and that it is not restricted to including the direct sound of the target speech. The baseline model, the bBSIM-STI_{NCV}, and the bBSIM-NI-STOI predicted the measured increase of SRTs for α close to one. The bBSIM-SII was not able to predict this effect and predicted nearly constant SRTs over all reflection amplification factors.

In experiment VII ($D_0 R_\pi N_0$, lower graphs), the IPD of the reflection was set to π. This enabled BU for the reflection, making it a potentially more dominant cue even if its amplitude is lower than the amplitude of the direct sound. The related decrease in SRTs was predicted by all models. For α larger than 1, bBSIM-SII did not underestimate the measured SRTs, whereas STI_{NCV} and NI-STOI showed underestimations similar to those in experiments II to V. For α equals 0.5, all models showed the largest underestimation of SRT. The baseline model predicted a peak at α equals 0.25, where the predicted SRT increased by about 6 dB. This prediction pattern may be explained by two contradicting effects: The IPD of the reflection enabled BU, making the reflection the dominant cue while, on the other hand, the reflection's amplitude was lower than that of the direct sound making the direct sound the dominant cue. The current models appear to use the direct sound as the more dominant cue, until the amplitudes are equal (α equals one), because they underestimated the SRTs more when α equals 0.5 than when alpha is larger than 1.

3.3 Multiple reflections with increasing delay time

Figure 5 shows the results of experiments VIII to XII with multiple reflections of the target speech. Reflections start at 10 ms and successively more reflections are added, up to a total of nine reflections with the last one at 200 ms. In each experiment, all reflections had identical IPD. In this way, the integration of multiple simultaneous reflections was evaluated. In experiment VIII ($D_0 R_0 N_0$, top left panel, upper graphs), direct sound, reflection(s), and noise had zero IPD. Compared to experiment I, the SRTs increased more with increasing number of reflections, due to the stronger masking by multiple reflections compared to a single reflection. The baseline model, bBSIM-NI-STOI, and bBSIM-STI_{NCV} predicted this increase. bBSIM-SII predicted nearly constant SRTs.

In experiments IX to XII, in addition to the increase in the number of reflections, the IPD was set to π, as indicated in Table 1. In experiment IX ($D_0 R_\pi N_0$, top right panel), the IPD of the reflections was set to π, as in experiment II. The measured SRTs were similar to those of experiment II up until seven reflections. For more than seven reflections, SRTs increased more rapidly with increasing number compared to experiment II, because masking was stronger with multiple reflections than with a single reflection. All models predicted this behavior correctly; however, bBSIM-NI-STOI strongly overestimated BRM.

In experiment X ($D_\pi R_0 N_0$, bottom left panel), the IPD of the direct sound was set to π, as in experiment III. SRTs increased with increasing delay time. All models could

predict this effect for up to five reflections. With seven and nine reflections, the predictions of bBSIM with all back-ends showed a flatter trend compared to the measured SRTs. For the baseline model, the predicted SRT decreased only at nine reflections, resulting in an underestimation of the SRT.

In experiments XI ($D_\pi R_\pi N_0$, bottom right panel) and XII ($D_0 R_0 N_\pi$, top left panel, lower graphs), the noise had a different IPD than the direct sound and its reflections, which enabled BU. In both experiments, SRTs increased with increasing number of reflections due to the increasingly detrimental effect of the later reflections. The overall pattern of the results was very similar to that of experiment VIII, except for a downward shift in SRTs by 7–9 dB. bBSIM-SII predicted nearly constant SRTs for all numbers of reflections in both experiments XI and XII. The baseline model, bBSIM-NI-STOI and the $\mathrm{STI_{NCV}}$ predicted the increase in SRTs quite well.

3.4 Multiple reflections with decreasing delay time

Figure 6 shows the results of experiments XIII to XVI with different numbers of speech reflections at different delays, starting with late reflections.

Experiment XIII ($D_0 R_0 N_0$, top left panel) showed an SRT pattern similar to experiment I: SRTs increased with increasing number of reflections even though the delays approached the direct sound. The bBSIM with the three back-ends showed trends similar to experiment I. bBSIM-SII did not clearly show the increasing tendency of the measured SRTs. bBSIM-NI-STOI and bBSIM-$\mathrm{STI_{NCV}}$ showed a small increase in the SRTs, which was biased towards too low SRTs. The increase was only predicted by the baseline model, with a slight underestimation of SRTs.

In experiment XIV ($D_0 R_\pi N_0$, top right panel), SRTs decreased by about 4.5 dB when a single or several late reflection(s) were added. Adding further reflections (i.e., increasing the period over which the delays were distributed) caused slightly increasing SRTs, which almost reached the SRT of the D condition. bBSIM-$\mathrm{STI_{NCV}}$ and bBSIM-NI-STOI predicted this trend, while bBSIM-SII showed a flatter trend and predicted nearly constant SRTs for an increasing number of reflections.

In experiment XV ($D_\pi R_0 N_0$, bottom left panel), only the IPD of the direct sound was set to π. SRTs first increased when reflections were added, and then stayed nearly constant for 5–9 reflections. The hybrid models predicted these nearly constant SRTs for 3–9 reflections, while bBSIM-NI-STOI showed a further increase with more reflections. All models predicted increasing SRTs from 0 to 5 reflections. In this experiment, the baseline model slightly underestimated SRTs for three and five reflections.

In experiment XVI ($D_\pi R_\pi N_0$, bottom right panel), the IPDs of direct sound and reflections were set to π. The results showed a trend similar to that in experiment XIII, except for a shift of 7–8 dB towards lower SRTs. Adding one reflection increased SRTs; adding more reflections did not influence SRTs any further. bBSIM-NI-STOI and $\mathrm{STI_{NCV}}$ predicted an increase in SRTs when adding one late reflection, but showed a peak at seven reflections. bBSIM-SII predicted nearly constant SRTs for all numbers of reflections and did not predict the increase when a single reflection was added. In this experiment, the baseline model slightly underestimated the measured SRTs for 1–5 reflections.

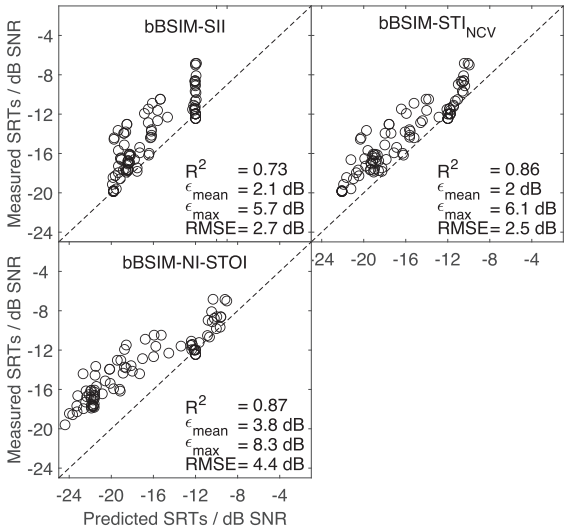

Figure 7. Overall prediction accuracy of bBSIM with three back-ends, with coefficient of determination (R^2), mean absolute prediction error (ϵ_{mean}), maximum absolute prediction error (ϵ_{max}), and root-mean-square error (RMSE) for all 16 experiments.

3.5 Overall comparison between back-ends

Figure 7 shows scatter plots of measured against predicted SRTs for all conditions in the 16 experiments. Each panel shows the predictions of bBSIM with one specific back-end together with the coefficient of determination (R^2), the root-mean-square error (RMSE), the highest (ϵ_{max}), the mean absolute (ϵ_{mean}) prediction error and identity as dashed lines. bBSIM-SII yielded an R^2 value of 0.73 with an RMSE of 2.7 dB; bBSIM-$\mathrm{STI_{NCV}}$ yielded an R^2 value of 0.86 with an RMSE of 2.5 dB, and bBSIM-NI-STOI yielded an R^2 value of 0.87 with an RMSE of 4.4 dB.

bBSIM-NI-STOI (bottom right panel) and bBSIM-$\mathrm{STI_{NCV}}$ (top right panel) yielded the highest R^2, whereas bBSIM-$\mathrm{STI_{NCV}}$ yielded a lower RMSE than bBSIM-NI-STOI. For SRTs higher than −12 dB SNR (which corresponds to the SRT for $D_0 R_0 N_0$), bBSIM-NI-STOI made very accurate predictions. Below −12 dB, SRTs were underestimated. Note that all conditions with an SRT below −12 dB include effects of BU. bBSIM-SII predicted constant SRTs over all reflection delays in experiments I–VI, VIII, IX, XI–XIII; this, is reflected as vertical patterns in the scatter plot.

The baseline model [20] showed the highest prediction accuracy, with an R^2 value of 0.91 and an RMSE of 1 dB. However, R^2 was only slightly higher than R^2 of bBSIM-NI-STOI. This is remarkable given the fact that the baseline model received auxiliary information about the BRIR, the target speech, and the noise and that furthermore the reference SII values and the binaural processing errors were adapted independently for the two groups of listeners and the two groups of experiments [20] (see Sect. 2.2.4). In contrast, bBSIM-NI-STOI does not use any auxiliary information, as it works blindly. NI-STOI requires only a reference value that corresponds to the NI-STOI value that belongs to the measured SRT of the reference condition D_0N_0 (no reflection) and which has been used for all conditions. Section 4.1 evaluates and discusses how the $\mathrm{STI_{NCV}}$ and the bBSIM-NI-STOI predictions can be improved by increasing bBSIM's binaural processing errors.

4 Discussion

The first main question of this study was, is bBSIM able to predict the BRM with an accuracy similar to the non-blind front-end used by Rennies et al. [20]? This can only be answered to a certain degree. Quite a number of the predicted SRTs were in good agreement with the observed data. But there was a general overestimation of the BRM, which depended on the back-end. (The overestimation was 3 dB for the SII, 2 dB for the $\mathrm{STI_{NCV}}$, and 5.5 dB for the NI-STOI for the $D_\pi N_0$ condition). This suggests that the overestimation of the BRM was not caused by the front-end alone, but rather by the combination of front-end and back-end. As hypothesized, the fully blind model ($R^2 = 0.87$) outperforms the hybrid model bBSIM-SII ($R^2 = 0.73$) at least with respect to R^2 (but not with respect to RMSE). bBSIM-$\mathrm{STI_{NCV}}$ ($R^2 = 0.86$) shows nearly the same R^2 as bBSIM-NI-STOI.

The second main question of this study was, is the NI-STOI able to estimate the useful information based on the output of bBSIM well enough to predict SRTs in complex listening conditions (with a single stationary interferer and different delays of the target speech) with an accuracy similar to the non-blind baseline model used by Rennies et al. [20]? This can be answered with yes with regard to the correlation between predictions and observations, as the NI-STOI achieved an R^2 similar to the baseline model ($R^2 = 0.91$). This is remarkable because the baseline model uses auxiliary information about the BRIR and about speech and noise in the EC process and in the SII. However, the overestimation of BRM biased all predictions that involve interaural phase inversions, which increased the RMSE from 1 dB for the baseline model to 4.4 dB for bBSIM-NI-STOI. Note that neither the binaural processing errors of the bBSIM nor the reference values of the NI-STOI were adjusted and that the overestimation of the BRM also occurred in the baseline model before the adjustment of the binaural processing errors. This overestimation led to the high RMSE of the NI-STOI.

4.1 Explanations for overestimating BRM and possible solutions

In order to compensate for the overestimation of BRM, Rennies et al. [20] increased the binaural processing errors of BSIM's front-end (see Sect. 2.2.4). This was motivated by the assumption that these overestimates were caused by the EC stage, which works more precisely than the binaural processing of the listeners. As the overestimates differed between the back-ends (see Fig. S1), we concluded that they did not originate from bBSIM alone, but rather from the combination of bBSIM and the back-end. One possible reason is the very low SRT of the AE matrix test [20], which is −12 dB for the D_0N_0 (no reflection) condition, which has no BRM. In other studies, however, using BSIM [6, 7] or bBSIM [8, 42, 43] with the German matrix test, we did not find an overestimation of BRM using the SII. Note that the German matrix test has an SRT of −7.1 dB. Such differences are not unusual. Kollmeier et al. [33] showed that SRTs differ significantly between different matrix tests, which can be attributed to differences between languages as well as between the talkers' articulation [14].

The SII reference of the AE matrix test is very low at 0.1, compared to values around 0.2 for the German Matrix test. It should be kept in mind that the SNR at the SRT of the speech material interacts with the calculation of the SII, as the dynamic range which is considered for the calculation of the SII is limited to −15 to +15 dB in each frequency band. Thus, for a low SRT of −12 dB, the SNR falls below −15 dB in many frequency bands, and these bands do not contribute to the SII. If EC processing introduces BRM, the SNR will exceed −15 dB in some of these frequency bands, and these bands will contribute to the SII. Consequently, the low reference SII will suddenly be exceeded, and a low SRT will be predicted. But if the SII reference is higher, as for the German matrix sentences, the SNR is above the −15 dB threshold in more frequency bands, and fewer bands kick in suddenly due to BRM. Overall, more frequency bands are required to exceed the higher reference index. This makes the BRM estimate more stable, as it is averaged across more frequency bands. This effect can be seen in the sudden increase in predicted BRM in Figure 1 for SII values below 0.2.

This very steep behavior of the BRM for low SII reference values is caused by the flat areas of the index input-output function (see Fig. 2 for SII only and Fig. S3 for all back-ends in the Supplemental Material).[1] While determining the SNRs to match a low SII reference value during the SRT prediction process (see Sect. 2.5), this led to an underestimation of the measured SRTs.

Furthermore, the MCSs introduce an uncertainty that, in combination with the flat curve, makes the SRT estimate very imprecise. The uncertainty of the SRT estimate is given by the uncertainty of the SI measure multiplied by the inverse of the slope of the SI curve. For that reason,

[1] Note that the back-end outputs, are integrated across frequencies and that the EC processing achieves BU only in a few frequency bands. This results in even shallower output functions (see Fig. S3, Supplemental Material).

measures with steeper curves produce more precise SRT estimates. In order to compensate for these effects, we adjusted the reference value of the SII (see Fig. 1), so that it is more central in the dynamic range (−15 to 15 dB) of the SII. Alternatively, the dynamic range of the SII could have been changed to, e.g., −20 to 10 dB, but we did not want to change the standardized SII [21]. For the SII, this method worked successfully because, with the optimal reference, the SII worked on a steeper area of its input–output function, eliminating the strong overestimation of BRM (see Fig. 1).

The reference SRT of the STI_{NCV} was not on the flat part of the input-output function (see Fig. S3 in the Supplemental Material), even when the lower limit of the STI_{NCV} was at −15 dB. One reason for this could be that the STI_{NCV} uses more input signal information than just the SNR, such as modulations. In addition, the STI splits the input signal into octave bands with center frequencies from 125 Hz to 8000 Hz, which includes only seven bands, which are wider than the 30 bands used by the SII. This may lead to more stable BRM predictions over different STI reference values (see Fig. S1). A shift of the original reference value of the STI_{NCV} to the optimal reference value corresponds to a shift of the reference from 0.1 to 0.6, which corresponds to a shift of the SRT reference by 11.4 dB. This is a rather extreme shift which brings only a small benefit and therefore we decided not to apply it in this study.

For the NI-STOI, however, the problem of overestimated BRM remained even when we shifted its reference value: The reference value of NI-STOI cannot be moved sufficiently far away from the flat area of the input-output curve because there is no steeper area in the whole input–output function of the NI-STOI (see Fig. S1 in the Supplemental Material). Furthermore, the NI-STOI has a very limited dynamic range. This effect could also be observed with the combination bBSIM-NI-STOI with German speech material [42]. Certainly, the reason for this flat curve is that NI-STOI uses no auxiliary information. In other words, it has to solve a much more complex task than the non-blind back-ends, and even at a higher SRT, it is still difficult to estimate the envelope of the target signal based on the degraded signal. A similar problem occurred previously [8] when trying to use the SRMR as the final back-end of bBSIM.

The overestimation of BRM could be compensated for by increasing bBSIM's binaural processing errors, as was done by [20], for both the STI_{NCV} and the NI-STOI. In order to evaluate this, we increased the ITD jitter by a factor of 1.6 for the STI_{NCV} and by a factor of 2.1 for the NI-STOI. This increased the R^2 to 0.88 and decreased the RMSE to 1.7 for the STI_{NCV} and increased the R^2 to 0.91 and decreased the RMSE to 1.9 dB for the NI-STOI. The resulting scatter plot of all tested conditions is shown in Figure S4 in the Supplemental Material. The use of these increased binaural processing inaccuracies certainly improves the predictions. However, it contradicts the intention of this study to evaluate a blind binaural front-end that works independently of the back-end and without any auxiliary information about the listening condition. It is not clear why the required increase of the EC processing inaccuracies depends on the back-end. Further research and development is required in order to find a combination of binaural front-end and back-end that works for arbitrary conditions without further adjustments of the model's parameters.

4.2 Limitations of this study and outlook

The original NI-STOI [32] comes with a voice activity detection (VAD) feature that was not used in this study. However, we made informal simulations using the VAD and found no difference to the predictions shown here. This is certainly due to our experimental design, in which there were no passages without target speech.

As mentioned above, bBSIM-NI-STOI cannot be expected to work optimally with modulated maskers, because the SRMR in the front-end and the NI-STOI analyze the speech-like modulations and will be disturbed by modulations of speech(-like) interferers. Note that the intrusive version of STOI (which is similar to the STI_{NCV} used in this study) also does not work well for modulated maskers [44] compared to other models [45, 46]. In this study, we did not use modulated noise. An evaluation using modulated noises would certainly require the use of a short-term version of bBSIM similar to the short-term version of BSIM introduced in [7] preferably with the extension that takes human binaural sluggishness into account [47].

In future studies, bBSIM should be combined with more elaborate back-ends. A first investigation has been done in [43], where bBSIM was combined with a blind back-end based on a deep neural network from an automatic speech recognition (ASR) system, which was combined with an entropy measure, the mean temporal distance (MTD) [48]. The MTD used by [43] is based on the assumption that with a degraded signal, the phonemes are less distinctive, leading to a smeared phoneme probability over time. The MTD identifies degradation of the signal by measuring the diversity of phoneme vectors over time. This combination was introduced as binaural ASR-based prediction of speech intelligibility (BAPSI) [43]. Predictions with BAPSI showed high prediction accuracy for reverberant conditions [43]. However, this approach is computationally much more demanding and would go beyond the scope of this study. As a further alternative, the combination of bBSIM with the simulation framework for auditory discrimination experiments (FADE) [49] is promising and has been evaluated successfully in [42]. FADE does not calculate an index like SII, STI, STOI, or MTD, but instead calculates recognition rates directly. However, this has the drawback that FADE requires a transcription of the test sentences, whereas back-ends like NI-STOI [32] and MTD [48] do not. Furthermore, FADE requires a very extensive training using the same kind of sentences used in the final test.

One of the largest challenges we see for future work are speech-in-speech conditions with speech interferers that cause informational masking [50]. As far as we know, there is no model that is able to predict human speech intelligibility in binaural conditions with speech maskers, at least for

conditions that are strongly influenced by informational masking. Modelling speech intelligibility for speech-in-speech conditions will require taking into account, which of the competing speech sources is the target and which is the interferer. This will probably require inclusion of effects like localization and auditory scene analysis into the model, as done for example by [51], which both go far beyond the scope of this study.

5 Conclusions

In this study, we modelled SRTs for listening conditions with complex reflections of the target speech and different IPD settings in stationary noise. The blind front-end bBSIM was combined with three different back-ends, resulting in two hybrid models (bBSIM-SII and bBSIM-STI_{NCV}) and one fully blind model (bBSIM-NI-STOI) that predict the interaction between binaural and temporal processing. These are the main conclusions of this study:

- The very low SRTs of the applied speech material caused an overestimation of BRM. As a workaround, it is possible to increase the reference value of the model back-end. This worked very well for the SII and we recommend this method for all studies where the SII is applied to speech with a very low SRT.
- The fully blind model (bBSIM-NI-STOI) achieved a higher R^2 compared to the hybrid model bBSIM-SII and an R^2 similar to the hybrid model bBSIM-STI_{NCV}.
- The fully blind model (bBSIM-NI-STOI) achieved an R^2 similar to the non-blind baseline model, which receives information about the BRIR, the target speech signal and the noise. However, the fully blind model overestimated the BRM, resulting in an RMSE of 4.4 dB, compared to an RMSE of 1 dB for the baseline model. In the baseline model, overestimates of BRM were avoided by increasing the binaural processing errors of the EC stage.
- With the increase of the binaural processing errors of the EC stage, the RMSE of the bBSIM-STI_{NCV} decreased to 1.7 and the RMSE of the bBSIM-NI-STOI decreased to 1.9 dB without a relevant decrease in R^2. However, this adjustment of the binaural processing errors contradicts the intention of a blind binaural front-end that works independently of the back-end.

Supplemental material

Supplementary material is available at https://acta-acustica.edpsciences.org/10.1051/aacus/2022009/olm

Figure S1: Predicted binaural release from masking (BRM, shown on the y-axis) defined as the difference between SRTs of the first condition of experiment I (D_0N_0) and the first condition of experiment III ($D_\pi N_0$) using bBSIM-SII (black line), STI_{NCV} (light gray line), and NI-STOI (dark gray line). The back-end reference values are shown on the x-axis. The red line shows the measured BRM of 7.8 dB. The yellow circles show the optimal references, which allows predicting accurately the measured BRM for the SII and the STI_{NCV} (see Fig. S3). For the NI-STOI the optimal reference allows predicting the measured BRM with less overestimation (see Fig. S3).

Figure S2: Measured (gray circles connected with gray solid line) and predicted SRTs for experiment III with bBSIM-SII uses the optimal reference (black stars connected with black dashed line) and the original reference (black diamonds connected with dashed lines). See Section 2.5 for details. This figure shows a single reflection of the target speech, which was varied in its delay time. Note that the first condition is always direct sound only.

Figure S3: Difference between predictions of the BRM using the original reference (gray circle) and the optimal reference (yellow circle) for the SII (upper panel), STI_{NCV} (middle panel), and NI-STOI (lower panel). Measured BRM of 7.8 dB is marked as red line. See Section 2.5 for details.

Figure S4: Overall prediction accuracy of bBSIM-STI_{NCV} and bBSIM-NI-STOI with increased binaural processing errors (by a factor of 1.6 for STI_{NCV} and 2.1 for NI-STOI), with coefficient of determination (R^2), mean absolute prediction error (ϵ_{mean}), maximum absolute prediction error (ϵ_{max}), and root-mean-square error (RMSE) for all 16 experiments.

Data availability statement

Implementation of the model bBSIM-SII [8] is publicly available as part of the Auditory Modeling Toolbox as (hauth2020) (AMT, https://www.amtoolbox.org) [52] in the release of the version 1.0.0 available as a full package for download (https://sourceforge.net/projects/amtoolbox/files/AMT%201.x/amtoolbox-full-1.0.0.zip/download) [53]. The current model version is freely available here: http://medi.uni-oldenburg.de/BSIM_2020/.

Conflict of interest

The authors declare that they have no conflicts of interest in relation to this article.

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 352015383 – SFB 1330 A1.

References

1. E.C. Cherry: Some experiemnts on the recognition of speech, with one and with two ears. The Journal of the Acoustical Society of America 25 (1953) 975–979.
2. 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.
3. L. Rayleigh: On our perception of sound direction. Philosophical Magazine 13 (1907) 214–223.

4. N.I. Durlach: Equalization and cancellation theory of binaural masking-level differences. The Journal of the Acoustical Society of America 35, 8 (1963) 1206–1218.
5. A.H. Andersen, J.M. De Haan, Z.H. Tan, J. Jensen: Predicting the intelligibility of noisy and nonlinearly processed binaural speech. IEEE/ACM Transactions on Audio Speech and Language Processing 24, 11 (2016) 1908–1920.
6. R. Beutelmann, T. Brand: Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. The Journal of the Acoustical Society of America 120, 1 (2006) 331–342.
7. R. Beutelmann, T. Brand, B. Kollmeier: Revision, extension, and evaluation of a binaural speech intelligibility model. The Journal of the Acoustical Society of America 127, 4 (2010) 2479–2497.
8. C.F. Hauth, S.C. Berning, B. Kollmeier, T. Brand: Modelling binaural unmasking of speech using a blind binaural processing stage. Trends in Hearing 24 (2020) 1–16.
9. S. Jelfs, J.F. Culling, M. Lavandier: Revision and validation of a binaural model for speech intelligibility in noise. Hearing Research 275, 1–2 (2011) 96–104.
10. M. Lavandier, J.F. Culling: Prediction of binaural speech intelligibility against noise in rooms. The Journal of the Acoustical Society of America 127, 1 (2010) 387–399.
11. M. Lavandier, S. Jelfs, J.F. Culling, A.J. Watkins, A.P. Raimond, S.J. Makin: Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources. The Journal of the Acoustical Society of America 131, 1 (2012) 218–231.
12. R. Wan, N. Durlach, H.S. Colburn: Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers. The Journal of the Acoustical Society of America 128, 6 (2010) 3678–3690.
13. E.L.J. George, S.T. Goverts, J.M. Festen, T. Houtgast: Measuring the effects of reverberation and noise on sentence intelligibility for hearing-impaired listeners. Journal of Speech, Language, and Hearing Research 53, 6 (2010) 1429–1439.
14. S. Hochmuth, T. Jürgens, T. Brand, B. Kollmeier: Talker- and language-specific effects on speech intelligibility in noise assessed with bilingual talkers: Which language is more robust against noise and reverberation? International Journal of Audiology 54, March 2016 (2015) 23–34.
15. J. Rennies, T. Brand, B. Kollmeier: Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet. The Journal of the Acoustical Society of America 130, 5 (2011) 2999–3012.
16. A. Warzybok, J. Rennies, T. Brand, S. Doclo, B. Kollmeier: Effects of spatial and temporal integration of a single early reflection on speech intelligibility. The Journal of the Acoustical Society of America 133, 1 (2013) 269–282.
17. I. Arweiler, J.M. Buchholz: The influence of spectral characteristics of early reflections on speech intelligibility. The Journal of the Acoustical Society of America 130, 2 (2011) 996–1005.
18. J.S. Bradley, H. Sato, M. Picard: On the importance of early reflections for speech in rooms. The Journal of the Acoustical Society of America 113, 6 (2003) 3233.
19. J.P.A. Lochner, J.F. Burger: The influence of reflections on auditorium acoustics. Journal of Sound and Vibration 1, 4 (1964) 426–454.
20. J. Rennies, A. Warzybok, T. Brand, B. Kollmeier: Measurement and prediction of binaural-temporal integration of speech reflections. Trends in Hearing 23 (2019) 1–22.
21. ANSI: ANSI S3.5-1997, American national standard methods for calculation of the speech intelligibility index. Am. Natl. Stand. Institute, New York, 1997.
22. T. Leclère, M. Lavandier, J.F. Culling: 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, 6 (2015) 3335–3345.
23. J. Rennies, A. Warzybok, T. Brand, B. Kollmeier: Modeling the effects of a single reflection on binaural speech intelligibility. The Journal of the Acoustical Society of America 135, 3 (2014) 1556–1567.
24. S. Cosentino, T. Marquardt, D. McAlpine, J.F. Culling, T.H. Falk: A model that predicts the binaural advantage to speech intelligibility from the mixed target and interferer signals. The Journal of the Acoustical Society of America 135, 2 (2014) 796–807.
25. M. Geravanchizadeh, A. Fallah: Microscopic prediction of speech intelligibility in spatially distributed speech-shaped noise for normal-hearing listeners. The Journal of the Acoustical Society of America 138, 6 (2015) 4004–4015.
26. M. Dietz, S.D. Ewert, V. Hohmann: Auditory model based direction estimation of concurrent speakers from binaural signals. Speech Communication 53, 5 (2011) 592–605.
27. T. Dau, D. Püschel, A. Kohlrausch: A quantitative model of the "effective" signal processing in the auditory system. I. Model structure. The Journal of the Acoustical Society of America 99, 6 (1996) 3615–3622.
28. H. Sakoe, S. Chiba: Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 26, 1 (1978) 43–49.
29. H.J.M. Steeneken, T. Houtgast: A physical method for measuring speech-transmission quality. The Journal of the Acoustical Society of America 67, 1 (1980) 318–326.
30. I. Holube, B. Kollmeier: Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. The Journal of the Acoustical Society of America 100, 3 (1996) 1703–1716.
31. C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen: An algorithm for intelligibility prediction of time – frequency weighted noisy speech. IEEE Transaction on Audio, Speech, and Language Processing 19, 7 (2011) 2125–2136.
32. A.H. Andersen, J.M. Haan, Z. Tan, J. Jensen: A non-intrusive short-time objective intelligibility measure, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, United States, 5 March, 2017, pp. 5085–5089.
33. B. Kollmeier, A. Warzybok, S. Hochmuth, M.A. Zokoll, V. Uslar, T. Brand, K.C. Wagener: The multilingual matrix test: Principles, applications, and comparison across languages: A review. International Journal of Audiology 54, December (2015) 3–16.
34. T. Brand, B. Kollmeier: Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. The Journal of the Acoustical Society of America 111, 6 (2002) 2801–2810.
35. V. Hohmann: Frequency analysis and synthesis using a Gammatone filterbank. Acta Acustica United with Acustica 88, 3 (2002) 433–442.
36. B.C.J. Moore, B.R. Glasberg: Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. Journal of the Acoustical Society of America 74, 3 (1983) 750–753.
37. H. vom Hövel: Zur Bedeutung der Übertragungseigenschaften des Aussenohrs sowie des Binauralen Hörsystems bei Gestörter Sprachübertragung [On the importance of the transmission properties of the outer ear and the binaural auditory system in disturbed speech transmission]. [PhD dissertation]. RWTH Aachen, Aachen, Germany, 1984.

38. J.F. Santos, M. Senoussaoui, T.H. Falk: An improved non-intrusive intelligibility metric for noisy and reverberant speech, in 2014 14th International Workshop on Acoustic Signal Enhancement, IWAENC 2014, Juan-les-Pins, France, September 8–11, 2014, pp. 55–59.
39. A.H. Andersen: Speech Intelligibility Predictors. Retrieved date: 2nd May 2022. http://ah-andersen.net/code/.
40. 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.
41. J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, V. Zue: TIMIT Acoustic-phonetic continous speech corpus. LDC93S1. Web Download. Lingistic Data Consortium, Philadelphia, 1993.
42. D. Hülsmeier, C.F. Hauth, S. Röttges, P. Kranzusch, J. Roßbach, M.R. Schädler, B.T. Meyer, A. Warzybok, T. Brand: Towards non-intrusive prediction of speech recognition thresholds in binaural conditions, in Speech Communication; 14th ITG Conference, Kiel, Germany, 29 September – 1 October, 2021, pp. 1–5.
43. J. Roßbach, S. Röttges, F.C. Hauth, T. Brand, B.T. Meyer: Non-intrusive binaural prediction of speech intelligibility based on phoneme classification, in ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, ON, Canada, 6–11 June, 2021, pp. 369–400.
44. W. Schubotz, T. Brand, B. Kollmeier, S.D. Ewert: Monaural speech intelligibility and detection in maskers with varying amounts of spectro-temporal speech features. The Journal of the Acoustical Society of America 140, 1 (2016) 524–540.
45. S. Jørgensen, S.D. Ewert, T. Dau: A multi-resolution envelope-power based model for speech intelligibility. The Journal of the Acoustical Society of America 134, 1 (2013) 436–446.
46. K.S. Rhebergen, N.J. Versfeld, W.A. Dreschler: Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. The Journal of the Acoustical Society of America 120, 6 (2006) 3988–3997.
47. C.F. Hauth, T. Brand: Modeling sluggishness in binaural unmasking of speech for maskers with time-varying interaural phase differences. Trends in Hearing 22 (2018) 1–10.
48. H. Hermansky, E. Variani, V. Peddinti: Mean temporal distance: Predicting ASR error from temporal properties of speech signal, in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings, Vancouver, BC, Canada, 26–31 May, 2013.
49. M.R. Schädler, D. Hülsmeier, A. Warzybok, S. Hochmuth, B. Kollmeier: Microscopic multilingual matrix test predictions using an ASR-based speech recognition model, in 17th Annual Conference of the International Speech Communication Association (Interspeech 2016), San Francisco, CA, USA, September 8–12, 2016, pp. 610–614.
50. G. Kidd, C.R. Mason, V.M. Richards, F.J. Gallun, N.I. Durlach: Informational masking, in Auditory Perception of Sound Sources, Yost W.A., Popper A.N., Fay R.R., Editors. New York. Springer. 2008, pp. 143–190.
51. J. Mi, H.S. Colburn: A binaural grouping model for predicting speech intelligibility in multitalker environments. Trends in Hearing 20 (2016) 1–12.
52. P. Majdak, C. Hollomey, R. Baumgartner: Amt 1.x: A toolbox for reproducible research in auditory modeling. Acta Acustica 6 (2021) 19. https://amtoolbox.org/cite_us.php.
53. The AMT Team: The Auditory Modeling Toolbox Full Package (version 1.x) [Code]. 2021. https://sourceforge.net/projects/amtoolbox/files/AMT%201.x/amtoolbox-full-1.0.0.zip/download.

Cite this article as: Röttges S. Hauth CF. Rennies J. & Brand T. 2022. Using a blind EC mechanism for modelling the interaction between binaural and temporal speech processing. Acta Acustica, **6**, 21.