Original Article

Measurement and Prediction of Binaural-Temporal Integration of Speech Reflections

Trends in Hearing
Volume 23: 1–22

DOI: 10.1177/2331216519854267
journals.sagepub.com/home/tia

Jan Rennies[1,2], Anna Warzybok[3], Thomas Brand[3], and Birger Kollmeier[2,3]

Abstract

For speech intelligibility in rooms, the temporal integration of speech reflections is typically modeled by separating the room impulse response (RIR) into an early (assumed beneficial for speech intelligibility) and a late part (assumed detrimental). This concept was challenged in this study by employing binaural RIRs with systematically varied interaural phase differences (IPDs) and amplitude of the direct sound and a variable number of reflections delayed by up to 200 ms. Speech recognition thresholds in stationary noise were measured in normal-hearing listeners for 86 conditions. The data showed that direct sound and one or several early speech reflections could be perfectly integrated when they had the same IPD. Early reflections with the same IPD as the noise (but not as the direct sound) could not be perfectly integrated with the direct sound. All conditions in which the dominant speech information was within the early RIR components could be well predicted by a binaural speech intelligibility model using classic early/late separation. In contrast, when amplitude or IPD favored late RIR components, listeners appeared to be capable of focusing on these components rather than on the precedent direct sound. This could not be modeled by an early/late separation window but required a temporal integration window that can be flexibly shifted along the RIR.

Keywords

speech intelligibility, binaural hearing, temporal integration, reflections

Date received: 29 November 2018; revised: 10 April 2019; accepted: 9 May 2019

Introduction

In real rooms, acoustic signals are reflected from objects and room boundaries, which produces a generally complex pattern of sound propagation between any source and receiver, comprising the direct sound, the early reflections within the first 50 to 100 ms after the direct sound, as well as later reflections that ultimately combine to the late reverberation. In terms of speech intelligibility, it is generally agreed that early reflections are beneficial, that is, that they can be integrated (at least partially) with the direct sound and improve speech recognition (e.g., Arweiler & Buchholz, 2011; Bradley, Sato, & Picard, 2003; Lochner & Burger, 1964). In contrast, late reflections cannot be integrated with the direct sound and can be detrimental for speech intelligibility. This study investigates how the integration of speech reflections depends on their amplitude, their delay relative to the direct sound as well as the binaural information contained in them, such as the interaural level or time differences, and how these effects can be predicted by binaural speech intelligibility models (BSIMs).

The temporal integration of speech reflections and reverberation has been investigated in many studies. While late reverberation typically decreases speech intelligibility (e.g., George, Goverts, Festen, & Houtgast, 2010; Hochmuth, Jürgens, Brand, & Kollmeier, 2015; Rennies, Brand, & Kollmeier, 2011; Steeneken &

[1]Department of Speech, Language and Hearing Sciences, Boston University, Boston, MA, USA
[2]Project Group Hearing, Speech and Audio Technology, Fraunhofer Institute for Digital Media Technology IDMT, Cluster of Excellence Hearing4all, Oldenburg, Germany
[3]Medical Physics Group, Department of Medical Physics and Acoustics, Cluster of Excellence Hearing4all, University of Oldenburg, Germany

Corresponding author:
Jan Rennies, Fraunhofer IDMT, Hearing, Speech and Audio Technology, 26129 Oldenburg, Germany.
Email: jan.rennies@idmt.fraunhofer.de

Houtgast, 1980), reflections arriving shortly after the direct sound can be beneficial. For example, Lochner and Burger (1964) found that adding a delayed copy of the direct sound at different delays (while keeping the direct sound constant) improved speech intelligibility. For delays up to about 30 ms, this improvement was the same as measured for a 3-dB increase of speech level in conditions with no reflection, that is, the reflection could be perfectly integrated with the direct sound. For longer delays, the improvement decreased and disappeared at a delay of 95 ms. Similar durations of the temporal window for perfect integration (25–50 ms) were reported by other studies for a single reflection (Nábělek & Robinette, 1978; Warzybok, Rennies, Brand, Doclo, & Kollmeier, 2013) and multiple early reflections (Bradley et al., 2003). In contrast, other studies found that adding speech energy as early reflections was less beneficial than adding the same energy as direct sound (Arweiler & Buchholz, 2011; Parizet & Polack, 1992; Soulodre, Popplewell, & Bradley, 1989), that is, the temporal integration of early reflections was less than perfect even at very short delays.

A few recent studies explicitly investigated the interaction of temporal speech integration and binaural processing. Arweiler and Buchholz (2011) measured speech intelligibility in diffuse noise by varying the signal-to-noise ratio (SNR) by either increasing the direct sound energy or the energy of the early reflections. They employed binaural room impulse responses (BRIRs) containing 20 reflections, all arriving within 55 ms after the direct sound. The stimuli were presented via a 29-loudspeaker setup, which approximately maintained the azimuthal direction and elevation of the reflections as well as their spectral characteristics (which varied due to frequency-dependent absorption). As control conditions, Arweiler and Buchholz (2011) also presented all reflections from the frontal loudspeaker, that is, colocated with the direct sound, and, in addition, included monaural presentation where one ear was blocked by an insert earphone playing masking noise. The main findings were that an increase in direct-speech energy was more beneficial than an equivalent increase in reflection energy, and that this difference was smaller when all reflections were presented from the front. Arweiler and Buchholz (2011) concluded that temporal integration of early reflections was facilitated when they arrived from the same direction as the direct sound. In addition, they found that speech intelligibility was better in binaural than in monaural listening conditions by 2 to 3 dB, which could be explained by spatial unmasking in the presence of the diffuse masker. Since this spatial unmasking was the same for frontal and spatially distributed early reflections, Arweiler and Buchholz (2011) argued that the binaural system could not integrate early reflections more efficiently than the monaural system and that, therefore, temporal processing and binaural processing were independent.

This finding was confirmed by Warzybok et al. (2013) for conditions with frontal direct sound and a single, equally strong frontal reflection. They found that speech recognition thresholds (SRTs, i.e., the SNRs required to achieve 50% speech intelligibility) were shifted downward by a constant amount of about 4 and 8 dB when using diffuse noise or lateral noise, respectively, instead of colocated noise, independent of the delay of the reflection. This was no longer the case, however, when the reflection was not colocated with the direct sound. When the delay was short ($\leqslant$ 50 ms), SRTs were the same as for colocated reflections. In contrast, when a long delay of 200 ms was used, the detrimental effect observed for colocated reflections was considerably reduced when the reflection arrived from the same hemisphere as the lateral noise but not when it arrived from the opposite hemisphere. This was called *suppression effect* of a detrimental reflection by Warzybok et al. (2013) and was interpreted as an indication for an interaction of binaural and temporal processing. Interestingly, Warzybok et al. (2013) did not find evidence for the finding of Arweiler and Buchholz (2011) that useful reflections could be integrated more efficiently when they arrived from the same direction as the direct sound. Warzybok et al. (2013) suggested that this discrepancy might be due the fact that they had used a single reflection (while Arweiler & Buchholz, 2011, had used 20 early reflections) and that the binaural system might be unable to integrate more than one or a low number of non-colocated reflections. One goal of this study was therefore to extent the approach of Warzybok et al. (2013) by using BRIRs with more than one reflection.

A further goal of this study was to challenge current modeling approaches to predict speech intelligibility. Various studies as well as commercial tools employ basic room acoustical measures as predictors for speech intelligibility. Among the most widely used measures considering the contribution of early and late reflections to speech recognition are the *definition* (i.e., the ratio of the early components of a room impulse response (RIR) to the overall energy of the RIR) and the *clarity* measure (i.e., the ratio of the early-component energy to the late-component energy, typically expressed in decibel). These measures are based on the aforementioned concept that the RIR is split into an *early* part, which is assumed to be useful, and a *late* part, which is assumed to be detrimental. Typical values for the separation time between early and late part, t_e, are between 50 and 100 ms (see Rennies et al., 2011). Rennies et al. (2011) and also Leclère, Lavandier, and Culling (2015) extended this concept of monaural room acoustical measures to binaural prediction models: The binaural input signals were split into a

useful and a detrimental part based on the BRIR, and the detrimental part was simply added to the masking noise, while the early part was kept as target signal. This splitting was conducted prior to modeling binaural processing in both models. This approach was highly successful in predicting SRTs in various reverberant conditions (see Leclère et al., 2015; Rennies et al., 2011). In particular, Rennies, Warzybok, Brand, and Kollmeier (2014) showed that this approach could also predict the interaction of temporal and spatial processing as observed by Warzybok et al. (2013) (for details, see section "SRT predictions").

These model approaches, as well as established measures like definition and clarity, are based on the assumption that the useful part consists of the early BRIR components (i.e., direct sound plus reflections delayed by up to t_e), while the detrimental part consists of the late components (i.e., reflections delayed by $> t_e$). However, Leclère et al. (2015) found that using a fixed value t_e was not sufficient to accurately predict SRTs in rooms with different degrees of reverberation. They argued that this might be due to the capability of the auditory system to adapt to different listening environments. This would suggest a flexible spatio-temporal integration process which is not currently included in any speech intelligibility model. Furthermore, while assuming a useful early part and a detrimental late part is reasonable for most realistic listening conditions due to the typically decaying nature of the BRIRs over time, there may be conditions in which it is not appropriate. For example, when listening to a voice amplified by a public-address system, the amplified version may be considerably delayed but much higher in level than the direct sound. In the extreme case, the direct sound becomes barely audible and the (delayed) playback voice becomes the only target speech source. Hence, it seems reasonable to assume that—at some point—the delay of the reflection is no longer relevant because the reflection will dominate speech intelligibility and the direct sound will become a barely perceptible preecho. In other words, the detrimental effect of a late reflection should be strongest when direct sound and reflection have a similar level and should disappear when either of these two components has a much lower level than the other. While this effect seems rather predictable, it cannot be modeled by current room acoustical measures assuming that the useful window starts at the direct sound. It could, however, be at least partly modeled by assuming that the component of the BRIR with the highest level determines the position of the temporal window to extract the useful components from the RIR. However, choosing the useful part based only on the amount of energy might also be wrong, for example, when a reflection is not dominant in terms of energy but carries an interaural phase difference (IPD) different than that of the masker. We are not aware of data to quantitatively test the role of late reflections with favorable energy or binaural information. The experimental conditions measured in this study were therefore designed to test these effects and their predictability by current models.

It is unclear how speech perception is affected when reflections and direct sound differ in the "binaural advantage" they contribute to intelligibility. In an extreme case, stimuli could be designed in which only a late (and normally detrimental) reflection comprises a binaural advantage relative to a given masking noise, while the (normally useful) direct sound does not. It is possible that under such conditions, the reflection becomes the dominant component for speech intelligibility (like in the case when it has a considerably higher level). Some conditions of this study were designed to test this hypothesis. If it is true, then this means that a simple *maximum detector* scanning of the RIR at each ear would not be sufficient to determine the optimal position of the temporal integration window of useful speech components. To test this, all measured conditions of this study were compared with model predictions of the model proposed by Rennies et al. (2014). As stated earlier, this model assumes that the useful window starts at the direct sound and includes only early BRIR components. It should, thus, fail in conditions in which the experimental data are strongly affected by late reflections being the dominant speech source.

To provide a sound data base for assessing the integration of spatial, temporal, and energetic properties of speech reflections as well as the interactions between these properties, this study measured SRTs in a total of 86 different combinations in which the reflection delay, reflection amplitude, number of reflections, as well as the relative binaural advantage of the reflections and direct sound in comparison to a stationary masker were varied. The focus here was on binaural processing and temporal integration, that is, unlike in the studies of Warzybok et al. (2013) and Arweiler and Buchholz (2011), this study minimized head-shadow effects and (frequency-dependent) better-ear listening. While these are relevant for real listening conditions, they may also dominate the overall spatial benefit (e.g., Rennies & Kidd, 2018; Warzybok et al., 2013). Therefore, binaural information was introduced by modifying IPDs to create stimuli with no better-ear advantage.

The analyses and discussions of the present data are organized in 20 experiments, where Experiments I to VII investigate the effects of delay, IPD, and amplitude of a single reflection and Experiments VIII to XVI investigate the effects of successively increasing the number of reflections (with different delays and IPDs). In addition, Experiments XVII to XX investigate the effects of varying the IPD-advantage per reflection within BRIRs comprising a fixed number and temporal configuration

of early and late reflections. These last four experiments are described in the provided Supplementary Material.

Methods

Listeners

The experimental conditions of this study were divided in two parts (A and B, see later). For each part, data of eight listeners were collected. Listeners participating in the first part were between 18 and 27 years of age. Four of them also participated in the second part, for which listeners ranged in age from 18 to 31 years. All had English as their native language and had pure-tone thresholds not exceeding 20 dB hearing level at audiometric frequencies between 250 Hz and 8 kHz. Listeners conducted the measurements in several sessions of 1 to 2 hr, were paid for their participation, and gave an informed consent. All procedures were approved by the Boston University Institutional Review Board (Protocol 2633E).

Stimuli and Conditions

The target speech was uttered by a female talker and consisted of sentences taken from the American English matrix sentence test (Kollmeier et al., 2015), for example, *Peter has eight green sofas*. These sentences always have the fixed five-word structure name-verb-numeral-adjective-object. For each word group, 10 alternatives are available, which can be randomly combined to produce syntactically correct but semantically unpredictable sentences. The test material consists of 90 such sentences, which are combined to lists of 20 or 30 sentences. Due to the lack of semantic predictability, each sentence appears to the listeners as one of the 10^5 possible random combinations, and memorizing any of the ninety sentences is not likely, allowing for multiple measurements with the same target material. The sentence lists have been optimized to produce highly homogeneous SRTs (see Kollmeier et al., 2015). The masker consisted of stationary speech-shaped noise generated from the speech material so that the long-term noise spectrum matched that of the speech material.

To generate the desired combinations of speech components (direct sound and one or several reflections), the target speech was convolved with artificially created BRIRs. The BRIRs were created based on the BRIR for frontal sound incidence in anechoic conditions (i.e., without reflections) employed by Warzybok et al. (2013). This BRIR had been simulated with the CATT Acoustic software v8.0 a (CATT, Gothenburg, Sweden) by using an omnidirectional source in an anechoic room and modeling the receiver as a head-and-torso simulator (KEMAR; G.R.A.S., Sound & Vibration, Holte, Denmark) at a distance of 5 m from the source. This BRIR comprising only direct sound was used as the basic component to create new BRIRs by introducing identical copies at specific delays to produce the desired reflections, resulting in BRIRs with between 1 (direct sound only) and 10 components (direct sound plus nine reflections). The delay Δt of the reflections was varied systematically and was 10, 25, 50, 75, 100, 125, 150, 175, and 200 ms. In addition, the IPD of the different components was manipulated, that is, the direct sound could be either diotic (D_0) or have an IPD of 180° (D_π). The IPD manipulation was realized by inverting the phase at the left ear. The same IPD manipulation was conducted for some or all of the individual reflections (denoted by R_0 or R_π in the following) and the masking noise (N_0 or N_π). In most experiments of this study, all components of the BRIR had the same level, but in some experiments, the relative level of direct sound and reflections was varied. This was achieved by multiplying the reflection by an amplitude amplification factor α before copying the BRIR component at the desired delay of the BRIR. This manipulation ensured that both speech and noise always had the same level at both ears and, hence, that there was no monaural SNR advantage at either ear.

Altogether, 86 different conditions were created with different parameter variations. These conditions were combined to 20 experiments, where some conditions served as anchor points for several experiments (e.g., the direct sound-only condition was included as reference point in several experiments as described in the following). Table 1 summarizes the different combinations of BRIR components, the reflection delays relative to the direct sound, their amplitude amplification factor α, their IPD, and the IPD of the noise. In Experiments I to VII (Part A), target speech always consisted of the direct sound and a single reflection, and the IPDs (of D, R, and N), the reflection delay, as well as the reflection amplification factor α were varied. Part B (Experiments VIII to XX) comprised BRIRs with several reflections. In Experiments VIII to XI, the effect of adding an increasing number of reflections was investigated, and the reflections were successively added starting from the lowest delay (i.e., at 10 ms and then increasing). Experiments XIII to XVI also explored the effect of adding an increasing number of reflections, but here the reflections were successively added from largest delay (i.e., at 200 ms and then decreasing). Experiment XII was included to investigate the effect of multiple reflections in N_π noise. In Experiments XVII to XX (see Supplementary Material), the number of reflections was fixed at nine (i.e., there was a reflection at each delay between 10 and 200 ms), but the IPDs of the reflections differed. In Experiment XVII, for example, an increasing number of early reflections (i.e., starting with the reflection at 10 ms) had an IPD of 0, while the later reflections had an IPD of π.

Table 1. Overview of Measurement Conditions.

Experiment	D-IPD	N-IPD	No. of reflections	Reflection delay Δt/ms	Reflection amp. α	R-IPD
I	0	0	1	10, 50, 100, 150, 200	1	0
II	0	0	1	10, 25, 50, 75, 100, 150, 200	1	π
III	π	0	1	10, 25, 75, 150, 200	1	0
IV	0	π	1	10, 100, 200	1	π
V	π	π	1	10, 100, 200	1	0
VI	0	0	1	200	0, 0.25, 0.75, 1.0, 1.25, 2.0, 2.5	0
VII	0	0	1	200	0, 0.25, 0.5, 1.0, 1.75, 2.5	π
VIII	0	0	1	10	1	0
			2	10, 25		
			3	10, 25, 50		
			5	10, 25, 50, 75, 100		
			7	10, 25, 50, 75, 100, 125, 150		
			9	10, 25, 50, 75, 100, 125, 150, 175, 200		
IX	0	0	1	10	1	π
			3	10, 25, 50		
			5	10, 25, 50, 75, 100		
			7	10, 25, 50, 75, 100, 125, 150		
			9	10, 25, 50, 75, 100, 125, 150, 175, 200		
X	Same as IX, but with D_π and R_0 instead of D_0 and R_π					
XI	Same as IX, but with D_π instead of D_0					
XII	0	π	3	10, 25, 50	1	0
			5	10, 25, 50, 75, 100		
			9	10, 25, 50, 75, 100, 125, 150, 175, 200		
XIII	0	0	3	150, 175, 200	1	0
			5	100, 125, 150, 175, 200		
			9	10, 25, 50, 75, 100, 125, 150, 175, 200		
XIV	0	0	1	200	1	π
			2	175, 200		
			3	150, 175, 200		
			5	100, 125, 150, 175, 200		
			7	50, 75, 100, 125, 150, 175, 200		
			9	10, 25, 50, 75, 100, 125, 150, 175, 200		
XV	π	0	1	200	1	0
			3	150, 175, 200		
			5	100, 125, 150, 175, 200		
			7	50, 75, 100, 125, 150, 175, 200		
			9	10, 25, 50, 75, 100, 125, 150, 175, 200		
XVI	Same as XV, but with D_0 instead of D_π					
XVII–XVIII	0	0	9	See Supplementary Material		
XIX–XX	π	0	9	See Supplementary Material		

Note. The second and third columns indicate the IPD of the direct sound (D) and noise (N), respectively, while the remaining columns indicate the properties of the reflection(s). IPD = interaural phase difference.

It is important to note that the overall speech level was always calculated including all speech components. This means that, for speech stimuli with at least one reflection, the absolute level of the direct sound was reduced at a given overall speech level, because the energy was spread across the BRIR components. This reduction of direct sound level depended on reflection amplitude and number of reflections, as illustrated in Table 2. The top part shows the increase in overall level of an arbitrary target sentence consisting of direct sound and a single reflection delayed by 200 ms when increasing the reflection amplification factor α. This increase is equivalent to the level attenuation required to restore the same overall level as for the same sentence consisting only of direct sound. The indicated overall level increase was about 3 dB for $\alpha = 1$, and smaller (larger) for $\alpha < 1$ ($\alpha > 1$). For the largest value of α employed in this study ($\alpha = 2.5$), the level difference was about 8 dB. The bottom part of Table 2 shows the overall level increase when successively adding more reflections (all with the same amplitude, i.e., $\alpha = 1$), starting at the lowest delay of 10 ms. For a single reflection added, the overall level increase was about 3 dB as for the 200-ms reflection and $\alpha = 1$. For more than one reflection added, the level difference increased up to about 10 dB for the largest number of reflections employed in this study. This means that a speech signal convolved with the BRIR including all nine reflections was attenuated by about 10 dB to achieve the same overall level as a speech signal convolved with the BRIR consisting of the direct sound only.

Calibration and Equipment

In all conditions of this study, the masker level was fixed at 65 dB sound pressure level (SPL, as in Warzybok et al., 2013) and the speech level was adjusted to produce the desired SNRs. All stimuli were generated and controlled using Matlab (Natick, MA). The AFC-Matlab framework of Ewert (2013) was used to measure SRTs. The digital output was D/A converted via an RME HDSP 9632 (ASIO) 24-bit sound card (RME, Chemnitz, Germany) and delivered to the listeners via HD280 pro headphones (Sennheiser, Wedemark, Germany) in a sound-attenuated booth. The setup was calibrated to SPL using a Brüel and Kjær (B&K, Nærum, Denmark) 4153 artificial ear, a B&K 4947 1/2 in. microphone, a B&K ZC-0032 preamplifier, and a B&K 2250 sound level meter. The right ear served as reference point for the calibration, but the level at the two ears was always the same within the limits of the headphones and the simulated BRIRs.

Procedure

The SRT measurements were conducted using a closed-set procedure, that is, listeners selected the words they had recognized after each sentence on a graphical user interface, which consisted of the entire matrix of 50 words. Listeners confirmed their choices by pressing a button, which triggered the presentation of the next sentence. Two initial training SRTs were measured using lists of 20 sentences to familiarize the listeners with the speech material and the task to reduce training effects typical for matrix sentence test (Kollmeier et al., 2015). Subsequently, the experimental conditions were measured, each with a new random list of 20 sentences. The initial SNR of the adaptive track was 0 dB. The SNRs of the subsequent sentence presentations varied adaptively using the procedure described by Brand and Kollmeier (2002) to converge to the SRT. The 20 experiments were grouped in two parts (Experiments I–VII and Experiments VIII–XX). Within each group, all conditions of all experiments were pooled and randomized. The SRT for each condition was measured completely before proceeding to the next condition. All measured SRTs were tested for normality using Shapiro–Wilk tests and analyzed by means of repeated-measures analyses of variance (ANOVAs) with Greenhouse–Geisser corrections. Post hoc tests were Bonferroni corrected for multiple comparisons.

SRT Predictions

Validating BSIM with early/late separation. To investigate the conditions of this study with a model-based approach,

Table 2. Exemplary Illustration of the Overall Level Increase With Increasing Reflection Amplification Factor α for the Case of a Single Reflection Delayed by 200 ms Relative to the Level of the Direct Sound Only (Top Part), and of the Overall Level Increase When Successively Adding Reflections With $\alpha = 1$ (Bottom Part).

α	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.50
ΔL/dB	0.0	−0.1	0.7	1.6	2.8	3.8	4.8	5.8	6.7	8.3
No. of reflections	0	1	2	3	4	5	6	7	8	9
ΔL/dB	0.0	2.8	4.9	6.4	7.4	8.3	8.9	9.4	10.0	10.4

Note. These level increases are equivalent to the attenuation required to restore the same overall level as for stimuli consisting only of direct sound.

SRT predictions were made using the model and parameters proposed by Rennies et al. (2014). This model was developed based on the BSIM proposed by Beutelmann, Brand, and Kollmeier (2010). BSIM receives the left and right ear signals of speech and noise as input and combines an equalization-cancellation (EC) processing stage (Durlach, 1963) with the Speech Intelligibility Index (SII, American National Standards Institute, 1997) as back end. The EC processing is conducted independently in each of 30 Gammatone filters representing auditory filters, that is, in each filter, an interaural delay parameter and an interaural amplification parameter are optimized such that the noise is equalized. These equalized left and right ear signals are then subtracted from each other (cancellation), which can result in an improved SNR compared with the individual ear signals for listening conditions in which speech and noise differ in ITD or IPD or ILD. The accuracy of the cancellation is limited by the two internal processing errors ϵ and δ for the amplification and delay equalization, respectively. These processing errors are assumed to be of zero mean and normally distributed. The standard deviations are defined as the sum of a minimum value ($\sigma_{\epsilon 0}$ and $\sigma_{\delta 0}$, respectively) and a term increasing with the actual equalization parameters (for details, see Beutelmann et al., 2010). The binaurally enhanced SNRs are then fed into the SII, from which an SRT is derived by modifying the input SNR until a specified reference SII is reached. BSIM has been shown to provide highly accurate SRT predictions in conditions with near-field speech and various degrees of spatial unmasking (Beutelmann & Brand, 2006; Beutelmann et al., 2010; Hauth & Brand, 2018; Rennies et al., 2011). However, during the processing in the model (optimization of the EC parameters, SII calculation), the entire speech signal is considered as useful. In particular, late reflections and reverberation are considered to contribute to speech intelligibility in the same way as the direct sound and early reflections. Rennies et al. (2011, 2014) showed that this leads to prediction inaccuracies in conditions with strong reverberation and late reflections (see also Leclère et al., 2015). To overcome this limitation, Rennies et al. (2011) proposed a model extension based on the concept of separating the speech signal into an early (useful) and a late (detrimental) part by multiplying the BRIRs with weighting factors. In this concept, only the clean speech signal convolved with the early part of the binaural BRIRs is used as input speech, while the clean speech convolved with the late part of the BRIR is added to the noise input. Rennies et al. (2014) fitted the model parameters to the data of Warzybok et al. (2013). The early and late windows were always defined to be complementary, that is, their sum was always equal to 1. It was found that highest prediction accuracy was achieved for a temporal integration window with a transition time between early and late part of $t_e = 100$ ms, and a linear transition ramp with a decay duration (DD) of 200 ms to fade out the early part and fade in the late part. The shape of such a temporal window is illustrated in the left panel of Figure 1. This modeling

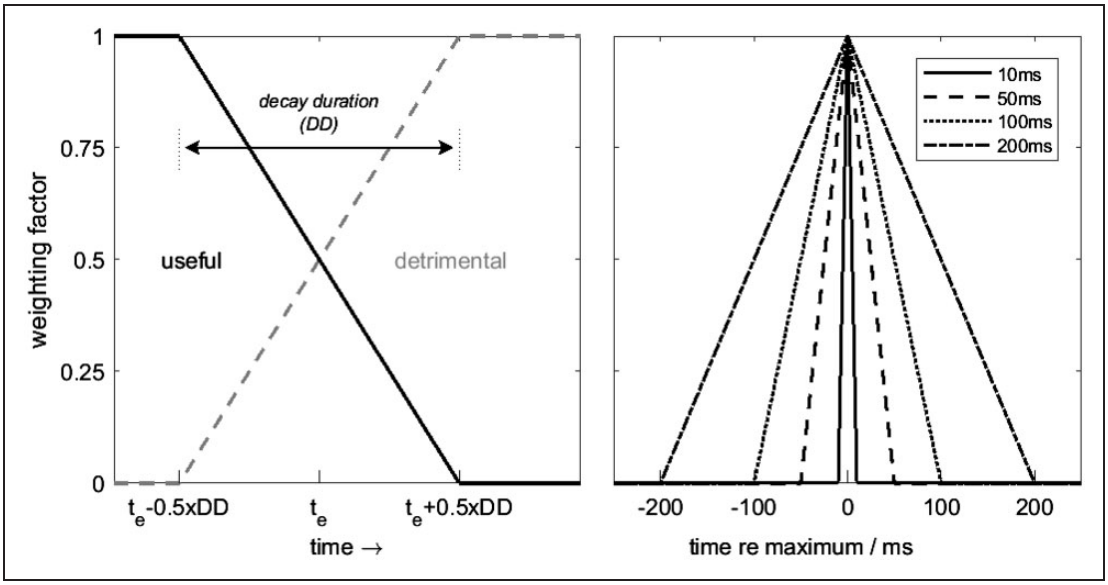

Figure 1. Left panel: illustration of the window shape to separate early and late part of the BRIRs by assuming that the BRIR components are either fully useful (up to $t_e - 0.5 \times$ DD), fully detrimental (after $t_e + 0.5 \times$ DD), or partially useful (during DD). Right panel: illustration of the window shape to separate useful and detrimental part of the BRIRs by assuming that a flexible window can be shifted flexibly along the BRIR. Different line styles illustrate different RDs.

approach thus assumes that the early part (E) is useful and the late part (L) is detrimental (referred to as BSIM-EL in the following). BSIM-EL achieved very high prediction accuracy ($R^2 = .95$, root-mean-square error [RMSE] < 1 dB) for all conditions tested by Warzybok et al. (2013). In particular, BSIM-EL predicted the interaction of spatial and temporal characteristics of all listening conditions with a single reflection.

One goal of this study was to extend the validation of the model to conditions with a larger variety of reflection properties and a larger number of reflections. In particular, the model was also challenged in conditions in which a late reflection (i.e., with a delay outside the temporal integration window for useful components) was assumed to be important for speech intelligibility. Furthermore, Rennies et al. (2014) found that different durations and shapes of the temporal integration window led to similar prediction accuracy. Another goal of this study therefore was to extend the database for fitting the model parameters and thereby get a clearer picture of the shape of the effective temporal integration window in binaural speech intelligibility.

In a first step, BSIM-EL was used as proposed by Rennies et al. (2014). The only adaptation was to adjust the reference SII to the listener group and speech material employed here. This was done by using the stimuli of the least complex data point of this study as reference condition (D_0N_0, no reflection), and then varying the reference SII until the predicted SRT matched the mean SRT for this data point. For the first part (Part A) of this study, this resulted in a reference value of 0.098. For the second part (Part B) with a partly different group of listeners (and, hence, a slightly different reference SRT), the reference SII value was 0.083. These values were then kept constant for predicting SRTs in all other conditions of the respective experiments. This approach revealed that BSIM-EL systematically overestimated the binaural benefit observed in the data. This overestimation calculated by comparing SRTs in the D_0N_0- and in the $D_\pi N_0$-conditions was about 2 and 2.5 dB for the two listener groups. Since we were interested in prediction deviations that occurred due to the model's (in-)capability to predict spatio-temporal interaction, we decided to conduct another adjustment by increasing the minimum standard deviation of the ITD processing error $\sigma_{\delta 0}$ by a factor of 1.6 and 1.8 for the two listener groups, respectively, to match the predicted binaural benefit in conditions not including temporal integration of reflections (D_0N_0 vs. $D_\pi N_0$) to the observed benefit. This ensured that any observed deviations between model predictions and experiments could be interpreted to reflect fundamental shortcomings to predict spatio-temporal integration rather than a general offset. Using these fixed values, SRTs of all conditions were predicted to validate the model proposed by Rennies et al. (2011) using the new data.

In a second step, the parameters t_e and DD of the temporal integration window were systematically varied to determine the best fit to the data set. This aimed to extend previous results of Rennies et al. (2014) and Leclère et al. (2015) as to measuring the effective shape of the early/late temporal integration window. Values of t_e ranged from 10 to 150 ms (step size 10 ms). For each t_e, five values of DD were tested by multiplying t_e by 0, 0.5, 1.0, 1.5, and 2.0. DD = 0 ms corresponds to the classic rectangular window. DD = 2.0·t_e corresponds to a ramp that starts decreasing immediately after the direct sound (i.e., without a constant weighting of 1 at the beginning).

For all predictions, speech-shaped noise was used to simulate the speech signal as in previous studies applying BSIM (Beutelmann & Brand, 2006; Beutelmann et al., 2010; Rennies et al., 2011, 2014). The duration of speech-simulating noise and masker was set to 3 s, and speech and masker noises were convolved with the same BRIRs as used in the experiments. Because the focus of this study was on stationary noise maskers, the long-term version of BSIM was used, that is, the capability of BSIM in the version of Beutelmann et al. (2010) to account for beneficial masker envelope fluctuations by employing short-term calculations similar to the concept of the extended SII (Rhebergen & Versfeld, 2005) was not used.

Temporally flexible useful or detrimental separation. The systematic parameter variation of BSIM-EL revealed that, for some conditions, the trend in the experimental data could not be well predicted by any combination of t_e and DD. This was presumably due to the fact that the late reflections (which are considered detrimental in the EL-approach) were relevant for speech intelligibility due to their level or IPD relative to direct sound and noise. To test this hypothesis within the framework of BSIM, a flexible temporal integration window was implemented. As in the EL-approach, this window consisted of linear ramps, but here both increasing and decreasing ramps were used resulting in a triangular shape of the integration window as illustrated in the right panel of Figure 1. For simplicity, symmetric ramps were used, that is, increasing and decreasing ramps always had the same ramp duration (RD). As before, the useful part of the BRIR was extracted by multiplying the BRIR with the useful (U) window (as illustrated in Figure 1), and the detrimental (D) part was extracted by multiplying the BRIR with the complementary window. These weighted BRIRs were then used to generate the target input (U) and an additional masker component (D) added to the external masker in the same way as in the EL-approach (this model is referred to as BSIM-UD in the following). The important difference between BSIM-EL and BSIM-UD was that, for BSIM-UD, the peak of the temporal window

was not fixed at the direct sound but could be moved flexibly along the BRIR. This approach was chosen to determine how model predictions compared with experimental data when allowing for full temporal flexibility, including a complete ignoring of the direct sound and focus on very late components which are normally assumed to be detrimental. Note that the previous EL approach is a special case of the flexible temporal window, that is, the peak of the flexible window could still be at the beginning of the BRIR (in which case only the decreasing ramp of the window would be used). Values of RD between 10 and 300 ms were tested (step size 10 ms), as well as very long RD up to 500 ms. This concept of temporal flexibility was based on the assumption that, for conditions with N_0-masker, the π-IPD of the BRIR components is potentially beneficial while, for conditions with N_π-masker, the 0-IPD of the BRIR-components is potentially beneficial. For each experiment and RD, the temporal position of the peak was placed such that the IPD-related binaural benefit from the BRIR could be maximally exploited by the model as follows:

- Create two combined RIRs by adding and subtracting the RIRs for the left and right channel, respectively;
- Shift the temporal window along each of these combined RIRs in steps of 2 ms, starting with the peak at the direct sound (i.e., ignoring the part of the increasing ramp which preceded the direct sound);
- For each step, compute the sum of the sample-wise product of the squared combined RIRs and the temporal window;
- Determine the temporal position at which this sum is maximal for each of the combined RIRs (i.e., sum or difference between left and right ear). If the temporal position is not unique because the sum is maximal at several temporal positions, pick the earliest position;
- For each experiment, select one of these two temporal positions based on the IPD of the noise employed, that is, select the temporal position resulting from the difference-RIR for experiments using N_0, and the temporal position resulting from the sum-RIR for experiments using N_π.

It is obvious that this conceptual approach included a number of steps requiring oracle knowledge about the stimuli and listening conditions which are not accessible to the listener and cannot be used to enhance the model's applicability to practical applications (the implications and limitations for enhancing binaural model implementations are addressed in the general discussion). The motivation for testing this approach was to provide a model-based assessment of the degree of flexibility in spatio-temporal integration which might be required to account for the present data and to provide an estimate of the effective temporal integration duration under the assumption of a maximum degree of flexibility.

Results and Discussion

SRT Measurements

Integration of a single reflection. Experiments I to VII explored the spatio-temporal integration of a single reflection with respect to its delay, IPD, and amplitude relative to the direct sound. Experiment I served as a reference condition (diotic stimuli, single reflection with the same amplitude as the direct sound and varied delays) and essentially replicated the reference condition of Warzybok et al. (2013) for a subset of delays. Mean SRTs are shown as circles connected by solid lines in the top left panel of Figure 2. Error bars represent standard errors. As expected, SRTs increased with increasing reflection delay from −12.0 dB SNR (direct sound only) to −8.6 dB SNR (delays of 150 and 200 ms). An ANOVA confirmed that SRTs significantly depended on reflection delay, $F(1.857,\ 12.996) = 29.722$, $p < .001$. Paired post hoc comparisons showed that the SRT for the direct sound-only condition was significantly lower than in all other conditions except for a reflection delay of 10 ms. In addition, SRTs measured for reflections at delays of 10 and 50 ms were significantly lower than SRTs measured for reflections at the two longest delays (150 and 200 ms). This dependence on reflection delay was in line with previous data of Warzybok et al. (2013), who reported that early reflections (up to about 25 ms) could be perfectly integrated with the direct sound in their reference condition and that larger delays caused an SRT increase by about 4 dB. There was an overall shift of SRTs of the present data (mean direct sound-only SRT of −12.0 dB) compared with the data of Warzybok et al. (2013) (mean direct sound-only SRT of −7.7 dB). This was slightly larger than expected from the differences in reference SRT values for normal-hearing listeners between the American English matrix test (−10.0 dB) and the German matrix test (−7.1 dB, see Kollmeier et al., 2015). While the differences between matrix tests between languages likely reflect differences in talker intelligibility, the small additional difference observed in this study might be due to the different listener groups. Apart from this offset, the temporal integration of a single reflection seemed to be highly comparable between both studies, and the overall increase in SRT indicated that the energy of a very late reflection was no longer available for speech recognition (which would produce an SRT increase of 3 dB), and could even contribute a small additional detrimental effect due to its property of being an identical (intelligible) copy of the direct sound.

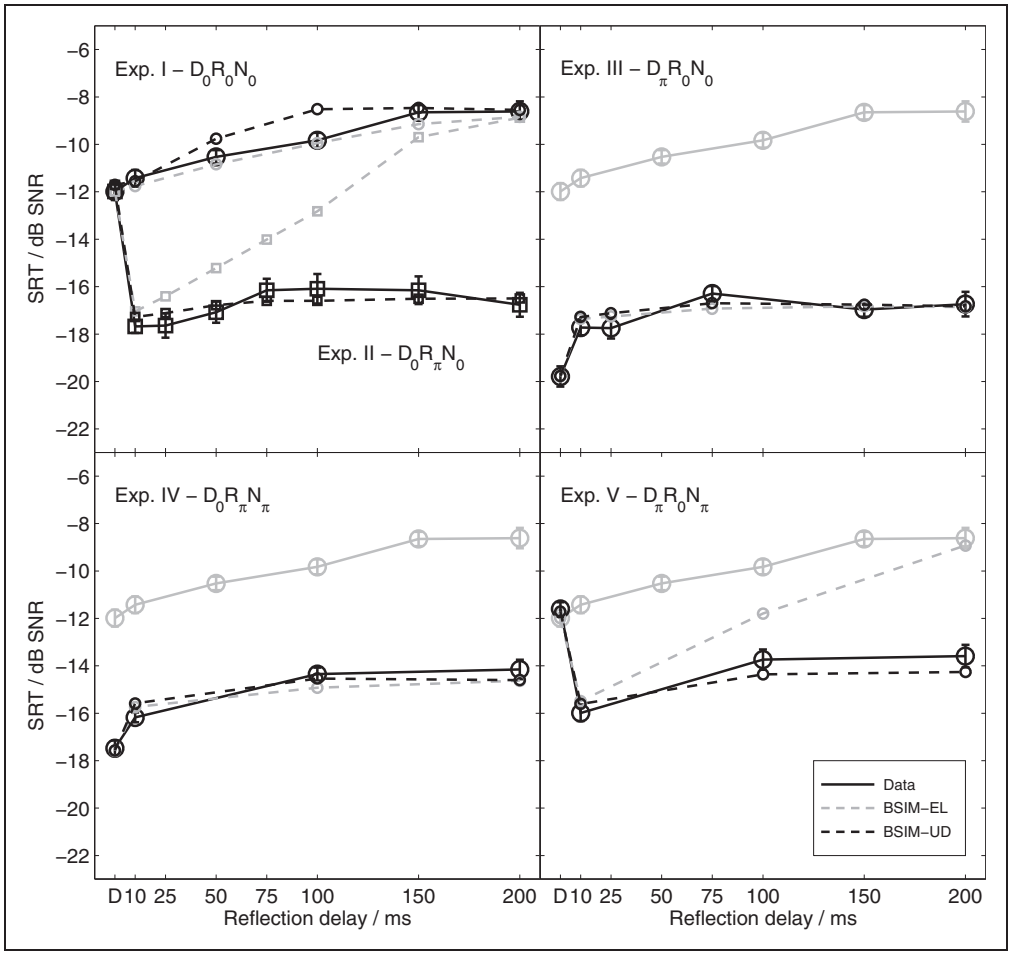

Figure 2. Mean SRTs across listeners (and standard errors) of Experiments I to V, which comprised a single reflection with varying delay. Denotes conditions with direct sound only. Dashed lines show predictions of BSIM-EL (gray) and BSIM-UD (black). Gray solid lines with circles represent data replotted from Experiment I. SRT = speech recognition threshold; BSIM = binaural speech intelligibility model; SNR = signal-to-noise ratio.

Experiment II explored the temporal integration of a single reflection (with $\alpha = 1$) for dichotic speech in which only the reflection carried an IPD-advantage (R_π) in the N_0 masker, while the direct sound did not (D_0). Mean SRTs are shown as squares connected by solid lines in the top left panel of Figure 2. The data differed markedly from SRTs of Experiment I (R_0): Adding an R_π-reflection to the direct sound resulted in a considerable SRT improvement by about 6 dB, and SRTs were quite similar (between −18 and −16 dB SNR) for all reflection delays, that is, they remained low even for long delays. An ANOVA confirmed a significant main effect of delay, $F(3.179,\ 22.252) = 23.986$, $p < .001$, but only because a delay of 0 ms (i.e., direct sound-only) was included in the analyses: Post hoc test showed that the SRT in this condition was significantly higher than in all other conditions, but that the conditions with a reflection delayed by 10 ms or more did not differ from each other. This result is interesting because it suggests that the reflection "takes over" the role of the primary source of information from the preceding direct sound regardless of its delay, while the direct sound (which has the same amplitude as the reflection) is ignored. An IPD-advantage thus seems to dominate speech intelligibility also when the level of the direct sound and reflection is the same and the reflection would produce an increase in SRTs, at each ear considered separately, as observed in Experiment I.

Experiments III to V were designed to follow up on these conditions by varying the IPD of the direct sound, the noise, and the reflection ($\alpha = 1$ in all cases). The corresponding SRTs are shown in the other panels of Figure 2. In Experiments III (top right) and IV (bottom left), the reflection had the same IPD as the noise, while the direct sound carried an IPD-advantage. In both experiments, SRTs were significantly lower than in the reference Experiment I. For the direct sound-only conditions, SRTs were −19.8 and −17.5 dB SNR for

Experiments III and IV, respectively, compared with −12.0 dB in Experiment I. This decrease reflects the well-known IPD-advantage for speech. The difference in IPD-advantage between the two experiments of about 2 dB is in line with previous studies reporting better intelligibility for $S_\pi N_0$-conditions than for $S_0 N_\pi$-conditions (e.g., Feldmann, 1963; Licklider, 1948). For longer reflection delays, SRTs in both experiments increased by about 2 to 3 dB. In both experiments, the main effect of reflection delay was significant, Experiment IV: $F(3.315,\ 23.202) = 18.531$, $p < .001$; Experiment V: $F(1.764,\ 12.346) = 34.791$, $p < .001$. Post hoc tests showed that, in Experiment III, SRTs in the direct sound-only condition were significantly lower than in all other conditions. In addition, SRTs for delays of 10 and 25 ms were significantly lower than for a delay of 75 ms. In Experiment IV, SRTs in the direct sound-only condition were significantly lower than in all other conditions, and the SRT for a delay of 10 ms was also significantly lower than for delays of 100 and 200 ms.

In contrast to these two experiments, BRIRs in Experiment V were such that the reflection carried an IPD-advantage, while the direct sound did not (similar to Experiment II, but with inverted IPDs). The SRTs are shown in the bottom right panel of Figure 2. SRTs in the direct sound-only condition were similar to those in Experiment II (−11.6 vs. −12.0 dB). SRTs decreased by about 4.5 dB when adding a reflection delayed by 10 ms. This decrease was again slightly smaller than observed for inverted IPDs in Experiment II. For longer delays, SRTs increased slightly by about 2 dB. These findings were supported by an ANOVA, which confirmed a significant influence of reflection delay, $F(2.319,\ 16.230) = 28.996$, $p < .001$. Post hoc tests showed that SRTs for direct sound-only were significantly higher than all other SRTs. In addition, the SRT for a delay of 10 ms was significantly lower than for delays of 100 and 200 ms. Unlike in Experiment II, the differences in SRTs across reflection delays were statistically significant, although the magnitude of the SRT increase between short and long delays was similar (about 2 dB). One possible reason is that Experiment II comprised more conditions and, hence, the significance level was reduced more using the Bonferroni corrections to avoid Type I errors in the analysis. In this light, the data of Experiments II and V may be considered comparable.

Experiment I confirmed that a reflection of the same amplitude as the direct sound and delayed by 200 ms could not be integrated with the direct sound, which resulted in increased SRTs. Experiment VI investigated how this depended on the relative amplitude of the reflection at a fixed delay of 200 ms. Solid lines and circles in Figure 3 show mean SRTs for α values from 0 (direct sound only) to 2.5. As a visual guide, the SRT for $\alpha = 0$ was copied to the other extreme of the scale ($\alpha = \text{inf.}$), because in both cases, the BRIR only consisted of a single component and the only difference was a temporal shift of the speech signal by 200 ms, which was assumed to be irrelevant in the stationary masker employed here. The data followed a pattern of approximately constant SRTs for small α values, followed by an increase for α values close to 1, and a decrease for larger α values. The ANOVA showed that SRTs depended significantly on α, $F(3.137,\ 21.958) = 32.7$, $p < .001$. Post hoc comparisons indicated that SRTs for α values of 0.75, 1.0, and 1.25 were significantly higher than for α values of 0, 0.25, and 2.5. In addition, SRTs were significantly different between $\alpha = 2.0$ and α of 0.25, 0.75, and 1.0. The other differences, including those between small α values (0.25, 0.5) and the largest α value of 2.5, were not significant. This pattern seems intuitive in that one of the two components dominates speech intelligibility when its amplitude is markedly larger than that of the other component, irrespective of the delay. When both components are similar in amplitude, an integration of both components is not possible and SRTs increase. This effect reached a maximum of about 3.5 dB for $\alpha = 1.0$, but the differences in the mean SRT between α values of 0.75, 1, and 1.25 were not statistically significant.

As the final experiment with a single reflection, Experiment VII extended Experiment VI and investigated the role of reflection amplification for a reflection

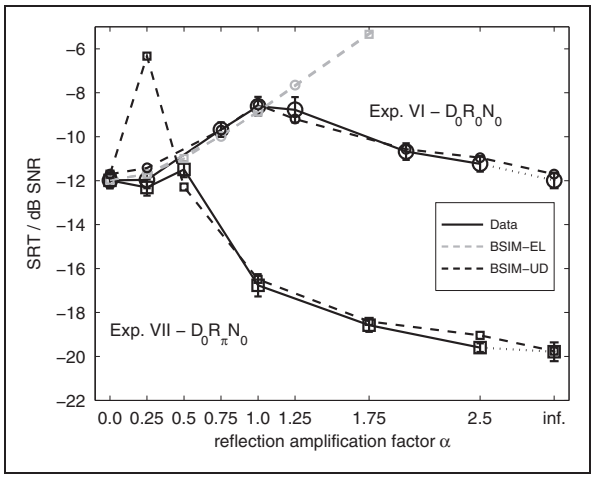

Figure 3. Mean SRTs across listeners (and standard errors) of Experiments VI (black circles) and VII (black squares), which comprised a single reflection delayed by 200 ms and varying reflection amplification factor α. Dashed lines show predictions of BSIM-EL (gray) and BSIM-UD (black). Note that predictions of BSIM-EL are almost identical for both experiments. "inf." indicates the condition consisting of the reflection only (data copied from $\alpha = 0$ for Experiment VI, and taken from the direct sound-only condition of Experiment III for Experiment VII). SRT = speech recognition threshold; BSIM = binaural speech intelligibility model; SNR = signal-to-noise ratio.

with a potentially detrimental delay of 200 ms. In Experiment VII, however, the reflection carried an IPD-advantage. Squares connected by solid lines in Figure 3 show the resulting SRTs, which differed considerably from Experiment VI (circles): For small reflection amplitudes (α of 0.25 and 0.50), SRTs were similar to the direct sound-only condition ($\alpha = 0$). For $\alpha \geqslant 1$, however, SRTs decreased considerably and approached the value measured in the $D_\pi N_0$-condition in Experiment III (included in Figure 3 as "$\alpha = \text{inf.}$"). This was supported by an ANOVA, which showed that the main effect of α was significant, $F(3.009,\ 21.066) = 121.002$, $p < .001$. Post hoc tests showed that SRTs for the three lowest α values (0, 0.25, 0.5) were significantly higher than for higher α values. In addition, the SRT for $\alpha = 1.0$ was significantly higher than for $\alpha = 2.5$ ($\alpha = \text{inf.}$ was not included in the ANOVA). This SRT pattern further supports the notion discussed earlier that the reflection became the dominant component for speech recognition regardless of its long delay. For α values of 1 and larger, the differences in SRTs between Experiments VI and VII were very similar (about 8 dB), indicating that the binaural advantage resulting from the reflection IPD and the energetic differences between direct sound and reflection affected SRTs independently from each other.

Integration of multiple reflections with the same IPD. In Experiments VIII to XI, SRTs for conditions with more than one reflection were measured where all reflections had the same IPD, and reflections were successively added starting from the shortest delay of 10 ms. In other words, the speech energy was spread across more and more components as more reflections were added, and the temporal window across which the target energy was spread increased. In Experiment VIII, diotic stimuli were used, that is, direct sound, reflections, and masking noise all had an IPD of 0. The resulting SRTs are shown as circles connected by a solid line in the top left panel of Figure 4. SRTs were similar for a small number of reflections and then gradually increased as more (and later) reflections were included. When all nine reflections were included, SRTs were about 6 dB higher than in the direct sound-only condition. This was supported by an ANOVA, which showed that the effect of the number of reflections was significant, $F(2.789,\ 19.523) = 37.538$, $p < .001$. Post hoc tests showed that SRTs for zero, one, two, and three reflections did not significantly differ from each other but were lower than SRTs for five or more reflections. In addition, SRTs for five reflections were significantly lower than for all nine reflections. This suggests that the auditory system is capable of perfectly integrating more than one early reflection, and that the temporal window for perfect integration is similar in duration as found for a single reflection in earlier studies (e.g., Nábělek & Robinette, 1978; Warzybok et al., 2013), which is in line with data of Bradley et al. (2003). When reflections with delays outside this integration window were added, integration was no longer perfect and SRTs increased.

In Experiments IX to XI, the same temporal structure of BRIRs was used (i.e., reflections were added starting at short delays), but the IPDs of direct sound and reflections were varied (the masker was always diotic). In each case, the main effect of the number of reflections was significant, IX: $F(2.023,\ 14.222) = 19.417$, $p < .001$; X: $F(3.632,\ 25.421) = 71.666$, $p < .001$; XI: $F(2.208,\ 6.153) = 15.121$, $p < .001$, and the post hoc comparisons are reported in the following. The top right panel of Figure 4 shows SRTs for conditions in which the reflections carried an IPD-advantage (R_π) in an N_0-masker, but the direct sound did not (D_0). The data for direct sound-only and a single reflection delayed by 10 ms were similar as reported for Experiment II (which used partly different listeners) and indicated a considerable SRT decrease when a reflection with IPD-advantage was added. When more reflections were added, SRTs first remained relatively constant and then increased when reflections at very long delays were included. For the maximum number of reflections, SRTs were about 1.5 dB below SRTs in the direct sound-only condition, but this difference was not significant according to post hoc tests. In contrast, SRTs for one, three, five, and seven reflections were significantly lower than the SRT in the reference condition but not statistically different from each other. In addition, the SRT for nine reflections was significantly higher than for one and three reflections. The fact that SRTs were statistically equivalent for BRIRs containing a single reflection with IPD-advantage and BRIRs with 7 R_π-reflections spread over delays from 10 to 150 ms suggests that the auditory system not only exploits a single reflection carrying an IPD-advantage but integrates such reflections over a relatively long temporal window. Only when the temporal distance between reflections became too long (last data point for nine reflections), the SRT increased.

The bottom left panel of Figure 4 shows SRTs measured in Experiment X, that is, for the same temporal BRIR configuration but with the IPD-advantage confined to the direct sound ($D_\pi R_0 N_0$). In contrast to Experiment IX, SRTs tended to increase monotonically with increasing number of reflections compared with the direct sound-only condition. The SRT for the direct sound-only condition was significantly lower than for all conditions including three or more reflections but not statistically different from the SRT for a single reflection. In addition, the following pairs of SRTs differed significantly from each other: one versus five, seven, and nine reflections; three versus seven and nine reflections; and five versus nine reflections. This SRT pattern seems plausible when considering the relative energy of

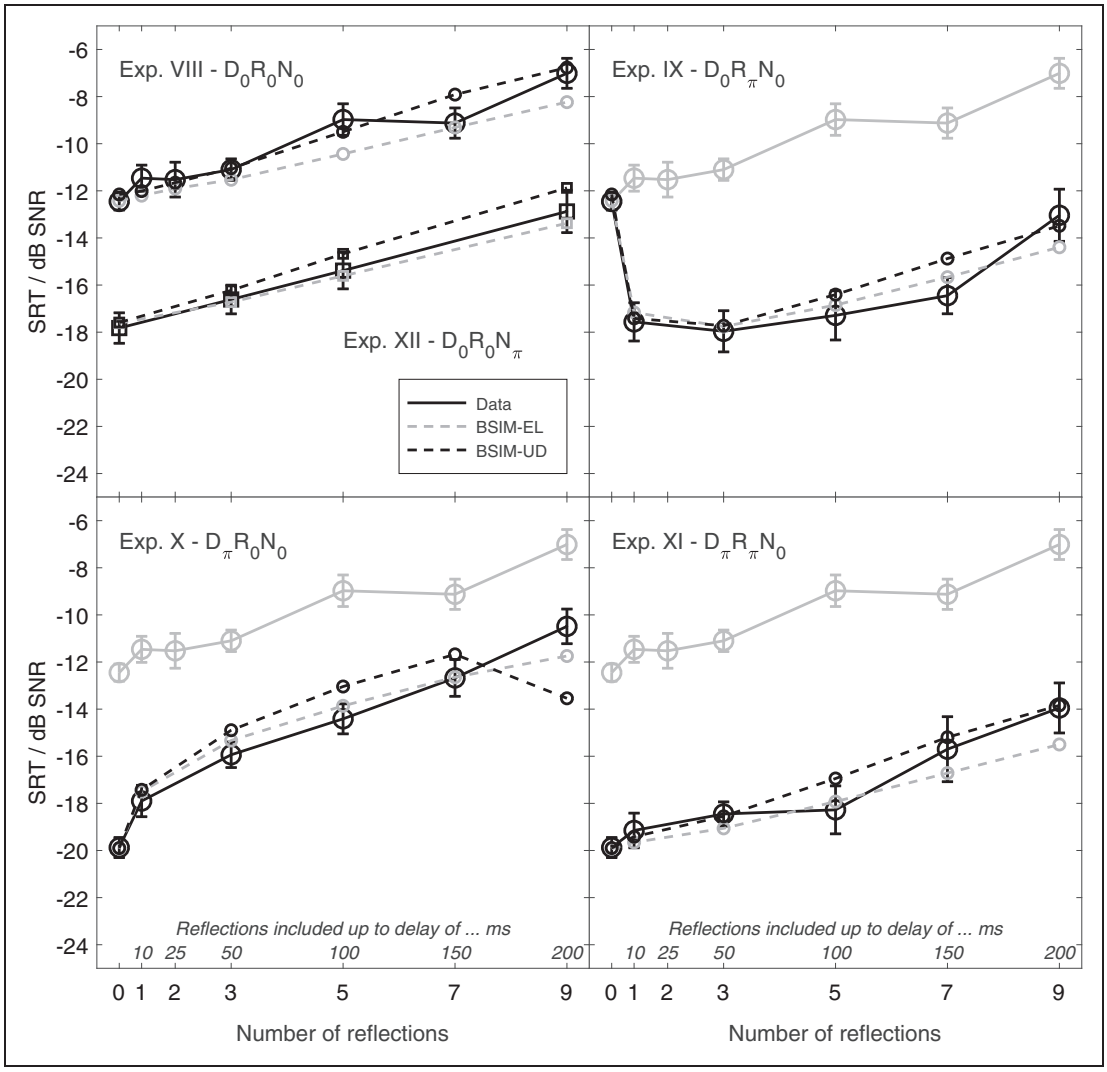

Figure 4. Mean SRTs across listeners (and standard errors) of Experiments VIII, IX, X, XI (black circles), and XII (black squares), in which reflections were successively added starting from the shortest delay. The number of reflections is given at the abscissas. For comparison, data of Experiment VIII are replotted in each panel in gray (solid line). Dashed lines show predictions of BSIM-EL (gray) and BSIM-UD (black). Gray solid lines with circles represent data replotted from Experiment VIII. SRT = speech recognition threshold; BSIM = binaural speech intelligibility model; SNR = signal-to-noise ratio.

the BRIR components. SRTs were lowest in the condition in which the entire speech energy was in the component carrying the IPD-advantage (i.e., the direct sound). As more and more components without IPD-advantage were added (and the energy of the direct sound decreased due to the level normalization), SRTs increased correspondingly. At the largest number of reflections, SRTs were about 9 dB higher, which is close to the energetic reduction of the direct sound when adding nine reflections to restore the overall sound pressure level (see Table 2).

The bottom right panel of Figure 4 shows SRTs of Experiment XI in which both direct sound and reflections carried and IPD-advantage ($D_\pi R_\pi N_0$). The data pattern showed that SRTs were approximately constant for conditions with direct sound only and lower numbers of reflections and then increased for seven and nine reflections. Paired comparisons indicated that the only significant differences were between the SRT for nine reflections and SRTs for zero, one, three, and five reflections. In general, this pattern of initially constant and then increasing SRTs was qualitatively similar to the data of Experiment VIII ($D_0R_0N_0$), although the significance of the individual paired comparisons differed. The main difference between both experiments was that SRTs were shifted downward by about 7 to 9 dB due to the IPD-advantage, while the dependence on the number of reflections was similar in both experiments. This supports the conclusions of Warzybok et al. (2013) and Arweiler and Buchholz (2011) that binaural and temporal processing

operate independently for "colocated" direct sound and reflections, that is, when direct sound and reflections comprise the same binaural information.

Experiment XII ($D_0R_0N_\pi$) was included to further validate this finding. The corresponding SRTs are shown as squares in the top left panel of Figure 4. It is evident that SRTs followed the same dependence on the number of reflections as in N_0-noise, that is, the binaural unmasking of 7 to 8 dB was independent from the temporal integration. This was confirmed by a two-way ANOVA (conducted for the subset of zero, three, five, and nine reflections measured in both experiments), which showed no significant interaction between the number of reflections and the noise IPD, $F(2.163, 15.141) = 0.987$, $p = .401$.

Experiments XIII to XVI (Figure 5) also explored the influence of varying the number of reflections in the presence of an N_0 masker, but in contrast to the previous experiments, reflections were successively added starting from the longest delay. As before the main effect of the number of reflections was significant for all experiments, XIII: $F(2.594, 18.159) = 19.506$, $p < .001$; XIV: $F(3.207, 22.450) = 13.185$, $p < .001$; XV: $F(3.551, 24.860) = 73.273$, $p < .001$; XVI: $F(3.047, 21.328) = 11.118$, $p < .001$, and the results of the post hoc comparisons are reported in the following.

In Experiment XIII ($D_0R_0N_0$, top left panel), SRTs increased significantly by about 5 to 6 dB relative to the direct sound-only condition for all conditions with three or more reflections (which showed no statistically

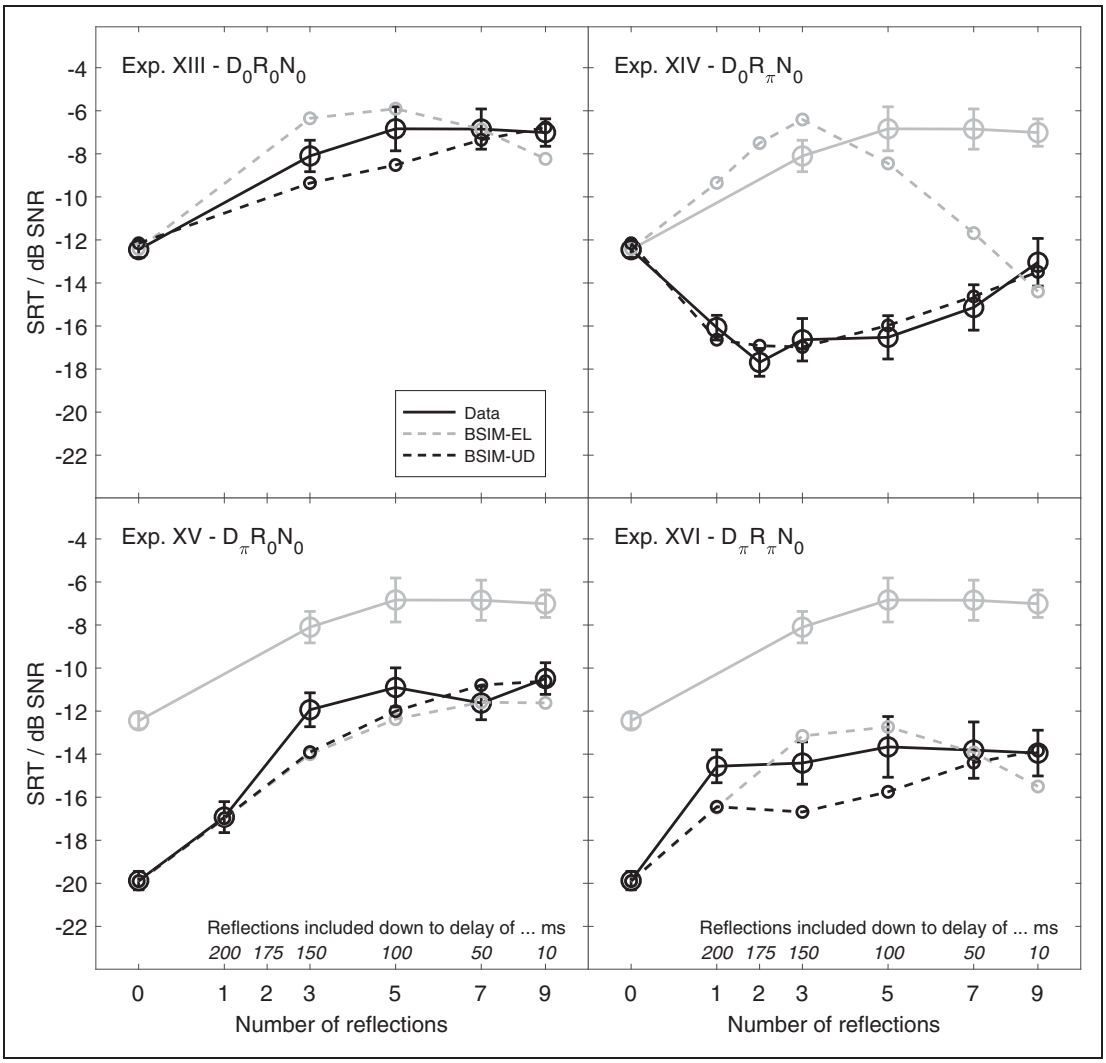

Figure 5. Mean SRTs across listeners (and standard errors) of Experiments XIII to XVI (black) in which reflections were successively added starting from the longest delay. Dashed lines show predictions of BSIM-EL (gray) and BSIM-UD (black). Gray solid lines with circles represent data replotted from Experiment XIII. SRT = speech recognition threshold; BSIM = binaural speech intelligibility model; SNR = signal-to-noise ratio.

significant differences). This pattern is interesting: When adding a single late reflection to the direct sound, SRTs were shown to increase by about 4.5 dB in Experiment I. SRTs of this experiment suggest that SRTs do not increase further when more target energy is put into multiple late reflections at the cost of direct sound-energy: SRTs for three reflections (at 200, 175, and 150 ms) were also about 4.5 dB higher than for direct sound only. This suggests that direct sound energy does not drive SRTs in these conditions (which would lead to a further increase in SRTs). One possible reason for the observed SRT pattern could be that the "center of temporal attention", that is, the temporal position of the BRIR which is most relevant for speech recognition, shifts from the direct sound to the later part of the BRIR. These late components (in combination) comprise more speech energy than the direct sound and are temporally separated by only 50 ms, while the direct sounds precedes these components by 150 ms, which is longer than the (perfect) temporal integration window as observed in previous experiments and studies. Along this line, the fact that SRTs remained constant when adding further (and increasingly early) reflections may then be a result of the early components "recapturing" the role of the dominant speech components but at the cost of losing the late component for the temporal integration so that, in sum, SRTs remain constant.

SRTs of Experiment XIV ($D_0R_\pi N_0$) are shown in the top right panel. Here, SRTs decreased when adding late reflections with an IPD-advantage relative to the D_0N_0-condition. This decrease was about 4.5 dB when adding a single reflection with a delay of 200 ms (compared with 5 dB observed in Experiment VII for the same conditions) and was then approximately constant for two to five reflections. For seven and nine reflections, SRTs increased slightly. This was confirmed by post hoc comparisons, which indicated that SRTs for the direct sound-only condition were significantly higher than for conditions with one, two, three, and five reflections, but not higher than for seven and nine reflections. In addition, the SRT for nine reflections was significantly higher than for two and three reflections. This SRT pattern suggests that it is beneficial to have one or more BRIR components with an IPD-advantage even when they are considerably delayed relative to the direct sound, and that this benefit decreases when the components are spread across a longer range of delays, which is in line with the data of Experiment IX discussed earlier.

SRTs of Experiments XV and XVI are shown in the bottom panels of Figure 5. When only the direct sound carried an IPD-advantage (Experiment XV, bottom left), SRTs increased significantly when adding one or three late reflections. This can be understood in energetic terms since the direct sound (i.e., the only component with IPD-advantage) comprised increasingly less energy. For larger numbers of reflections, SRTs were constant at about −12 to −10 dB SNR and did not differ statistically from the condition with three reflections. It is interesting to observe that SRTs for three and more reflections were always about 4 dB lower than SRTs in Experiment XIII (replotted on gray). The only difference between these experiments was the IPD of the direct sound. It seems that the π-IPD of the direct sound component in Experiment XV still provided a benefit even when its energy was considerably reduced because the speech energy was distributed across multiple BRIR components. This residual IPD-benefit was about half as large as the maximum IPD-advantage observed when comparing the direct sound-only conditions.

When both direct sound and reflections carried an IPD-advantage (Experiment XVI, bottom right panel), the SRT pattern was very similar to that in the $D_0R_0N_0$-condition (Experiment XIII), except that SRTs were constantly lower by about 7 to 8 dB. A two-way repeated measures ANOVA conducted for the subset of zero, three, five, seven, and nine reflections confirmed that the interaction of *experiment* (XIII and XVI) and *number of reflections* was not significant, $F(2.749, 19.246) = 0.248$, $p = .846$. These data further support the previous finding that temporal processing and binaural processing are independent when direct sound and reflections stem from the same (virtual) position in space. Post hoc tests for Experiment XVI showed no significant differences between SRTs for BRIRs with one to nine reflections, while all SRTs were significantly higher than for zero reflections. The initial SRT increase of about 5 dB when adding a single, late reflection was in line with the previous conclusion that a reflection delayed by 200 ms cannot be integrated with the direct sound and can even be detrimental (increase > 3 dB). As in Experiment XIII, it is interesting to observe that SRTs did not increase further when additionally adding reflections at 175 and 150 ms. Such an increase may have been expected based on energetic considerations, since the direct sound component contained increasingly lower energy. Hence, these data further support the idea of a shifting "focus of temporal attention" toward later components for conditions in which this is energetically beneficial.

SRT Predictions

Validating the concept of early/late separation. Gray dashed lines in Figures 2 to 5 show predictions of BSIM-EL with $t_e = 100$ ms and DD = 200 ms as proposed by Rennies et al. (2014). With respect to conditions including a single reflection, predicted SRTs were in quantitative agreement with the data for Experiments I, III, and IV. For Experiment I, this was expected because the model

had already been shown to predict the data in the same reference condition of Warzybok et al. (2013). The good agreement in Experiments III and IV shows that the IPD-advantage was correctly predicted (after adjusting the ITD processing error, see section "SRT prediction"). Similarly, the model predicted the observed inability to integrate even an early reflection when its IPD differs from the direct sound and is the same as the noise-IPD. In contrast, there were considerable deviations between data and predictions for Experiments II, V, VI, and VII. As expected, BSIM-EL failed to predict that a late reflection dominates speech intelligibility when its energy is much larger than that of the direct sound (Experiment VI), because the reflection ($\Delta t = 200$ ms) was outside the temporal window extracting the early components. Accordingly, predictions of BSIM-EL were the same whether the late reflection had an IPD of 0 (Experiment VI) or π (Experiment VII, lines overlap in Figure 3). BSIM-EL also failed to predict the observation that SRTs were independent of reflection delay when the reflection carried the IPD-advantage but the direct sound did not (Experiments II and V). Instead, the model predicted increasing SRTs because the advantageous reflection gradually moved out of the early window as its delay increased.

In line with these general findings, predicted SRTs were in good agreement with the data of Experiments VIII to XI, where BRIRs comprised multiple reflections which were added starting from the shortest delay (see gray dashed lines in Figure 4). This confirmed that BSIM-EL correctly predicted the independence of temporal and spatial processing (e.g., Experiment VIII vs. XI vs. XII) when direct sound and reflections had the same binaural information. The model also predicted the trends in Experiments IX and X, which can probably be attributed to an interaction of the temporal configuration of reflections and their relative energy (which decreases with increasing number of reflections) as discussed earlier.

In contrast, not all trends were correctly predicted when the reflections were successively added starting at long delays (Experiments XIII to XVI, see gray dashed lines in Figure 5). For diotic stimuli (top left panel), the SRT increase when adding three late reflections to the direct sound was slightly overestimated by the model. This could support the idea that the auditory system focuses on these late reflections rather than on the direct sound (because this is energetically beneficial), while BSIM-EL focuses on the direct sound and only considers a considerably attenuated portion of the late reflections as useful. The same trend was also observed for three reflections in Experiment XVI (bottom right panel). Altogether, the magnitude of the deviations was rather small for Experiments XII, XV, and XVI, although the SRT patterns as a function of reflection number differed somewhat from the experimental data as noted earlier. However, for Experiment XIV, large deviations occurred as BSIM-EL predicted a tent-shaped SRT pattern, while the data followed a V-shaped pattern. This also supports the notion discussed earlier that the listeners appeared to be able to focus on the late reflections because they carried the binaurally beneficial information. This was not possible for BSIM-EL because the late components were considered as only partially useful and were strongly attenuated by the early window. Only as the number of reflections was increased to 9, that is, when a considerable number of early reflections carried the IPD-advantage, model predictions were again similar to the data.

In summary, BSIM-EL could not predict the observed trends for conditions in which the components carrying the binaurally beneficial information were outside the early window but provided good predictions in all conditions in which the binaurally advantageous components were within the temporal window extracting the early (and assumed useful) components. It appeared that the parameters of $t_e = 100$ ms and DD = 200 ms as fitted to the data of Warzybok et al. (2013) were suitable also for the new data in these conditions, that is, they generalized to another language and group of listeners. The overall prediction accuracy of BSIM-EL with $t_e = 100$ ms and DD = 200 ms including all 20 experiments is summarized in the left panel of Figure 6, which shows the scatter plot of measured against predicted SRTs. Most data points clustered around the diagonal representing perfect agreement between data and predictions, but there were significant outliers, the most notable of which originated from Experiment VII where the late reflection was much higher in energy than the direct sound and additionally carried an IPD-advantage but was outside the early window and hence considered detrimental. Correspondingly, the resulting mean absolute prediction error ϵ_{mean} as well as the maximum absolute error (ϵ_{max}), the RMSE, and the coefficient of determination (R^2) indicated a rather poor overall prediction accuracy when including all conditions tested in this study.

To investigate the prediction accuracy for other combinations of t_e and DD, these parameters were systematically varied as described in section "SRT prediction." The left panels of Figure 7 illustrate ϵ_{mean} for Part A (top, experiments with a single reflection), Part B (middle, experiments with several reflections), and the combined experiments of both parts (bottom). Note that, for this analysis, Experiments II, V, VI, VII, XIV, and XVII were *not* included. This was done to avoid the prediction error being dominated by conditions in which an early/late separation is obvious to fail because the late reflections (beyond t_e) dominate speech perception and, hence, the comparability to previous studies would be

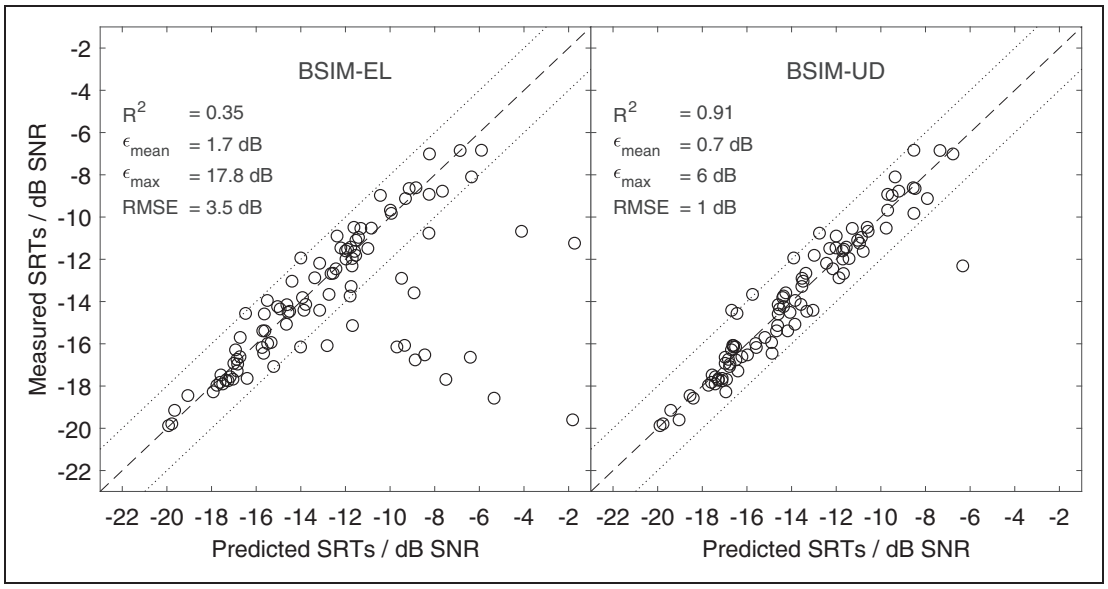

Figure 6. Scatter plot of measured versus predicted SRTs for BSIM-EL ($t_e = 100$ ms, DD = 200 ms, left panel) and BSIM-UD (RD = 100 ms, right panel) for all 20 experiments. The dashed and dotted lines represent perfect agreement and deviations of ±2 dB, respectively. SRT = speech recognition threshold; BSIM = binaural speech intelligibility model; SNR = signal-to-noise ratio; RMSE = root-mean-square error.

difficult. In each panel, the prediction error is shown as contour plot, where the magnitude of ϵ_{mean} (in dB) is given as labels at the contours. For (the reduced) Part A, the error pattern was generally such that the error was larger at very short and very long t_e (ordinate) and the minimum roughly occurred between 60 and 100 ms. For each value of t_e (except for small values), the error decreased with increasing DD. The smallest prediction error of about 0.4 dB was found for an area between about 60 ms $\leqslant t_e \leqslant$ 100 ms and DD $\geqslant 1.5 \cdot t_e$. In other words, the best parameter combination was not very sharply defined and included the values of $t_e = 100$ ms and DD = 200 ms for which predictions are shown in Figures 2 and 3. For (the reduced) Part B, the magnitude of ϵ_{mean} was generally larger and followed a similar pattern (minimum for intermediate t_e, decreasing error with increasing DD). Best predictions ($\epsilon_{\text{mean}} < 0.9$ dB) were observed in a narrow area around $t_e = 100$ ms and DD = 200 ms. This pattern was very similar when considering all experiments except those with intelligibility-dominating late reflections (bottom panel), where the smallest prediction error of $\epsilon_{\text{mean}} < 0.8$ dB was observed in the same area.

Altogether, this parameter analysis is in line with our previous studies investigating BRIRs without intelligibility-dominating late reflections (Rennies et al., 2011, 2014) and supports the finding that the effective spatio-temporal integration window can be characterized by an early/late limit of about 100 ms and a fade out of about 200 ms. It should be noted that optimal values for early/late limits and temporal window length have been shown to differ somewhat across listening rooms (Leclère et al., 2015).

Flexible temporal integration window. The SRT predictions of BSIM-UD, that is, the conceptual approach with temporally flexible window to separate useful and detrimental BRIR components are shown as black dashed lines in Figures 2 to 5. These predictions were made with RD = 100 ms (the role of RD on prediction accuracy is discussed later). For a single reflection varied in delay (Figure 2), predictions closely matched the data except for a slight overestimation of SRTs in Experiment I for intermediate delays. Specifically, BSIM-UD also predicted the quasi-independence of SRTs from reflection delay when the reflections carried an IPD-advantage, but the direct sound did not (Experiments II and V). Similarly, BSIM-UD also quantitatively predicted SRTs measured when the level of a late reflection ($\Delta t = 200$ ms) was varied (Experiments VI and VII, see Figure 3). The only exception was observed for a reflection amplification factor of $\alpha = 0.25$ in Experiment VII, that is, when the reflection IPD favored the late reflection over the direct sound, but the amplitude did not. This discrepancy indicates that the applied simple maximum selection method is oversimplified for some combinations of reflection delay, IPD and amplitude. For $\alpha = 0.5$, this simplified maximum selection also favors the late reflection, and the fact that predictions were close to the measured data indicates that the binaural advantage gained from this reflection compensated for the reduced energy of the reflection.

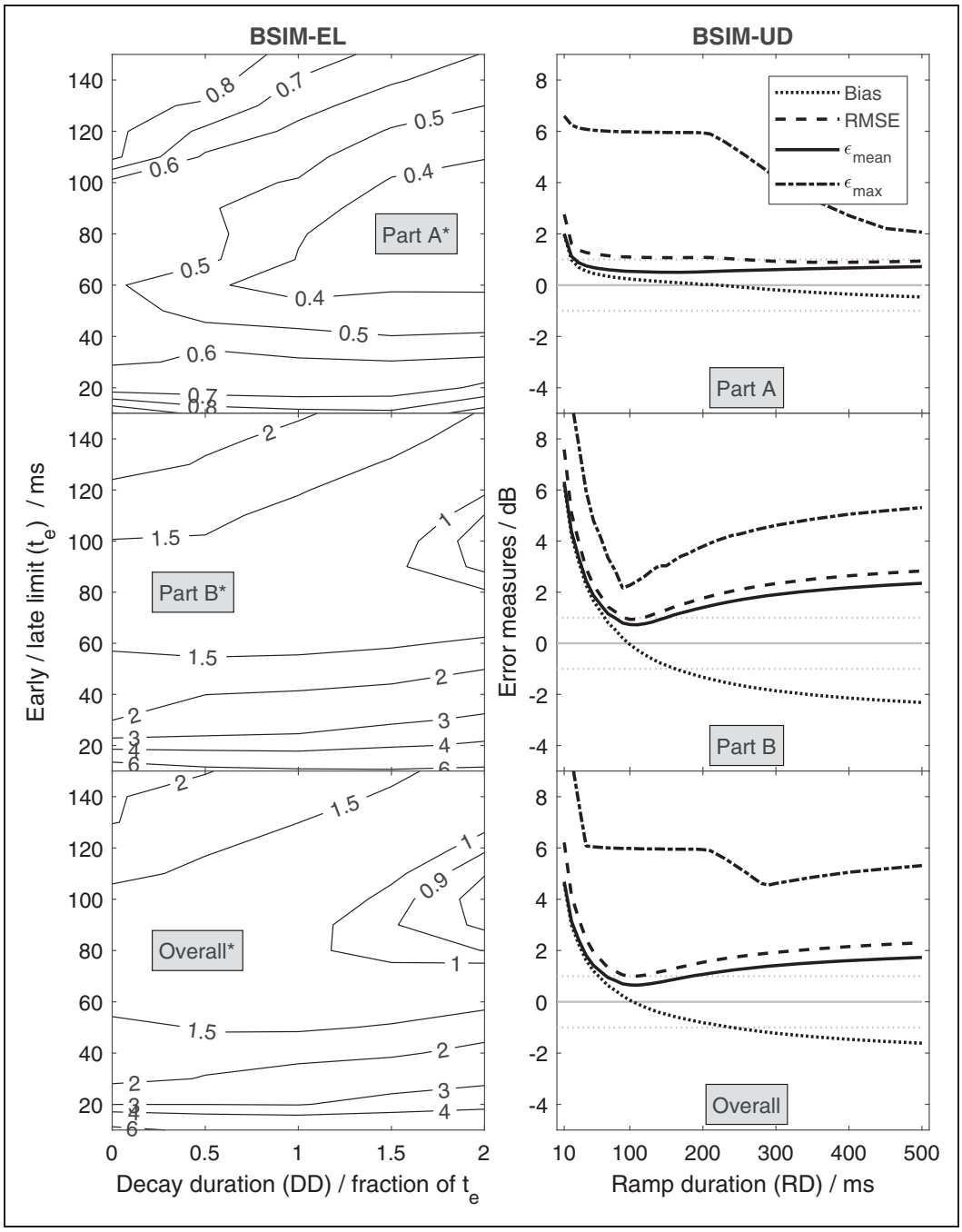

Figure 7. Left panels: Contour plots of the mean absolute prediction error ϵ_{mean} of BSIM-EL for different combinations of early/late limit (t_e, ordinates) and DD (abscissae) for Part A (top), B (middle), and both parts combined. Values at the contours indicate the magnitude of ϵ_{mean} in dB. The asterisk indicates that not all experiments were included in this analysis (see text). Right panels: different prediction error measures (in dB) for BSIM-UD as a function of the RD. Gray horizontal lines indicate optimal performance (solid) and deviations by ± 1 dB (dotted). BSIM = binaural speech intelligibility model; RMSE = root-mean-square error.

For conditions with increasing number of reflections (Figures 4 and 5), most predictions were also accurate. Notably, this also included Experiment XIV, where large deviations were observed for BSIM-EL, which can be attributed to the capability of BSIM-UD to focus on the late components rather than on the direct sound and exploit their IPD-advantage. Among these experiments, two further comparisons appeared worth looking at more closely: First, the SRT for the last condition of Experiment X was underestimated by the model, while the other SRTs tended to be slightly overestimated. A second underestimation of the measured SRTs was

observed in Experiment XVI when adding a cluster of late reflections to the direct sound. This might suggest that the implemented fully flexible temporal window is oversimplified and that, for example, the ''switch of temporal attention'' to later components of the BRIR cannot be made without a certain cost in all conditions.

The overall prediction accuracy of BSIM-UD is illustrated as scatter plot in the right panel of Figure 6. With a single exception (observed in Experiment VII as discussed earlier), all predicted SRTs were within 2 dB of the measured data. Accordingly, all error measures (ϵ_{mean} and RMSE $\leqslant 1$ dB, $R^2 = .91$) were considerably improved compared with BSIM-EL and were comparable to previous evaluations of BSIM (Beutelmann & Brand, 2006; Beutelmann et al., 2010) and BSIM-EL (Rennies et al., 2011, 2014) for conditions in which the potentially useful contribution of late reflection was not as pronounced as in some conditions of this study.

The dependence of prediction accuracy of BSIM-UD on RD is illustrated in the right panels of Figure 7. In addition to the error measures employed earlier, the prediction bias is also shown (dotted line). It was computed as the y-intercept of a linear fit to the scatter plot of experimental data against model predictions with unity slope and can be interpreted as a general prediction offset (negative or positive values indicate generally under- or overestimated SRTs). Note that all 20 experiments were included in this analysis. For Part A (top), prediction errors were largest for very short RD and then decreased. This decrease resulted in a very broad minimum of ϵ_{mean} around 100 ms, but in general the error values were small (<1 dB) over a large range. At very long RD, RMSE and ϵ_{mean} approached an asymptote at around 1 dB, which corresponds to the prediction accuracy of the original BSIM without useful or detrimental separation. For Part B (middle), a different error pattern was observed: ϵ_{mean}, ϵ_{max}, and RMSE decreased to a relatively sharp minimum around RD = 100 ms and increased gradually for longer RD. The bias was positive for very short RD (indicating that SRTs were overestimated when the model could not integrate over several reflections), then crossed 0 dB around RD = 100 ms and decreased for longer RD, indicating that SRTs were underestimated when the model integrated reflections over a too long temporal window.

General Discussion and Conclusions

The main conclusions that can be drawn from this study are:

1. Direct sound and one or several speech reflections can be perfectly integrated when they have the same IPD (Experiments I and VIII). In such conditions, temporal processing and spatial processing are independent from each other (Experiments VIII, XI, XII).
2. In contrast, a single or small number of early reflections with the same IPD as the noise (but not as the direct sound) cannot be perfectly integrated with the direct sound even at short reflection delays (Experiments III, IV, X).
3. All conditions in which the dominant speech information is within the direct sound—or the early reflections—can be well predicted by the established approach to separate the BRIR into an early (assumed useful) and a late (assumed detrimental) part.
4. When energy (Experiments VI, VII, XIII, XVI) or IPD (II, V, VII, IX, XIV, XVII) make late components of the BRIRs a more dominant cue, the auditory system appears to be capable of focusing on these components rather than on the precedent direct sound and the subsequent early components. This cannot be modeled with the classic approach which assumes that early reflections of the BRIRs are useful and late reflections are detrimental. Instead a temporal integration window is required that can be flexibly shifted along the BRIR.

The first two conclusions contribute to the existing knowledge about spatio-temporal integration as discussed earlier. The third conclusion is in line with results of previous studies with more realistic RIRs, which found that the temporal integration window for binaural speech intelligibility can be characterized by an early/late limit and a rather shallow fade-out/fade-in. While the present data could be well modelled by an early/late limit of about 100 ms and an RD of 200 ms, these values may be different in real rooms and may even be room-dependent (Leclère et al., 2015). The data underlying the fourth conclusion provide direct evidence that the binaural auditory system is capable to prioritize later components of BRIRs over the direct sound or early reflections when this is beneficial. The benefit can be energetic in nature (when the normally detrimental late reflection has considerably higher energy) or can result from an advantage in IPD confined to the late component(s). Especially the second aspect highlights that this flexibility is not limited to picking the strongest component observed at either ear but involves a comparison between ears to select which components provide the benefit in a given masking condition. In a simplified approach, such a flexible effective spatio-temporal integration was tested within the framework of BSIM. The analysis showed that, basically, all trends observed in the data could be predicted by using an integration window that was applied with a maximum degree of flexibility (the peak could theoretically be anywhere along the BRIR) and a significant amount of oracle knowledge

(positioning the temporal window based on the known noise IPD). A systematic parameter variation showed that the duration of the flexible window for BSIM-UD that best matched the experimental data (RD of about 100 ms for both the increasing and the decreasing ramp) was comparable to the overall duration of the best-matching early/late separation window in BSIM-EL (DD = 200 ms). While this study does not reveal insights into the physiological mechanisms underlying the spatio-temporal integration of speech components, it highlights the degree of flexibility that is required to effectively model the observed data.

Some degree of flexibility in temporally integrating speech reflections was also proposed by Leclère et al. (2015), who found that binaural SRTs in rooms with different reverberation could not be well predicted by a model assuming a fixed early/late separation limit. Instead, they found that predictions were improved when adjusting the early/late limit for each room. It is unclear how the high degree of flexibility observed in this study is related to a room-dependent optimal early/late separation, but it would be interesting to extend the present results to RIRs including late reverberation. With respect to the particular capability of the auditory system to focus on late components, the observed temporal flexibility is probably limited to specific listening scenarios in practice. One example mentioned in the introduction is to listen to amplified speech while still hearing the direct sound (which could result in echoes with very high relative energy). In theory, this could also include differences in physical location of echo and direct sound and, hence, differences in the binaural information. However, playback in real rooms will always include some degree of coloration and head-shadow effects, which limit the relative role of IPD (which was maximized in the artificial stimuli of this study). It is thus questionable if an attempt to implement the flexible temporal integration approach as tested here will benefit models designed for practical applications, especially since the current implementation would considerably increase the amount of required a priori knowledge (binaural relation between masker and BRIR components). It therefore seems that the classic approach of early/late separation, which was shown to again produce highly accurate predictions (with the exceptions discussed earlier), is more suitable for most practical applications, especially when the potential issue of the room-dependency of this separation has been addressed (e.g., Kokabi, Brinkmann, & Weinzierl, 2018).

Another aspect related to the model's applicability is the adaptation of the processing parameters. In this study, the model parameter limiting the spatial benefit related to the ITD processing in the EC stage were slightly adjusted to account for the 2- to 2.5-dB smaller benefit observed in the experimental data. This was done to ensure that deviations between predictions and data could be interpreted as resulting from a miss in the employed interaction of spatio-temporal processing rather than from a general model offset. The original parameter had been adapted by vom Hövel (1984) from the concept of Durlach (1963) to better predict pure-tone binaural masking level differences at 500 Hz (Langford & Jeffress, 1964). Beutelmann and Brand (2006) pointed out the critical role of the processing errors to avoid severe underestimations of binaural SRTs. It is remarkable to observe that previous evaluations of BSIM (Beutelmann & Brand, 2006; Beutelmann, Brand, & Kollmeier, 2009; Beutelmann et al., 2010; Hauth & Brand, 2018; Rennies et al., 2011, 2014) did not indicate any need to adjust the parameters originally employed by vom Hövel (1984) to achieve good prediction accuracy for normal-hearing listeners. While some previous studies indicated a trend for a slight overestimation of the spatial benefit (see, e.g., Figure 2 in Beutelmann & Brand, 2006), the deviations were usually smaller than the 2.5 dB observed in this study. The reasons for this are not clear but may be related to the different speech material used in this study (American English matrix test) compared with the German matrix test employed in the previous studies. One notable difference is a considerably lower reference SRT in the American matrix test as discussed earlier (−12.0 dB vs. −7.7 dB). Hochmuth, Kollmeier, and Shinn-Cunningham (2018) showed that the female talker used as target in this study was considerably better intelligible than other talkers uttering the same speech material in the same masking noise, while the original talker of the German matrix test was within the medium range of other German talkers. It is possible that such a low SRT already in the reference condition (without any spatial cues) leaves a somewhat smaller room for unmasking benefit. This would be supported by the fact that the correction of $\sigma_{\delta 0}$ applied in this study was slightly larger for the listeners of Part B, which had a slightly lower SRT (by 0.5 dB on average) in the reference condition than listeners of Part A (correction factor 1.8 vs. 1.6). Another possible reason could be the different language (i.e., a general difference in binaural unmasking beyond speaker-specific factors), although SRT-differences between talkers within one language were found to be considerably larger than general language differences when using highly comparable speech material (Hochmuth et al., 2015). The role of talker effects in binaural unmasking may therefore be interesting to investigate in more detail.

Acknowledgments

The authors thank HörTech gGmbH for providing the speech material.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Grant No. RE 4160/1-1 as well as Projektnummer 352015383 - SFB 1330 A1, and by the National Institutes of Health-National Institute on Deafness and other Communication Disorders Grant No. R01 DC04545.

Supplemental Material

Supplemental material for this article is available online.

ORCID iD

Jan Rennies https://orcid.org/0000-0002-0291-7723

References

American National Standards Institute. (1997). *ANSI S3.5-1997 Methods for calculation of the speech intelligibility index*. New York, NY: Author.

Arweiler, I., & Buchholz, J. M. (2011). The influence of spectral characteristics of early reflections on speech intelligibility. *Journal of the Acoustical Society of America*, *130*(2), 996–1005. doi:10.1121/1.3609258

Beutelmann, R., & Brand, T. (2006). Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, *120*(1), 331–342. doi:10.1121/1.2202888

Beutelmann, R., Brand, T., & Kollmeier, B. (2009). Prediction of binaural speech intelligibility with frequency-dependent interaural phase differences. *Journal of the Acoustical Society of America*, *126*(3), 1359–1368. doi:10.1121/1.3177266

Beutelmann, R., Brand, T., & Kollmeier, B. (2010). Revision, extension, and evaluation of a binaural speech intelligibility model. *Journal of the Acoustical Society of America*, *127*(4), 2479–2497. doi:10.1121/1.3295575

Bradley, J. S., Sato, H., & Picard, M. (2003). On the importance of early reflections for speech in rooms. *Journal of the Acoustical Society of America*, *113*(6), 3233–3244. doi:10.1121/1.1570439

Brand, T., & Kollmeier, B. (2002). Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *Journal of the Acoustical Society of America*, *111*(6), 2801–2810. doi:10.1121/1.1479152

Durlach, N. I. (1963). Equalization and cancellation theory of binaural masking-level differences. *Journal of the Acoustical Society of America*, *35*(8), 1206–1218. DOI 10.1121/1.1918675

Ewert, S. D. (2013). AFC—A modular framework for running psychoacoustic experiments and computational perception models. In *Proceedings of the international conference on acoustics* AIA-DAGA, German Acoustical Society (DEGA) (pp. 1326–1329).

Feldmann, H. (1963). Untersuchungen über das binaurale Hören unter Einwirkung von Störgeräuschen [Investigating binaural hearing under the influence of noise]. *Archiv fúr Ohren-, Nasen- und Kehlkopfheilkunde vereinigt mit Zeitschrift fúr Hals-, Nasen- und Ohrenheilkunde*, *181*, 337–374.

George, E. L. J., Goverts, S. T., Festen, J. M., & Houtgast, T. (2010). Measuring the effects of reverberation and noise on sentence intelligibility for hearing-impaired listeners. *Journal of Speech, Language, and Hearing Research*, *53*(6), 1429–1439. doi:10.1044/1092-4388(2010/09-0197)

Hauth, C., & Brand, T. (2018). Modeling sluggishness in binaural unmasking of speech for maskers with time-varying interaural phase differences. *Trends in Hearing*, *22*, 1–10. DOI 10.1177/2331216517753547

Hochmuth, S., Jürgens, T., Brand, T., & Kollmeier, B. (2015). 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*(sup2), 23–34. doi:10.3109/14992027.2015.1088174

Hochmuth, S., Kollmeier, B., & Shinn-Cunningham, B. (2018). The relation between acoustic-phonetic properties and speech intelligibility in noise across languages and talkers. In *Proceedings of German Annual Conference on Acoustics*, Munich, Germany: German Acoustical Society (DEGA) (pp. 628–629).

Kokabi, O., Brinkmann, F., & Weinzierl, S. (2018). Segmentation of binaural room impulse responses for speech intelligibility prediction. *Journal of the Acoustical Society of America*, *144*, 2793–2800. DOI 10.1121/1.5078598

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M., Uslar, V., Brand, T., & Wagener, K. (2015). The multilingual matrix test: Principles, applications, and comparison across languages: A review. *International Journal of Audiology*, *54*, 3–16. DOI 10.3109/14992027.2015.1020971

Langford, T., & Jeffress, L. (1964). Effect of noise crosscorrelation on binaural signal detection. *Journal of the Acoustical Society of America*, *36*(8), 1455–1458. DOI: 10.1121/1.2142529

Leclère, T., Lavandier, M., & Culling, J. (2015). Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking, and binaural de-reverberation. *Journal of the Acoustical Society of America*, *137*, 3335–3345. DOI 10.1121/1.4921028

Licklider, J. (1948). The influence of interaural phase relations upon the masking of speech by white noise. *Journal of the Acoustical Society of America*, *20*(2), 150–159. DOI 10.1121/1.1906358

Lochner, J. P. A., & Burger, J. F. (1964). The influence of reflections on auditorium acoustics. *Journal of Sound and Vibration*, *1*(4), 426–454. DOI 10.1016/0022-460X(64)90057-4

Nábělek, A. K., & Robinette, L. (1978). Influence of the precedence effect on word identification by normally hearing and hearing-impaired subjects. *Journal of the Acoustical Society of America*, *63*(1), 187–194. DOI 10.1121/1.381711

Parizet, E., & Polack, J. D. (1992). The influence of an early reflection upon speech intelligibility in the presence of a background noise. *Acustica*, *77*(1), 21–30.

Rennies, J., Brand, T., & Kollmeier, B. (2011). Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet. *Journal of the Acoustical Society of America*, *130*(5), 2999–3012. doi:10.1121/1.3641368

Rennies, J., & Kidd, G. (2018). Benefit of binaural listening as revealed by speech intelligibility and listening effort. *Journal of the Acoustical Society of America*, *144*, 2147–2159. DOI 10.1121/1.5057114

Rennies, J., Warzybok, A., Brand, T., & Kollmeier, B. (2014). Modeling the effects of a single reflection on binaural speech intelligibility. *Journal of the Acoustical Society of America*, *135*, 1556–1567. DOI 10.1121/1.4863197

Rhebergen, K. S., & Versfeld, N. J. (2005). A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *Journal of the Acoustical Society of America*, *117*(4), 2181–2192. doi:10.1121/1.1861713

Soulodre, G. A., Popplewell, N., & Bradley, J. S. (1989). Combined effects of early reflections and background noise on speech intelligibility. *Journal of Sound and Vibration*, *135*(1), 123–133. doi:10.1016/0022-460X(89)90759-1

Steeneken, H. J. M., & Houtgast, T. (1980). A physical method for measuring speech transmission quality. *Journal of the Acoustical Society of America*, *67*(1), 318–326. DOI 10.1121/1.384464

vom Hövel, H. (1984). *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 thesis). RWTH Aachen, Germany.

Warzybok, A., Rennies, J., Brand, T., Doclo, S., & Kollmeier, B. (2013). Effects of spatial and temporal integration of a single early reflection on speech intelligibility. *Journal of the Acoustical Society of America, 133*, 269–282. DOI 10.1121/1.4768880