Omid Kokabi, Fabian Brinkmann, and Stefan Weinzierl

# Segmentation of binaural room impulse responses for speech intelligibility prediction



The following article appeared in

Kokabi, O., Brinkmann, F., & Weinzierl, S. (2018). Segmentation of binaural room impulse responses for speech intelligibility prediction. The Journal of the Acoustical Society of America, 144(5), 2793–2800.

and may be found at

https://doi.org/10.1121/1.5078598

# Segmentation of binaural room impulse responses for speech intelligibility prediction

Omid Kokabi,[a] Fabian Brinkmann, and Stefan Weinzierl
*TU Berlin, Audio Communication Group, Einsteinufer 17c, 10587 Berlin, Germany*

The two most important aspects in binaural speech perception—better-ear-listening and spatial-release-from-masking—can be predicted well with current binaural modeling frameworks operating on head-related impulse responses, i.e., anechoic binaural signals. To incorporate effects of reverberation, a model extension was proposed, splitting binaural room impulse responses into an early, useful, and late, detrimental part, before being fed into the modeling framework. More recently, an interaction between the applied splitting time, room properties, and the resulting prediction accuracy was observed. This interaction was investigated here by measuring speech reception thresholds (SRTs) in quiet with 18 normal-hearing subjects for four simulated rooms with different reverberation times and a constant room geometry. The mean error with one of the most promising binaural prediction models could be reduced by about 1 dB by adapting the applied splitting time to room acoustic parameters. This improvement in prediction accuracy can make up a difference of 17% in absolute intelligibility within the applied SRT measurement paradigm.

## I. INTRODUCTION

The most important binaural mechanisms for the perception of speech in acoustic environments with competing noise sources are better-ear listening and binaural unmasking of spatially separated sources (Middlebrooks *et al.*, 2017). Head shadowing and the ears' spatial sensitivity cause different signal-to-noise ratios (SNRs) at the listeners' left and right ear. Better-ear listening refers to the fact that the auditory system primarily extracts information from the ear signal with the more favorable signal-to-noise ratio (Edmonds and Culling, 2006). Binaural unmasking refers to reducing the strength of a masking sound source on a speech target when the two are spatially separated (Kock, 1950). Although there is no clear interpretation of how both mechanisms are exactly combined in the auditory system, additivity proved to be a successful candidate (Jelfs *et al.*, 2011).

Different auditory models have been developed to represent the two mechanisms. Among these, the Oldenburg model (Beutelmann and Brand, 2006; Beutelmann *et al.*, 2010) and the Cardiff model (Jelfs *et al.*, 2011; Lavandier and Culling, 2010) seem to be most promising (Culling *et al.*, 2013). Both models combine an SNR/speech intelligibility index (SII) (ANSI S3.5, 1997) based better-ear evaluation with a modeling stage for binaural unmasking based on the equalization-cancellation (EC) theory (Durlach, 1963). The model input is either a binaural stream of the speech and masker ear signals, or a binaural room impulse response (BRIR), describing the transfer path between the speech and masker sources and the human receiver.

In typical rooms, the speech signal is a combination of the direct signal, a series of early distinct room reflections and late diffuse reverberation. While distinct reflections shortly following the direct sound are generally considered to improve speech intelligibility (Bradley *et al.*, 2003), reverberation is known to have a detrimental effect by increasing the temporal masking due to a reduced depth in the temporal modulation of running speech.

In both models mentioned above, however, the entire speech signals is considered as useful, thus ignoring the detrimental effect of reverberation on speech reception. To account for this, it was proposed to split the BRIR into an early, useful and a late, detrimental part (Rennies *et al.*, 2011), referred to as the U/D-approach in the remainder of this document. Both parts are fed separately into the model and are considered as the speech target and as an additional masker. The U/D-concept can also be found in many room acoustic parameters such as Clarity $C_{80}$ or Definition $D_{50}$, which are used to predict the transparency of speech and music (ISO 3382-1, 2010). However, different U/D-limits ranging from 35 to 95 ms are applied (Bradley, 1986; Lochner and Burger, 1964).

By extending the Oldenburg model with a U/D-approach, the prediction accuracy could be improved both for a simple case consisting of a direct signal and one lateral or frontal reflection (Rennies, 2014) as well as for a more complex sound field with non-negligible levels of reverberation (Rennies *et al.*, 2011). Improved performance was also observed for the U/D-extended Cardiff model (Leclère *et al.*, 2015). The optimal U/D-limit was found to depend on the properties of the room, which was considered as a general downside of this approach. A link between the respective U/D-limits and room acoustic properties, however, was not investigated so far.

The present work tries to fill this gap by predicting optimal U/D-limits for different room acoustical environments

[a]Electronic mail: kokabi@tu-berlin.de

and source-receiver configurations using room acoustic parameters, thus increasing the precision and the generalizability of binaural models for speech perception. Therefore, SRTs in quiet were measured for a virtual room with systematically varied acoustic properties.

## II. METHOD

### A. SRT measurements

#### 1. Subjects

Eighteen native German speakers (13 male, 5 female; average age 30.4) with normal hearing [ISO 8253-1 hearing levels (HLs) between −10 and + 20 dB HL] participated in the tests on a voluntary basis. Except for two, all subjects had experience with psychoacoustic listening tests.

#### 2. Procedure

The Oldenburg sentence test (OLSA) (Kühnel *et al.*, 1999; Wagener *et al.*, 1999a,b) was used to measure SRTs in quiet, i.e., without additional masking noise sources, by finding the sound pressure level that is required for 50% correctly understood words. For this purpose, test sentences consisting of five words at a natural speech rate with a fixed syntax (name–verb–number–adjective–object) but unpredictable semantics were presented to the participants. The participants were asked to repeat the test sentence, after which the experimenter adaptively adjusted the level of the successive sentence according to the number of correctly understood words in steps from ±1 to ±3 dB for sentences 2–5, and from ±1 to ±2 dB for sentences 6–31 (HörTech gGmbH, 2011). The test converges at the SRT (50% correctly understood words) within a set of 30 test sentences per condition. The OLSA corpus is comprised of 120 different sentences, which are combined into 40 test lists of 30 sentences per list.

Rennies *et al.* (2011) found a significant correlation between pure tone thresholds and measured SRTs in quiet even for listeners with normal HLs <20 dB HL, i.e., subjects with lower overall hearing sensitivities tend to show higher (= worse) SRTs. As the current study focused on the effect of reverberation on SRTs and not on the effect of hearing sensitivity, it seemed desirable to compensate the measured SRTs for the latter to achieve a clearer display of the experimental data. To do so, HLs were measured for every subject by means of individual pure tone audiograms for both ears and frequencies between 125 Hz and 8 kHz (IEC 60645-1, 2017; ISO 8253-1, 2010). For each subject, the pure tone average (PTA = mean dB HL) at 0.5, 1, and 2 kHz was calculated taking the ear data with the lower hearing level per band (assuming better-ear listening in speech perception). These better-ear PTAs ranged between −6 dB HL (most sensitive subject) and + 8 dB HL (least sensitive subject). To compensate the SRTs for these inter-individual differences, the better-ear PTAs were subtracted from the measured SRTs. A correlation analysis between each subjects better-ear PTA and his/her mean SRT across conditions revealed a high correlation ($r \approx 0.71$, $p < 0.001$), confirming the findings by Rennies *et al.* (2011).

Four test conditions with different acoustic conditions discussed below with 30 sentences per condition were prepared for every participant. The participants were positioned in a hemi-anechoic chamber at TU Berlin with the experimenter in the adjacent control room. The stationary noise level in the hemi-anechoic chamber was below 20 dB(A) (logged during the entire session with an NTI XL2 sound level meter, NTI MA220 Mic-preamp, and an NTI MA2230 microphone, calibrated via Larson Davis CAL200 acoustic calibrator). The stimuli were played back via a Focusrite Scarlett 18i20 USB interface, and closed, circumaural Beyerdynamic DT770Pro headphones. The headphones were calibrated to absolute sound pressure levels via a B&K Artificial Ear type 4152, a preamplifier B&K type 2609, and a B&K sound level calibrator type 4230. Audio playback was controlled by a laptop running MATLAB in the control room. For the audiogram test, the participant directly responded via a generated MATLAB user interface. For the SRT measurement, the participant made a spoken response via an Omnitronic GMTS100 intercom terminal with talk-back microphone.

The test started with the pure tone audiogram, followed by the SRT measurements for the four test conditions in randomized order. To familiarize the participants with the task and the stimuli, training was performed prior to the actual tests. The procedure with instruction, training, and filling out the questionnaire took about 70 min per participant.

The (re)-positioning of headphones slightly changes the frequency-dependent stimulus level at the listener's ear drum and causes an uncertainty in pure tone audiometry (Paquier *et al.*, 2012) and an audible coloration of the stimulus (Paquier and Koehl, 2015). To reduce this source of error, the participants were instructed to not move or touch the headphones during the entire test. This way, the measured hearing levels are sufficiently accurate with respect to the presentation level of the OLSA sentences.

#### 3. Stimuli

The physical response of a room is characterized by the reflection pattern (temporal structure and amplitude) arriving at the listener's ears. While the temporal structure is related to the room geometry and the positions of source and receiver, the amplitudes of the individual reflections are mainly determined by the boundary conditions (absorption, scattering) of the surfaces. To be able to independently vary the room geometry and surface properties, all BRIRs were simulated with the geometrical acoustics software RAVEN (Schröder and Vorländer, 2011). The acoustic environment for which BRIRs were generated was based on the geometry of an existing, medium sized auditorium with shoebox design featuring diffusing wall and ceiling elements with an elevated stage and an audience area (Fig. 1).

In a first step, BRIRs for seven different room configurations were simulated by scaling the room size and absorption coefficients (combinations of four volumes $V = \{500, 1000, 2000, 4000\}$ m$^3$ at a fixed reverberation time of $T_{20,m} = 1$ s and four reverberation times $T_{20,m} = \{0.5, 1, 2, 4\}$ s at a fixed volume of $V = 1000$ m$^3$). An informal listening test showed

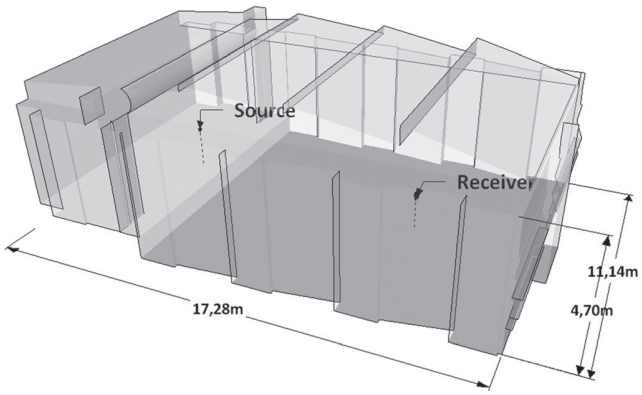


FIG. 1. Three-dimensional room model ($V = 1000\ m^3$) with dimensions and source/receiver position. The gray shade denotes the surface type (seating and residual).

a stronger impact on speech intelligibility when scaling the absorption coefficients for a room of fixed size than vice versa. As a consequence, the SRT measurements were conducted for four conditions with varying reverberation times by scaling the surface absorption coefficients. Absorption values maintained a typical behavior both in size and in frequency dependence under all test conditions.

BRIRs were calculated for a source at the center of the stage and a binaural receiver in the audience area at a distance of approximately 9 m corresponding to about three times the critical distance at the lowest reverberation level. For the source, the directivity of a male singer was applied (average directivity index $Q = 1.5$ for 500 Hz and 1 kHz octaves). Measured head related transfer functions (HRTFs) of the FABIAN head-and-torso simulator with a resolution of 2° in azimuth and elevation were used as receiver directivity (Brinkmann *et al.*, 2017b). Binaural auralizations of the OLSA sentence corpus were calculated via convolution with the generated BRIRs. To avoid coloration due to the frequency response of the headphones, we used an inverse filter of the Beyerdynamic DT770Pro headphones from the FABIAN database (Brinkmann *et al.*, 2017a).

The applied absorption and scattering coefficients as well as the resulting frequency dependent reverberation statistics and the calculated BRIRs (headphone filter *not* applied) are accessible in Kokabi *et al.* (2018).

## B. SRT prediction

### 1. General prediction procedure

The generated BRIRs were applied to the Cardiff binaural model (Jelfs *et al.*, 2011) implemented in the auditory modeling toolbox (Søndergaard and Majdak, 2013). The Cardiff model was chosen due to (a) its computational efficiency, (b) its open source availability, and (c) the fact that no parameter-fitting is involved in the implemented modelling stages for better-ear listening and binaural unmasking—apart from the JND-jitter implementation introduced in the original EC-model (Durlach, 1963). The model was extended by a temporal U/D-classification as suggested, e.g., by Rennies *et al.* (2011), implemented by the authors. For the latter, each BRIR was multiplied with two time windows: an early window consisted of a flat (weight = 1) part from the time of arrival up to the considered U/D-limit, and a linear fade-out with a length of 1 ms. A late window starting with zeros up to the considered U/D-limit, followed by a fade-in of length 1 ms, and continued with a flat part (weight = 1) until the end of the BRIR. The early (useful) part was used to generate the speech target and the late (detrimental) part was used to generate the masker. Both were separately fed into the model.

The model output is a SNR in dB predicting the benefit of binaural listening over listening to an omnidirectional receiver at the same position. As suggested by Jelfs *et al.* (2011), the predicted benefit was converted to an SRT by a multiplication by −1, and by scaling every benefit by the same factor until the average across all predictions matches the average across all measured SRTs. By doing so, the model output can directly be compared to measured SRTs in the respective condition. It is important to note that the model is only able to predict relative SRT differences between test conditions due to the matching of the means of measured and predicted data. Due to the fact that only relative differences between conditions can be predicted by the model, the compensation applied to the measured SRTs based on each subjects' better-ear PTA (cf. Sec. II A 2) has no effect on the prediction accuracy of the model. In addition, the prediction accuracy with fixed and room-dependent U/D-limits was also tested for an external dataset with SRTs in quiet measured for two conditions S0 (source in front of the listener) and S90 (source to the right of the listener) in a virtual rectangular room (length: 10 m, width: 15 m, height: 3 m) with reverberation times of about 2 s, simulated with CATT-Acoustic v8. The rationale for incorporating this additional dataset in the present evaluation was to further validate the derived prediction method on data which were not part of the derivation process. The two test conditions of the external dataset each feature four source-receiver distances, ranging from $d = 0.5$ m to $d = 13.0$ m (Rennies *et al.*, 2011). This dataset is referred to as RS11 in the remainder of this document.

### 2. Fitted U/D-limits

U/D-limits fitted to the measured SRT values were determined by calculating SRT predictions with the method given above, whereby for every condition (BRIR), 19 different U/D-limits from 20 to 200 ms with 10 ms steps were used, resulting in $19^4$ predicted SRT sets for each participant of the listening test. All U/D-limits leading to a mean absolute error (MAE) between measurement and prediction of $<1$ dB across all four conditions were selected. From this subset, the mean was calculated for each test condition, and taken as the fitted U/D-limit. Since differences between MAEs were quite small, this method was regarded as more robust than considering only the U/D-combination with the smallest MAE.

### 3. Room acoustical prediction of U/D-limits

To predict U/D-limits from room acoustic parameters, a linear regression analysis was performed with the room

acoustic parameters as independent variables, and the fitted U/D-limits as dependent variable. Since binaural dereverberation in speech perception was shown to be correlated to monaural acoustic parameters as well as binaural parameters assessing the similarity between both ear signals (Ellis *et al.*, 2015), three parameters were used as predictors in the regression analysis: Clarity ($C80_m$, ISO 3382-1, 2009) and the direct-to-reverberant energy ratio ($D/R$) as monaural predictors, and $IACC_m$ as a binaural predictor, where $m$ denotes the average over the 500 Hz and 1 kHz octave values. The room acoustic parameters $D/R$ and $C80_m$ were calculated from room impulse response (RIRs) with omnidirectional source and receiver directivities at the same positions used for the BRIR calculation. In case of the data from RS11, these parameters were calculated from the BRIRs (mean across ears), as monaural RIRs were not available. The *IACC* was always calculated from the BRIRs. $C80_m$ and $IACC_m$ were calculated using the ITA-Toolbox (Dietrich *et al.*, 2010). $D/R$ was calculated as the energy ratio of the direct to reverberant part of the RIR with a time limit of 2.5 ms to separate the two parts (Zahorik, 2002).

The results of the regression analysis were then used to predict U/D-limits from the room acoustic parameters. These predicted U/D-limits were tested against two fixed U/D-limits: 50 ms (recommendation in ISO 3382-1 for Clarity for speech) and 100 ms (better prediction than with 50/ 80 ms in Rennies *et al.*, 2011).

## III. RESULTS

Measured SRTs and predicted SRTs with predicted and two fixed U/D-limits (50 ms and 100 ms) are shown in Fig. 2 for all four test conditions and averaged across participants. The MAE averaged across test conditions is given in Table I. To test for systematic differences in measured SRT data between conditions, a one-way repeated measures analysis of variance (ANOVA) with a significance level of 0.05 and Greenhouse-Geisser correction was applied. The results reveal a significant effect of the level of reverberation on the measured SRTs for the four test conditions [$F(1.4, 24.4) = 206.3$, $p < 0.001$]. *Post hoc* tests using Bonferroni correction revealed that the measured SRTs at all tested levels of reverberation were significantly different from each other ($p < 0.001$). For completeness, the statistical analysis was repeated without compensation of the SRT data where the ANOVA also showed a significant effect of level of reverberation on measured SRTs [$F(1.4, 24.4) = 206.3$, $p < 0.001$]. As in the case with SRT compensation, *post hoc* tests using Bonferroni correction revealed that measured SRTs at all tested levels of reverberation were significantly different from each other ($p < 0.001$).

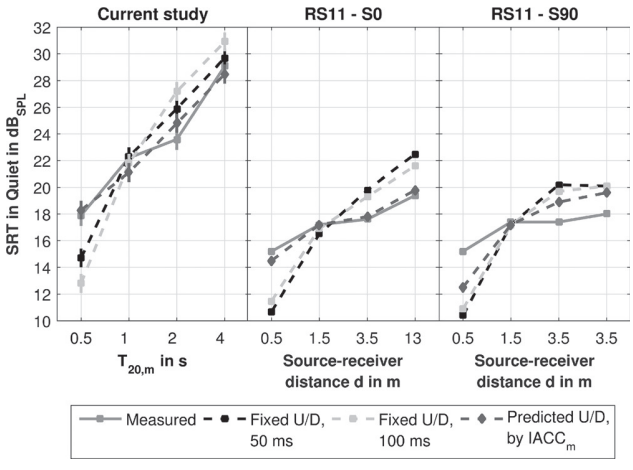


FIG. 2. Measured and predicted SRTs with fixed and predicted U/D-limits averaged across participants. Standard errors are shown as vertical bars.

TABLE I. MAEs in dB from fixed, fitted, and predicted U/D-limits in ms.

| | U/D-limit | | | | | |
|---|---|---|---|---|---|---|
| | Fixed | | | Predicted by | | |
| | 50 | 100 | Fitted | *D/R* | *C*80 | *IACC* |
| Current study | 1.9 | 2.9 | 0.2 | 1.3 | 1.3 | 1.2 |
| RS11 – S0 | 2.6 | 2.0 | 0.3 | 1.3 | 1.5 | 0.3 |
| RS11 – S90 | 2.5 | 2.0 | 0.5 | 0.9 | 2.0 | 1.5 |
| ∅ | 2.3 | 2.4 | 0.3 | 1.2 | 1.6 | 1.0 |

### A. Fixed U/D-limits

The data of the current study (Fig. 2, left) show that measured and predicted SRTs with fixed U/D-limits increase with increasing level of reverberation. Comparing the prediction accuracy with fixed U/D-limits, it can be seen that the error for U/D = 50 ms ($MAE_{mean} = 1.9$ dB) is slightly lower than with U/D = 100 ms ($MAE_{mean} = 2.9$ dB) for the data from the current study. However, this trend is reversed for the RS11 data (U/D = 50 ms: $MAE_{mean} = 2.6$ dB; U/D = 100 ms: $MAE_{mean} = 2.0$ dB), cf. Table I.

Because the absolute level of the predicted SRTs has to be manually matched to the measured SRTs, only SRT differences between test conditions can be predicted by the model. They can be deduced from the gradient of the lines connecting any two test conditions. The under-/overestimation of SRT increase with both the data from the current study and the RS11 data and the prediction model with the fixed U/D-limits is depicted in Fig. 3. For the current study, the SRT-increase is overestimated in the low and medium reverberant conditions ($0.5 \le T_{20,m} \le 2$ s), but underestimated between the conditions with $T_{20,m} = 2$ s and $T_{20,m} = 4$ s for both fixed U/D-limits. A similar trend can be observed for the SRTs measured at different source distances (RS11 data) where larger overestimations can be observed between conditions for distances below 3.5 m. For source distances between 3.5 and 13 m, quite accurate predictions can be observed with both fixed U/D-limits (under-/overestimation <1 dB).

### B. Fitted and predicted U/D-limits

The fitted and predicted U/D-limits averaged across all participants are shown in Table II together with the values of the room acoustic parameters used for the prediction. As can be seen, the U/D-limits increase with increasing level of reverberation (current study) and with increasing distance from the source (RS11 data) in almost all cases. As

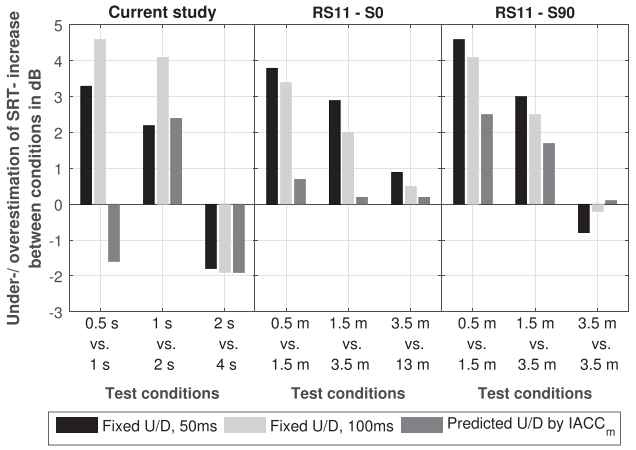


FIG. 3. Estimation error of SRT increase between test conditions. Values >0 dB denote an overestimation, values <0 dB and underestimation.

mentioned above, the predicted U/D-limits were obtained by means of regression analyses between the room acoustic parameters and the fitted U/D-limits. For both single-channel parameters $D/R$ and $C80_m$, significant regression equations could be found with $[F(1,\ 70) \approx 121,\ p < 0.001]$ and an adjusted $R^2$ of 0.62 for $C80_m$ and $[F(1, 70) \approx 126, p < 0.001]$ with an adjusted $R^2$ of 0.64 for $D/R$. The corresponding linear regression equations yield predicted U/D-limits with 54.9–6.3 ($D/R$) ms and 93.4–5 ($C80_m$) ms, respectively, both with a standard error of 21 ms. Slightly better results were obtained for the $IACC$ $[F(1, 70) = 191.7, p < 0.001]$, with an adjusted $R^2$ of 0.73. The corresponding linear regression equation yields predicted U/D-limits with 143–202 ($IACC_m$) ms, with a standard error of 18 ms.

The MAEs given in Table I show that the fitted U/D-limits clearly outperform the others with errors ≤0.5 dB. The MAEs based on predicted U/D-limits are smaller than those based on fixed limits and larger than results obtained with fitted limits. Noteworthy, improvements from 0.6 to 1.6 dB can be observed in comparison to the values from fixed U/D-limits for the current study and data from RS11, despite the fact that the regression formulae were calculated based on data from the current study only. The observed mean improvement in prediction accuracy of ≈ 1 dB can make up a difference of 17% in absolute intelligibility, which can be deduced from the slope of the discrimination function within the applied SRT measurement paradigm (Wagener *et al.*, 1999a,b). Moreover, the prediction of differences between test conditions improves, and systematic over- and underestimations are reduced/ disappear (cf. Fig. 3).

In the informal listening test, the scaling of the absorption coefficient of a room with fixed volume turned out to have a stronger effect on speech intelligibility than scaling the volume of a room with fixed absorption coefficients. This trend can be confirmed *post hoc* by calculating the predicted SRTs with U/D-limits as a function of $IACC_m$ for all seven rooms of the informal listening test (predicted SRT range 10 dB for scaling the absorption, 3 dB for scaling the volume, Fig. 4). This is yet another, albeit qualitative, indicator for the generalizability of the suggested U/D approach.

## IV. DISCUSSION

The prediction of speech intelligibility based on standard binaural models with better-ear identification and binaural unmasking can be improved by splitting the binaural impulse response at the input stage into an early, useful and a late, detrimental part (U/D-approach, Rennies, 2014; Leclère *et al.*, 2015). However, the use of *fixed* temporal U/D-limits tends to underestimate the level of intelligibility for signals with little reverberation and to overestimate the intelligibility for signals with much reverberation relative to values for medium reverberation. This was shown by measuring the SRT in rooms with different reverberation time (current study) and different source-receiver distances within the same room (RS11 data, Figs. 2 and 3). Based on these observations, one must conclude that there are obviously perceptual mechanisms that mitigate the deterioration of speech perception with increasing level of reverberation, which are not accounted for by a model with fixed U/D-limits.

With the current study, we were able to show that the prediction error for the SRT resulting from the Cardiff model for binaural speech perception (Jelfs *et al.*, 2011; Lavandier and Culling, 2010) can be reduced by about 1 dB by using U/

TABLE II. Fitted and predicted U/D-limits in ms and room acoustic parameters ($D/R$ and $C80_m$ in dB).

| | | Fitted U/D-limits mean (standard deviation) | Room acoustic parameters | | | Predicted U/D-limits | | |
|---|---|---|---|---|---|---|---|---|
| | | | $D/R$ | $C80_m$ | $IACC_m$ | $D/R$ | $C80_m$ | $IACC_m$ |
| Current study | $T_{20,m} = 0.5$ s | 59 (8) | −1.1 | 6.4 | 0.43 | 62 | 61 | 56 |
| | $T_{20,m} = 1.0$ s | 90 (14) | −6.6 | −0.7 | 0.22 | 96 | 97 | 99 |
| | $T_{20,m} = 2.0$ s | 142 (11) | −10.1 | −4.9 | 0.08 | 119 | 118 | 127 |
| | $T_{20,m} = 4.0$ s | 122 (24) | −13.1 | −8.7 | 0.06 | 137 | 137 | 131 |
| RS11 − S0 | $d = 0.5$ m | 48 (25) | 3.2 | 4.6 | 0.65 | 35 | 70 | 12 |
| | $d = 1.5$ m | 122 (37) | −3.4 | −0.5 | 0.29 | 76 | 96 | 84 |
| | $d = 3.5$ m | 162 (28) | −9.5 | −2.1 | 0.13 | 115 | 104 | 117 |
| | $d = 13.0$ m | 162 (26) | −20.8 | −3.7 | 0.10 | 186 | 112 | 123 |
| RS11 - S90 | $d = 0.5$ m | 35 (15) | 2.8 | 3.5 | 0.54 | 37 | 76 | 34 |
| | $d = 1.5$ m | 125 (34) | −5.5 | −2.2 | 0.28 | 90 | 104 | 86 |
| | $d = 3.5$ m | 171 (22) | −12.3 | −2.5 | 0.21 | 132 | 106 | 101 |
| | $d = 3.5$ m | 169 (24) | −15.2 | −2.5 | 0.26 | 151 | 106 | 90 |

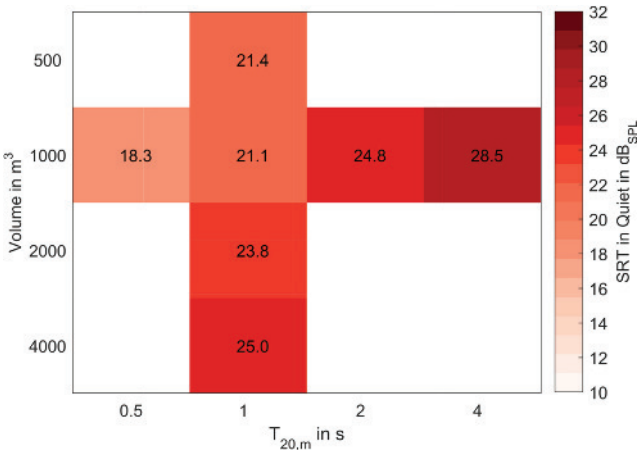


FIG. 4. (Color online) Predicted SRT with the binaural model and the U/D-extension as a function of $IACC_m$ for all seven rooms used in the informal listening test.

D-limits adapted to the acoustic environment compared to the model with fixed U/D limits. As the best room acoustic predictor for the adapted U/D-limit, we identified the $IACC_m$, which describes the similarity between the ear signals. Predictions of similar accuracy, however, can be reached with Clarity ($C_{80}$) and D/R as room acoustic parameters (cf. Table I). Since measurements of *IACC* are more complex than the measurement of energy ratios such as $D/R$ or $C_{80}$, the latter might be preferred for practical applications.

For a low $IACC_m$ (low $C_{80}$, low D/R), correlated with a high level of diffuse reverberation, the U/D-limit is increased raising the energy ratio between the early useful and the late detrimental components of the BRIR, i.e., the better-ear SNR calculated by the model. For a high $IACC_m$ (high $C_{80}$, high D/R), typical for dry signals with little diffuse reverberation, the U/D-limit is decreased, resulting in a reduced energy ratio between useful and detrimental components and a corresponding decrease in SNR.

Room-adapted U/D-limits can be considered as a functional extension of binaural models which reduce the prediction error. The trend that is reflected in the room-dependence, however, also indicates which perceptual mechanisms might be responsible for this effect. We see two potential candidates for this: binaural de-reverberation and room adaption.

Binaural de-reverberation, i.e., the partial suppression of room reverberation, leads to an improved signal recognition in a reverberant context when listening binaurally compared to monaurally. It has been shown by Gelfand and Hochberg (1976), Moncur and Dirks (1967), and Nábělek and Robinson (1982), that the extent of binaural de-reverberation depends on the absolute levels of reverberation apparent in the room. The largest benefits due to binaural listening could be observed for medium reverberant rooms, i.e., reverberation times of 1–2 s (test conditions ranged from 0 s to a maximum of 3 s in mentioned studies). For lower and higher levels of reverberation, this benefit vanished. A similar pattern can be observed in the U/D-limit we have to assume to correctly predict the measured SRTs (cf. Fitted U/D-limits in Table II): The U/D-limit increases from low to medium levels of reverberation ($T_{20,m} = 2$ s) where it reaches a maximum and slightly decreases again for higher levels of reverberation. Similar trends can be observed for the RS11 data, except for the slight decrease at large source distances.

Room adaptation refers to the partial suppression of the effect of reverberation on speech intelligibility with prior exposure to the reverberant environment compared to no prior exposure. Also there, the largest influence occurred at medium levels of reverberation of $T = 1$ s with a decrease in SRT of about 3 dB, vanishing to lower and higher levels of reverberation (Zahorik and Brandewie, 2016). This is in line with findings showing a lower consonant identification performance with increasing level of reverberation on the test word alone but an increasing performance when the context (i.e., preceding words) featured the same level of reverberation as the test word. Further, it was shown, that the identification performance increased with increasing duration of the reverberant context (Beeston *et al.*, 2014; Watkins, 2005a,b). The impact of room adaptation thus tends to exhibit the same dependence on room acoustic properties as the impact of binaural de-reverberation.

To account for this effect, a binaural model would need some knowledge about prior exposure to the acoustic environment. In its current implementation, there is no option to provide the model with such information. Moreover, there still seems to be too little knowledge about the relevant aspects driving the effect of room adaptation (speech rate, exposure time) and if this is a monaural or a binaural mechanism.

To account for the effect of binaural de-reverberation, some sort of binaural processing is required. In the applied model, the only candidate for this would be the EC-stage implemented. Initially developed based on observations of masking level thresholds as a function of ITD and ILD, it was implemented to account for the unmasking of spatially distributed, localized target and masker sources. The current EC-implementation is driven by interaural phase differences (IPDs) of the speech target and masker and weighted by the interaural coherence of the masker. In a fixed spatial configuration where target and masker are not co-located (i.e., target IPD $\neq$ masker IPD), a higher masker coherence is correlated with a higher binaural advantage, as both masker components in the left and right masker ear signal can be canceled more effectively.

With an increasing level of reverberation, the interaural coherence of the masker decreases, hence the binaural advantage according to the EC-theory decreases. This was shown in the unmasking study by Lavandier and Culling (2010), who calculated the binaural advantage with the same model as in the present study. To model de-reverberation, however, the binaural benefit would have to increase with increasing level of reverberation (up to a certain limit), i.e., with decreasing masker coherence. This is contrary to EC-theory, hence the binaural model in its current form *cannot* account for the effect of binaural de-reverberation. It also cannot be concluded that binaural de-reverberation is unmasking from the late, diffuse masking source (Leclère *et al.*, 2015) since unmasking and binaural de-reverberation are obviously inversely correlated with diffuse reverberation.

The relative importance of the individual mechanisms could further be evaluated with additional "knock-out" listening test conditions that try to deactivate a single perceptual mechanism: room adaptation could be deactivated following the procedure employed by Zahorik and Brandewie (2016), where the room (BRIR) was changed after each test sentence. Binaural de-reverberation might be deactivated by switching only the late reverberant part of the signal to monaural presentation leaving the early part of the signal binaural. Binaural unmasking—which is expected to be observed only for strong room reflections after an initial fusion time—might be deactivated by switching only the early part to a monaural presentation, leaving the late diffuse part binaural. However, in the latter two cases, the time that separates the binaural from the monaural part of the impulse response had to be subject of investigation itself. Moreover, these treatments might interact with each other to a certain amount. On the modelling side, binaural de-reverberation would need to be implemented as a pre-processing stage to the better-ear model, as the binaural suppression of late reverberation is expected to affect the SNR evaluated by the better-ear model. A potential candidate for implementation could be the (still speculative) model by Beeston (2015), which could at least qualitatively model binaural de-reverberation by dynamic-range adaptation of the internal signal representation as a function of reverberation. Room adaptation could be modelled therein by scaling the amount of adaptation as a function of exposure time.

## V. CONCLUSION

The present study showed that the binaural intelligibility model with its SII-weighted combination of a better-ear evaluation, an EC-stage to account for binaural unmasking, and a fixed U/D-limit to account for the effects of reverberation cannot fully model the room-dependent perceptual mechanisms affecting speech perception (with further competing sources being absent). Deviations between measured and modeled SRTs were observed. Two mechanisms, namely room adaptation and binaural de-reverberation were suspected to affect the measured SRTs. With the implementation of a room-dependent U/D-classification that was coupled to room acoustic parameters of the respective environment, a functional extension was presented which was able to reduce the prediction error by about 1 dB, which can make up a difference of 17% in absolute intelligibility within the applied SRT measurement paradigm. The extension was tested for data from different studies and proved to be robust against different acoustic conditions. This initial validation suggests that the presented U/D prediction based on the *IACC* or *D*/*R* might be applicable to a wide range of acoustic environments, making it a valuable tool as long as the binaural mechanisms with their impact on binaural speech perception in reverberant environments are not fully understood and implemented in the model.

## ACKNOWLEDGMENTS

The authors thank Jan Rennies-Hochmuth for providing binaural room impulse responses and two anonymous reviewers for their constructive comments on an earlier version of this text which further improved the manuscripts' quality.

ANSI (**1997**). S3.5, *Methods for the Calculation of the Speech Intelligibility Index* (Acoustical Society of America, New York).

Beeston, A. V. (**2015**). "Perceptual compensation for reverberation in human listeners and machines," Ph.D. thesis, University of Sheffield, Sheffield, UK.

Beeston, A. V., Brown, G. J., and Watkins, A. J. (**2014**). "Perceptual compensation for the effects of reverberation on consonant identification: Evidence from studies with monaural stimuli," J. Acoust. Soc. Am. **136**(6), 3072–3084.

Beutelmann, R., and Brand, T. (**2006**). "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **120**(1), 331–342.

Beutelmann, R., Brand, T., and Kollmeier, B. (**2010**). "Revision, extension, and evaluation of a binaural speech intelligibility model," J. Acoust. Soc. Am. **127**(4), 2479–2497.

Bradley, J. S. (**1986**). "Predictors of speech intelligibility in rooms," J. Acoust. Soc. Am. **80**(3), 837–845.

Bradley, J. S., Sato, H., and Picard, M. (**2003**). "On the importance of early reflections for speech in rooms," J. Acoust. Soc. Am. **113**(6), 3233–3244.

Brinkmann, F., Lindau, A., Weinzierl, S., Geissler, G., van de Par, S., Müller-Trapet, M., Opdam, R., and Vorländer, M. (**2017a**). "The FABIAN head-related transfer function data base," https://depositonce.tu-berlin.de//handle/11303/6153 (Last viewed November 5, 2018).

Brinkmann, F., Lindau, A., Weinzierl, S., Müller-Trapet, M., Opdam, R., and Vorländer, M. (**2017b**). "A high resolution and full-spherical head-related transfer function database for different head-above-torso orientations," J. Audio Eng. Soc. **65**(10), 841–848.

Culling, J. F., Lavandier, M., and Jelfs, S. (**2013**). "Predicting binaural speech intelligibility in architectural acoustics," in *The Technology of Binaural Listening* (Springer, New York), pp. 427–447.

Dietrich, P., Masiero, B., Müller-Trapet, M., Pollow, M., and Scharrer, R. (**2010**). "Matlab toolbox for the comprehension of acoustic measurement and signal processing," in *Fortschritte der Akustik–DAGA*, 15–18 March 2010, Berlin Germany, pp. 517–518.

Durlach, N. I. (**1963**). "Equalization and cancellation theory of binaural masking-level differences," J. Acoust. Soc. Am. **35**(8), 1206–1218.

Edmonds, B. A., and Culling, J. F. (**2006**). "The spatial unmasking of speech: Evidence for better-ear listening," J. Acoust. Soc. Am. **120**(3), 1539–1545.

Ellis, G. M., Zahorik, P., and Hartmann, W. M. (**2015**). "Using multidimensional scaling techniques to quantify binaural squelch," Proc. Mtgs. Acoust. **23**(1), 050007.

Gelfand, S. A., and Hochberg, I. (**1976**). "Binaural and monaural speech discrimination under reverberation," Int. J. Audiol. **15**(1), 72–84.

HörTech gGmbH (**2011**). "Oldenburger Satztest—Adaptive Sprachaudiometrie mit Sätzen in Ruhe und im Störgeräusch—Bedienungsanleitung für den manuellen Test auf CD," https://www.hoertech.de/images/hoertech/pdf/mp/produkte/olsa/HT.OLSA_Handbuch_Rev01.0_mitUmschlag.pdf (Last viewed November 5, 2018).

IEC (**2017**). IEC 60645-1, *Electroacoustics—Audiometric Equipment—Part 1: Equipment for Pure-Tone and Speech Audiometry* (IEC, Geneva, Switzerland).

ISO (**2009**). ISO 3382-1, *Acoustics—Measurement of Room Acoustic Parameters—Part 1: Performance Spaces* (ISO, Geneva, Switzerland).

ISO (**2010**). ISO 8253-1, *Acoustics—Audiometric Test Methods—Part 1: Pure-Tone Air and Bone Conduction Audiometry* (ISO, Geneva, Switzerland).

Jelfs, S., Culling, J. F., and Lavandier, M. (**2011**). "Revision and validation of a binaural model for speech intelligibility in noise," Hear. Res. **275**(1), 96–104.

Kock, W. E. (**1950**). "Binaural localization and masking," J. Acoust. Soc. Am. **22**(6), 801–804.

Kokabi, O., Brinkmann, F., and Weinzierl, S. (**2018**). "Assessment of speech perception based on binaural room impulse responses," depositonce.tu-berlin.de//handle/11303/7505.2 (Last viewed November 5, 2018).

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

Lavandier, M., and Culling, J. F. (**2010**). "Prediction of binaural speech intelligibility against noise in rooms," J. Acoust. Soc. Am. **127**(1), 387–399.

Leclère, T., Lavandier, M., and Culling, J. F. (**2015**). "Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking, and binaural de-reverberation," J. Acoust. Soc. Am. **137**(6), 3335–3345.

Lochner, J. P. A., and Burger, J. F. (**1964**). "The influence of reflections on auditorium acoustics," J. Sound Vib. **1**(4), 426–454.

Middlebrooks, J., Simon, J. Z., Popper, A. N., and Fay, R. R. (**2017**). *The Auditory System at the Cocktail Party* (Springer, New York).

Moncur, J. P., and Dirks, D. (**1967**). "Binaural and monaural speech intelligibility in reverberation," J. Speech Lang. Hear. Res. **10**(2), 186–195.

Nábělek, A. K., and Robinson, P. K. (**1982**). "Monaural and binaural speech perception in reverberation for listeners of various ages," J. Acoust. Soc. Am. **71**(5), 1242–1248.

Paquier, M., and Koehl, V. (**2015**). "Discriminability of the placement of supra-aural and circumaural headphones," Appl. Acoust. **93**, 130–139.

Paquier, M., Koehl, V., and Jantzem, B. (**2012**). "Influence of headphone position in pure-tone audiometry," in *Proceedings of the Acoustics 2012 Joint Congress (11ème Congrès Français d'Acoustique-2012 Annual IOA Meeting*), May 13–18, Hong Kong, pp. 3925–3930.

Rennies, J. (**2014**). "Modeling the effects of a single reflection on binaural speech intelligibility," J. Acoust. Soc. Am. **135**(3), 1556–1567.

Rennies, J., Brand, T., and Kollmeier, B. (**2011**). "Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet," J. Acoust. Soc. Am. **130**(5), 2999–3012.

Schröder, D., and Vorländer, M. (**2011**). "RAVEN: A real-time framework for the auralization of interactive virtual environments," in *Forum Acusticum*, https://www2.ak.tu-berlin.de/~akgroup/ak_pub/seacen/2011/Schroeder_2011b_P2_RAVEN_A_Real_Time_Framework.pdf (Last viewed November 5, 2018).

Søndergaard, P., and Majdak, P. (**2013**). "The auditory modeling toolbox," in *The Technology of Binaural Listening* (Springer, Berlin-Heidelberg), pp. 33–56.

Wagener, K., Brand, T., and Kollmeier, B. (**1999a**). "Entwicklung und Evaluation eines Satztests für die deutsche Sprache II: Optimierung des Oldenburger Satztests" ("Development and evaluation of a German sentence test II: Optimization of the Oldenburg sentence test"), Z. Audiol. **38**, 44–56.

Wagener, K., Brand, T., and Kollmeier, B. (**1999b**). "Entwicklung und Evaluation eines Satztests für die deutsche Sprache III: Evaluation des Oldenburger Satztests" ("Development and evaluation of a German sentence test III: Evaluation of the Oldenburg sentence test"), Z. Audiol. **38**, 8695.

Watkins, A. J. (**2005a**). "Perceptual compensation for effects of echo and of reverberation on speech identification," Acta Acust. united Ac. **91**(5), 892–901.

Watkins, A. J. (**2005b**). "Perceptual compensation for effects of reverberation in speech identification," J. Acoust. Soc. Am. **118**(1), 249–262.

Zahorik, P. (**2002**). "Direct-to-reverberant energy ratio sensitivity," J. Acoust. Soc. Am. **112**(5), 2110–2117.

Zahorik, P., and Brandewie, E. J. (**2016**). "Speech intelligibility in rooms: Effect of prior listening exposure interacts with room acoustics," J. Acoust. Soc. Am. **140**(1), 74–86.