Assessment of speech perception based on binaural room impulse responses

Master thesis

für die Prüfung zum Master of Science (M.Sc.) im Studiengang
Audiokommunikation und -technologie
an der Technischen Universität Berlin,
Fakultät I – Geisteswissenschaften
Vorgelegt von: Omid Kokabi
Matrikelnummer: 362718

Erstgutacher: Prof. Dr. Stefan Weinzierl
Zweitgutachter: Fabian Brinkmann

15.03.2018

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ohne unerlaubte fremde Hilfe und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

Berlin, den

_____

Omid Kokabi

Abstract

The two most important aspects in binaural speech perception — better-ear listening and spatial release from masking — can be modelled well with current prediction frameworks operating on binaural room impulse responses (BRIRs). To incorporate effects of reverberation, a model extension was recently proposed, splitting the BRIR into an early, useful and a late, detrimental part, before fed into the prediction framework. In a more recent work (Leclère et al., 2015) a relation 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 for four simulated rooms with systematically varied levels of reverberation and a constant room geometry. By linking the applied splitting time to room acoustic parameters, the mean prediction error with the binaural model by Jelfs et al. (2011) could be reduced by about 1 dB.
Further, the prediction accuracy with pseudo-binaural signals, which can be captured with existing microphone arrays allowing for the evaluation of different head orientations in a post-processing step, was tested. Results were close to predictions with BRIRs, illustrating its suitability for practical assessment of binaural speech perception in existing rooms.
All relevant data generated in the course is this work is publicly available from http://dx.doi.org/10.14279/depositonce-6725.

Zusammenfassung

Aktuelle raumimpulsantwortbasierte Sprachverständlichkeitsmodelle sind in der Lage, die beiden wichtigsten Wahrnehmungsaspekte – better-ear listening und spatial release from masking – mit hoher Genauigkeit nachzubilden. Um den Einfluss von Nachhall in diesen Modellen zu berücksichtigen, wurde zuletzt eine Modellerweiterung vorgeschlagen, welche die binaurale Raumimpulsantwort (BRIR) zunächst in einen frühen, nützlichen und einen späten, nachteiligen Anteil aufteilt, bevor sie in das Prädiktionsmodell eingeht. Auf der Grundlage von Hinweisen, dass sich eine solche Zeitgrenze und Raumeigenschaften auf den resultierenden Prädiktionsfehler auswirken (Leclère et al., 2015) wurde in der vorliegenden Arbeit der systematische Einfluss dieser Variablen untersucht. Hierzu wurden Sprachrezeptionsschwellen (engl. SRTs) in Ruhe in vier simulierten Räumen gemessen. Die Räume wurden systematisch hinsichtlich ihrer mittleren Nachhallzeit variiert, während die Raumgeometrie konstant gehalten wurde. Durch eine Erweiterung des binauralen Sprachverständlichkeitsmodells nach Jelfs et al. (2011) um raumabhängige Zeitgrenzen konnte im Ergebnis der mittlere Prädiktionsfehler um ca. 1 dB verringert werden.
Darüberhinaus wurde untersucht, inwieweit sich pseudo-binaurale Signale als Ausgangssignal bestehender Mikrofonanordnungen, die eine Auswertung beliebiger Kopforientierungen aus einer einzigen Aufzeichnung erlauben, zur Vorhersage von Sprachverständlichkeit eignen. Hierbei konnten Prädiktionsfehler in vergleichbarer Größenordnung zur Vorhersage mit BRIRs beobachtet werden, was für eine grundsätzliche Eignung dieser pseudo-binauralen Signale zur Vorhersage binauraler Sprachverständlichkeit in bestehenden Räumen spricht.
Alle relevanten, im Zuge dieser Arbeit generierten Daten sind öffentlich zugänglich unter http://dx.doi.org/10.14279/depositonce-6725.

Contents

1 Introduction **4**
- 1.1 Better-ear listening and binaural unmasking . . . 4
- 1.2 Binaural time/ spectral domain models . . . 5
- 1.3 Binaural envelope domain models . . . 5
- 1.4 The U/D approach . . . 6
- 1.5 Head orientation and speech perception . . . 6

2 Methods **7**
- 2.1 SRT measurement . . . 7
  - 2.1.1 Subjects . . . 7
  - 2.1.2 Procedure . . . 7
  - 2.1.3 Acoustic setup & stimuli generation . . . 8
- 2.2 SRT prediction . . . 10
  - 2.2.1 General prediction procedure . . . 10
  - 2.2.2 Calculation of optimum, room/ receiver dependent U/D-limits . . . 10
  - 2.2.3 Additional stimuli for SRT prediction . . . 11
  - 2.2.4 Statistical evaluation . . . 11

3 Results **11**
- 3.1 Prediction accuracy with fixed U/D limits . . . 11
- 3.2 Prediction accuracy with room-/ receiver dependent U/D limits . . . 13
  - 3.2.1 Calculated optimum, room dependant U/D limits . . . 13
  - 3.2.2 Prediction of optimum room/ receiver dependant U/D limits from room acoustical parameters . . . 14
  - 3.2.3 Resulting prediction error with room/ receiver dependent U/D limits . . . 14
- 3.3 Prediction accuracy with pseudo-BRIRs . . . 15

4 Discussion **16**
- 4.1 Prediction accuracy with fixed U/D limits . . . 16
- 4.2 Rationale for a room dependent U/D limit . . . 18
- 4.3 Prediction accuracy with pseudo-BRIRs . . . 18
- 4.4 Validity of the simulated sound field representations . . . 19

5 Conclusion **19**

1 Introduction

The most important perceptual mechanisms in daily speech perception in natural environments with competing noise sources can be described by better-ear listening and binaural unmasking of spatially separated sources (Middlebrooks et al., 2017). With fluctuating noise sources, e.g. different talkers, differences between fundamental frequencies and dip listening further help to segregate a specific speaker from the background.
For the prediction of speech perception based on the two former mechanisms, different models have been developed, with which experimental observations of both effects could be well reproduced. Among these models, the most promising ones can roughly be categorized into two groups: Models operating in the time/ spectral domain with the two most successful (Culling et al., 2013) models by Beutelmann et al. (2010) and Jelfs et al. (2011) and models operating in the envelope domain, e.g. the model by Chabot-Leclerc et al. (2016).
The model input is typically either a binaural stream of the speech/ masker signal apparent at both ears or a binaural room impulse response (BRIR), describing the transfer path between the speech/ masker source and the binaural receiver.
In typical rooms, the perceived speech signal is a combination of the direct signal, multiple distinct room reflections and late diffuse reflections, i.e. reverberation. While distinct reflections arriving in a short time interval after the direct sound are generally considered to improve speech perception (Bradley et al., 2003), reverberation is known to have a deteriorating effect by increasing the temporal masking of speech sounds on subsequent ones and reducing the depth of temporal modulation inherent in running speech.
To account for the effects of reverberation on speech perception in models operating in the time/ spectral domain, different model extension have been proposed in Rennies et al. (2011) with one splitting the BRIR into an early useful and a late detrimental part – referred to as U/D approach in the remainder of this document – which are then fed separately into the model.
A more recent work (Leclère et al., 2015) has pointed out, that prediction accuracy with the binaural model by Jelfs et al. (2011) and the U/D approach is affected by the temporal limit applied in the splitting process. Best prediction performance could only be achieved with room-/ receiver dependent U/D limits, which was considered as a downside of this approach limiting its applicability for speech intelligibility prediction in e.g. room acoustic evaluation. A connection between the respective temporal limits and specific room-/ receiver dependent aspects has not been drawn.
The present work thus tries to fill this gap by investigating the correlation between room/ receiver dependent aspects, the applied temporal limit in the U/D approach and the resulting prediction accuracy.

1.1 Better-ear listening and binaural unmasking

With a localized sound source outside the median plane, interaural differences in time and level (interaural time/ level differences, ITD/ ILD) between both ears can be observed. ITDs are subject to the direct path length differences from the source to both ears. ILDs occur due to head shadowing and the ears spatial sensitivity as a function of angle of incidence with the highest sensitivity being roughly at 40 - 60° azimuth on the respective ipsilateral side.
With spatially separated target and masker sources (while the latter might also be repetitions of the former, i.e. target energy reflected by room surfaces) interaural differences in signal-to-noise ratios (SNR) at the listeners' ears can be observed due to different target and masker ILDs. These interaural SNR differences can be evaluated by the auditory system in a better-ear fashion, i.e. information is primarily extracted from the ear signal with the higher SNR. This evaluation is done over the entire frequency range rather than on a per-band level (Edmonds and Culling, 2006). Better-ear listening can be considered as a monaural mechanism.
Different target and masker ITDs further help to segregate between the two reducing the strength of the masking effect on the speech target (Kock, 1950). This process is referred to as binaural

unmasking, indicating its binaural nature as interaural information (ITDs) is required as input to the auditory system to make use of this effect. Binaural unmasking is typically modelled by implementations following the Equalization-Cancellation (EC)-theory (Durlach, 1963). The EC model is designed as a "black box" model of the auditory system, meaning that all signal transformations from the outer ear to nerve impulses are ignored. The basic concept of EC assumes, that the auditory system tries to eliminate masking components in the total signal (target + competing masking sources) by transforming one ear signal in a way that its' masking components match the masking components apparent at the other ear (= Equalization stage), followed by a subtraction of one ear signal from the other (= Cancellation stage) improving the SNR when target and masker differ in ITD and/ or ILD.
Although there is no clear interpretation of how both mechanisms are exactly combined in the auditory system, there is a general agreement about (partial) additivity.

1.2 Binaural time/ spectral domain models

The two most successful time and frequency domain models which can account for better-ear listening and binaural unmasking are the Oldenburg model (Beutelmann and Brand, 2006) and the Cardiff model (Lavandier and Culling, 2010).
Both models have been designed to mimic observations on both perceptual mechanisms addressed rather than stating a correct psychophysical implementation of the auditory system. They both combine a better-ear evaluation and an EC-based modeling stage for binaural unmasking.
The Oldenburg model,initially developed by Beutelmann and Brand (2006) and revised in Beutelmann et al. (2010) combines a gammatone bandpass filter bank and an EC-stage per band serving as an input to a Speech Intelligibility Index (SII) model (ANSI, 1997) calculation. For interpretation with measured speech reception thresholds (SRTs), defined as the signal-to-noise ratio corresponding to 50% intelligibility, the calculated SIIs are mapped to SRTs based on a psychometric function matching. To allow for fluctuating maskers, the model calculates the SII/ SRT as a mean value across multiple frames of the time signals. In both the original and the revised version, the entire speech signals was considered as useful ignoring the deteriorating effect of reverberation on speech reception.
The Cardiff model,initially developed by Lavandier and Culling (2010) and revised in Jelfs et al. (2011) combines a gammatone bandpass filter bank, a SNR evaluation on a per band basis per ear and a EC-based implementation of a binaural unmasking stage (Culling et al., 2005). Both the SNR and the binaural unmasking stage are combined and frequency weighted according to the SII importance weighting. The model output is an effective target-to-masker ratio in dB, which can be compared to measured SRTs by e.g. average matching. Though it was tested in Lavandier et al. (2012) with multiple stationary masking sources and reverberation, the original as well as the revised version still featured the same limitation as the Oldenburg model considering the entire speech signal as useful for speech perception limiting its applicability for room acoustic evaluation with non-negligible levels of reverberation.

1.3 Binaural envelope domain models

The two best known binaural models operating in the envelope domain are the binaural extension of the Speech Transmission Index (STI) model (van Wijngaarden and Drullman, 2008) and the envelope-power spectrum (EPSM) model by Jørgensen et al. (2013).
The binaural STI has been designed to extend the original STI (IEC 60268-16, 2011) by a better-ear evaluation and a similarity evaluation of both ear signals via cross-correlation while keeping the model complexity as simple as possible to improve is applicability for room acoustic evaluation. The binaural STI handles many aspects affecting monaural speech perception including non-negligible levels of reverberation. The models' over-simplification of known auditory processes e.g. ignoring a closer modeling of the bandpass properties of the basilar membrane by means of e.g. ERB-spaced gammatone filters, ignoring aspects of binaural unmasking etc. is considered as a downside.
The EPSM, initially presented by Ewert and Dau (2000) in the context of amplitude modulation

detection has been extended by Jørgensen and Dau (2011) and revised in Jørgensen et al. (2013) for monaural speech intelligibility prediction. An extended model combining a better-ear-EPSM evaluation and an EC-based stage to account for binaural unmasking has been presented in Chabot-Leclerc et al. (2016). As reverberation affects the signals envelope, the EPSM approach can in principle account for the effects of non-negligible levels of reverberation. However, the models complexity and the implemented parameter fitting can be considered as a downside.

1.4 The U/D approach

In the present evaluation, the Cardiff model with its latest revision by Jelfs et al. (2011) has been chosen to be combined with the U/D approach mainly due to a) its computational efficiency (compared to all signal-based approaches), b) its open source availability (compared to the Oldenburg model) and c) the fact, that no parameter-fitting is involved in the entire process. The general concept of the U/D approach classifying early room reflections as useful and late reflections as detrimental can also be found in many room acoustic quantities (Clarity, definition, direct-to-reverberant energy ratio, useful-detrimental ratio (Bradley, 1986) etc.). Throughout the literature different limits for the time reflections can still be considered useful are used ranging from U/D = 35ms (Bradley, 1986) to U/D = 95ms (Lochner and Burger, 1964). The U/D approach has been introduced – besides a definition-based and an MTF-based approach – as a potential extension to the Oldenburg time/ spectral domain model in Rennies et al. (2011) to account for the effects of non-negligible levels of reverberation. After the U/D split, early (= useful) and late(= detrimental) signal components are separately fed into the model. The U/D approach (tested with two U/D limits: 50ms and 100ms) showed slightly better performance than the other two candidates improving the overall prediction accuracy. These results were confirmed with measured SRT data by Warzybok et al. (2013) with a more simplified sound field comprising only the direct signal and one lateral reflection as a function of reflection delay (Rennies, 2014). The generalisability of the U/D approach has further been tested by its implementation into the Cardiff time/ spectral domain model by Leclère et al. (2015) as a function of U/D limit and time/ shape of the transition between the early and late part used for the splitting process. The prediction accuracy with this model on the data by Rennies et al. (2011) and Lavandier and Culling (2008) could be improved with an U/D extension however for best prediction performance the applied U/D limit was found to be room-dependent limiting its generalizability.
In the present work the prediction accuracy as a function of the temporal limit applied in the U/D approach is evaluated for correlations with room-/ receiver dependent aspects. Therefore, SRTs in quiet were measured and predicted for a virtual room, whose room acoustical properties are systematically varied. In Quiet refers to the condition without additional masking noise sources. The SRT in quiet is analogous to the sound pressure level in dB_{SPL} required for 50% correctly understood words. For predictions, the Cardiff binaural intelligibility model in its latest revision by Jelfs et al. (2011) with the U/D approach implemented by the author is used. This is to a) show potential compensation for the mentioned drawback of the U/D approach b) to improve the models' applicability to room acoustic evaluation and c) to highlight and quantify remaining deviations between observed and modelled perceptual mechanisms.

1.5 Head orientation and speech perception

Head orientation can significantly affect speech perception, as both binaural unmasking and better-ear aspects can be improved within a given acoustic setup with an optimized head orientation.
In an anechoic environment, a SRT benefit due to an optimized head orientation (HOB = head orientation benefit) could be observed with up to 8 dB (Grange, 2016). Further, predictions with the Cardiff binaural intelligibility model in its latest revision by Jelfs et al. (2011) accurately reproduced the observed HOBs. In a more realistic restaurant environment with moderate reverberation, an HOB of about 3 dB could be observed (Grange and Culling, 2016), constituting a noticeable difference in intelligibility (McShefferty et al., 2015).
In the assessment of the suitability of room acoustic design for speech reproduction, it seems logical to incorporate these effects. However, for practical measurements in existing rooms, setting up and

rotating a dummy head at one or multiple receiver positions is tedious and time-consuming, limiting its applicability to evaluations based on room acoustic simulations.
In Bernschütz (2016) it has been shown, that BRIRs can also be calculated from spherical microphone array responses allowing for the incorporation and evaluation of different head orientations from a single microphone array measurement in a post-processing stage. To capture (and reproduce) a high spatial resolution, a dense grid of receivers is required to avoid aliasing. For e.g. HRIR data (= BRIR data within an anechoic environment) with persistent localization cues a spherical harmonics order of N = 35 is said to be required corresponding to a spherical grid of 1300 sensors due to the detailed structure of the human head and pinna causing fine spatial structures in the resulting sound field. This is not feasible for practical room acoustic applications.
While localization and speech perception partially make use of the same interaural cues (ITD, ILD) and spatial separation is known to affect speech perception, there is evidence, that the processing of the two is handled in parallel auditory pathways (Ahveninen et al., 2006). Further, due to the band limited nature of natural speech (with highest frequencies up to 10 kHz) localization cues for frequencies exceeding this range can be considered irrelevant. Hence, it can be questioned whether persistence of localization cues is an appropriate criterion for spatial sampling and reconstruction in the context of speech perception and prediction.
In the present work, it will thus be determined, whether two relatively simple and low-cost spatial measurement approaches without persisting localization cues, i.e. introducing manipulation on both ILD and ITD information can be used to produce estimations of binaural signals, which are sufficient to allow evaluation of HOBs in a postprocessing step.
Therefore, binaural signals based will be estimated by means of room acoustic simulation with two different approaches: a) A low-order sound field decomposition approach with the routines presented in Tervo et al. (2013) and b) ear signals based on the Motion Tracked Binaural (MTB) approach (Algazi et al., 2004).
The estimated pseudo-binaural signals shall serve as input to the Cardiff binaural intelligibility model in its latest revision by Jelfs et al. (2011) with the U/D approach implemented by the author. The predicted accuracy with these pseudo-binaural signals will be compared to the accuracy with distinctive binaural signals.
Though it is clear, that both approaches will produce binaural signal representations deviating from distinctively simulated binaural signals with e.g. a dummy head, it shall be assessed, whether these estimated signals feature a sufficient spatial resolution to be used for binaural speech perception prediction with an existing binaural model.

2 Methods

2.1 SRT measurement

2.1.1 Subjects

18 native German speakers (13 male/ 5 female, age mean = 30.4, age standard deviation = 2.9) without reported hearing impairment. All subjects participated in the tests on a voluntary basis. Except for two, all subjects had experience with psychoacoustic listening tests.

2.1.2 Procedure

For the acoustical conditions considered, SRTs were measured in Quiet. For the measurement of SRTs in Quiet, the Oldenburg sentence (OLSA) test (Kuehnel et al., 1999; Wagener et al., 1999b,a) was used. The OLSA test has been developed to measure SRTs with or without additional competing masking sources. The test sentences consist of five words at a natural speech rate with a fixed syntax (name - verb - number - adjective - object) but unpredictable semantics. The participants' task is to repeat the words of the test sentence he/she understood. Depending on the number of correctly understood words, the experimenter adaptively adjusts the signal level according to the OLSA manual (HörTech gGmbH, 2011) (step size ranging from ±1 dB to ±3 dB for sentences 2 - 5

and ±1 dB to ±2 dB for sentences 6 - 31) for the subsequent sentence converging to a threshold of 50% correctly understood words (= SRT) within a set of 30 test sentences per condition. The entire sentence corpus comprises 600 different sentences, corresponding to 20 sets of 30 sentences each.
For the measurement in quiet, individual pure tone audiogram data was additionally measured for frequencies between 125 Hz and 8 kHz according to DIN EN 60645-1 (2015) to compensate SRT results for individual hearing insensitivities. In Rennies et al. (2011) a significant correlation between pure tone thresholds and measured SRTs in Quiet even for listeners with normal hearing capabilities with dB HL (Hearing level, (DIN EN ISO 8253-1, 2011)) $<$ 20 dB could be observed. By compensating the measured SRTs for the individual hearing insensitivities, the between-subject variability in measured SRTs in Quiet can thus be significantly reduced. The individual octave-based dBHL values were therefore transformed into a sum-level adaptation, which was subtracted from the measured SRTs.
Four test conditions with systematically varied acoustic conditions discussed further below with 30 sentences per condition were prepared for every participant.
The tests were performed at the Institute of Fluid Mechanics and Engineering Acoustics (ISTA) at Technische Universität Berlin. The participant was positioned in the ISTAs' hemi-anechoic chamber with the conductor located in the adjacent control room. The stationary room noise level in the hemi-anechoic chamber was 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 closed, circumaural Beyerdynamic Headphones DT770Pro with headphone equalization, the latter provided within Brinkmann et al. (2017b,a).
The headphone playback level was 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. The headphone was connected to a Focusrite Scarlett 18i20 USB interface to a windows Laptop running MATLAB R2015B The MathWorks (2013), located in the control room. For intercom purpose, the conductor used a Omnitronic GMTS100 intercom terminal with a gooseneck microphone.
Both tests were implemented in MATLAB R2015B, with the paradigm according to DIN EN ISO 8253-1 (2011) for the audiogram test and HörTech gGmbH (2011) for the SRT measurement.
For the audiogram test, the participant responded by him/herself via a generated MATLAB graphical user interface (GUI). For the SRT measurement, the participant made a spoken response via a talkback microphone. Based on the number of correctly understood words, the experimenter applied the manual level adaptation for the subsequent sentence in the test script. After completion of one condition, there was a short pause before the next condition was tested. The different conditions were tested in random order. To familiarize the participants with the task and the stimuli, a training was performed prior to the actual tests.
The positioning of circumaural headphones over the listeners ears in psychoacoustic testing can significantly affect the results due to differences in the way sound waves are scattered at the ears fine structure, typically rising with rising frequency. This can both affect the magnitude of the stimulus level resulting in e.g. an increase in threshold variability in a pure tone audiometry test (Paquier et al., 2012) and introduce additional audible coloration between different headphone positions (Paquier and Koehl, 2015).
Both the measurement of pure tone thresholds and SRTs in Quiet can be considered critical to absolute stimulus level. Hence, to minimize the variance in sound representation level and/ or coloration due to repositioning the headphone, all participants were asked before the test to not remove/ reposition the headphone after initial positioning until the end of both the audiometry and SRT test. The entire test with instruction, training and filling out the questionnaire took about 70 min per participant.

2.1.3 Acoustic setup & stimuli generation

The BRIRs for the different test conditions were simulated on the concept of geometrical acoustics within the RAVEN software package (Schröder and Vorländer, 2011). For a list of applied simulation settings, refer to the published dataset (Kokabi et al., 2018).
The acoustic environment for which BRIRs were generated was based on the geometry of an

existing, medium sized auditorium[1]. The room can be considered as a shoebox room featuring diffusing wall/ceiling elements with a stage and a lowered audience area. Two acoustical surface properties were distinguished: Seating (audience area) and residual (all other planes) .
The perceptual impression of a room is mainly characterised by the reflection pattern (temporal structure and reflection amplitudes) arriving at the listeners ears. While the temporal structure is subject to the geometrical conditions (room dimensions, source and receiver positions), the amplitudes of the individual reflections are (aside from source and receiver directivity and air absorption) subject to the acoustic absorption and scattering properties of the rooms surfaces. The generation of the different perceptual conditions was done by systematically varying both the room dimensions and the surface properties independently. The former was done by uniformly scaling the room model, the latter by uniformly scaling the respective absorption coefficients, resulting in 16 different room configurations (four volumes $V = 500\ m^3$, $1000\ m^3$, $2000\ m^3$ and $4000\ m^3$ with each one featuring four reverberation times $T_{20,m} = 0.5$ s, 1.0 s, 2.0 s and 4.0 s with the subscript m denoting the mean across the 500 Hz and 1 kHz octave). To limit the number of test conditions and hence the test duration per participant, an informal listening test was initially performed, wherein four test conditions had been chosen for the actual test, covering the largest possible range of realistic reverberation scenarios (a room volume of $4000\ m^3$ is rarely found in reality with an $T_{20,m}$ of 0.5 s).
The four different conditions finally used for SRT measurement and prediction were generated with a fixed volume of $1.000\ m^3$ (Length: 17.3 m, width: 11.2 m, height: 5.2 m) and varying absorption properties of both materials, maintaining a realistic frequency dependence (higher T_{20} at lower frequencies and vice versa) typically observed in rooms in all test conditions based on the recommendations in DIN 18041 (2016).
The four test conditions differ in level of reverberation with mean reverberation times of $T_{20,m} =$ 0.5 s, 1.0 s, 2.0 s and 4.0 s with identical temporal reflection structures. The applied absorption and scattering coefficients as well as the resulting, frequency dependent reverberation statistics are shown in Kokabi et al. (2018).
BRIRs were calculated for every test condition for one source, located at the centre of the stage and one binaural receiver located at the audience area. The distance between source and receiver was approximately d ≈ 9 m, equivalent to about three times the critical distance with the applied source directivity, see below, at the lowest reverberation level tested. The exact geometrical conditions can be seen in Figure 1.

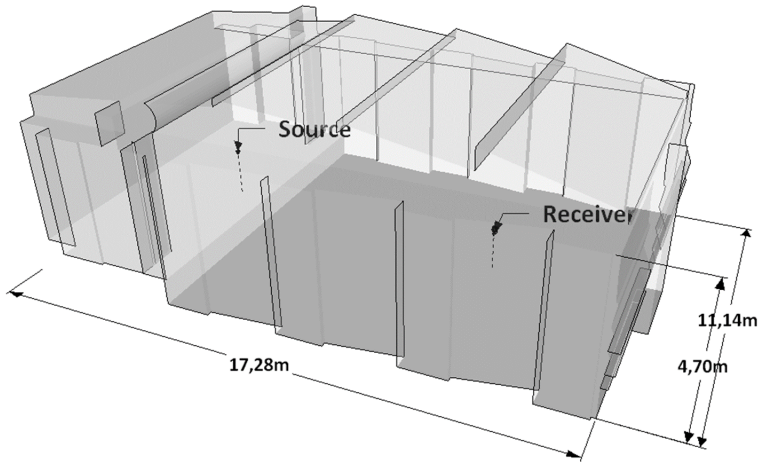

Figure 1: 3d room model

[1]Kammermusiksaal, Musikhochschule Köln Link: https://www.hfmt-koeln.de/veranstaltungen/veranstaltungsorte/konzertsaal-aachen.html

For the sound source, directivity data of a male singer (average directivity factor Q of 1.5 for 500 Hz and 1 kHz octaves) was used supplied within RAVEN. For the receiver, measured full spherical directivity data of the FABIAN HATS (Lindau et al., 2006) with an azimuth/ elevation resolution of 2° was used, which was measured in the anechoic chamber of the Carl von Ossietzky University Oldenburg (Brinkmann et al., 2017b). The HATS directivity data was converted into .daff format (Wefers, 2010) for use within RAVEN.
Binaural auralisations of the OLSA sentence corpus were calculated via convolution with the generated BRIRs within MATLAB.

2.2 SRT prediction

2.2.1 General prediction procedure

The generated BRIRs were applied to the binaural intelligibility model by Jelfs et al. (2011) provided as MATLAB code within Søndergaard and Majdak (2013) with a temporal U/D-classification as shown e.g. in Rennies et al. (2011), implemented by the author. For the latter, each BRIR was multiplied in the time domain with two windows: An early window and a late window. The early window comprised a flat (weight = 1) part up to the considered U/D limit in milliseconds (starting from the time of arrival) and a linear fadeout with a length of 1 ms. The late window started with a linear fade-in of length 1 ms at the considered U/D-limit and continued with a flat part (weight = 1) until the end of the respective BRIR, so that both windows summed up to 1. The early useful part was considered as the speech target and the late detrimental part as a masker. Both were separately fed into the model.
The model output is a signal-to-noise ratio in dB predicting the benefit of having a head with two ears instead of having an omnidirectional microphone in the center of the head with the head absent. The predicted benefit was converted to an SRT by a multiplication by -1 and scaling every single 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. This procedure was also used e.g. in Jelfs et al. (2011). It is important to note here, that with this method, the model is only able to predict relative differences in SRT between conditions (or receivers) due to the average-matching of measured and predicted data. In addition, the prediction accuracy with fixed and room-/ receiver dependent U/D-limits was tested with an external dataset (provided BRIRs calculated within CATT-Acoustic v8 (Dalenbäck, 2008) featuring KEMAR (Burkhard and Sachs, 1975) head related transfer function (HRTF) data and documented SRTs in Quiet) from (Rennies et al. (2011), conditions "S0" and "S90") where SRTs in Quiet were measured for two different test setups (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 levels of about 2.0s. The two test conditions each feature four source-receiver combinations differing in source-receiver distance d, ranging from d = 0.5 m to d = 13.0 m for condition S0. In condition S90, the fourth receiver has the same distance to the source as receiver three (d = 3.5 m) with different absolute positions in the room. For details, please refer to the original publication. This dataset is referred to as *RS11* in the remainder of this document.
To assess the suitability of pseudo-binaural signals as a substitute for BRIRs, additional stimuli were generated for SRT prediction, see section 2.2.3. These were applied to the model in the same way as the BRIRs. BRIRs and pseudo-BRIRs are publicly available (Kokabi et al., 2018).

2.2.2 Calculation of optimum, room/ receiver dependent U/D-limits

For the analysis of optimum, room-/receiver dependent U/D limits, the following method was applied: SRT predictions for every participant were calculated with the method depicted above, whereby for every condition (= BRIR) 19 different U/D-limits ranging from 20 ms to 200 ms with 10ms steps were used resulting in 194 different predicted sets of SRTs per participant.
Next, all U/D-combinations leading to a mean absolute error (MAE) < 1 dB over all four conditions per participant were used. From this subset of U/D-combinations, the mean across all U/D-limits

per condition was calculated. By doing so, the average U/D-shift between the test conditions leading to a minimum prediction error was calculated. The idea of this method is to capture the general trends of shifts in U/D-limits between conditions leading to a minimum prediction error. The so calculated limits are a more robust estimation of optimum U/D-limits per condition as e.g. considering only the U/D combination with the smallest absolute error.
To correlate the obtained optimum U/D limits with common room acoustical parameters, a linear/ multiple regression analysis was performed with room acoustical parameters as independent variables and the optimum U/D-limits as dependent variable. In Ellis et al. (2015), it has been pointed out, that aspects of binaural dereverberation in speech perception can be correlated to monaural acoustic parameters according to distance between source and receiver and binaural parameters assessing the similarity between both ear signals. Thus, the following parameters are used as potential predictors in the regression analysis: $C80_m$ and D/R as monaural predictors and $IACC_m$ as a binaural predictor, with the subscript m denoting the mean across the 500 Hz and 1 kHz octave values. The results from the regression analysis were used for the finally calculated optimum U/D limits from the room acoustical predictors.

2.2.3 Additional stimuli for SRT prediction

For SRT predictions, additional pseudo-binaural signals were simulated based on two different approaches: a) Based on the motion tracked binaural (MTB) approach (Algazi et al., 2004), pseudo-binaural signals are taken from two sensors flush mounted on opposite sides on the equatorial line of a rigid spherical microphone array with a diameter of d = 17.6 cm (Ackermann et al., 2016). To calculate the array response, full spherical directivity data of the spherical rigid microphone array was measured in the anechoic chamber of the ISTA at Technische Universität Berlin with the system described in Fuß et al. (2015). The array directivity data was converted into .daff format for use within RAVEN. The array output signals are referred to as *MTB-BRIRs* in the remainder of this document.
b) Based on B-Format responses calculated within RAVEN, binaural signals were estimated with the spatial decomposition method approach (SDM) described in Tervo et al. (2013) with the routines available as MATLAB scripts (Tervo, 2016). The concept of this method can be described as follows: From multichannel room impulse response (RIR) data and knowledge about the locations of the sensors in the array used for capturing this data, estimations of the incident angles of individual reflections of the RIR are calculated. This spatial information is used for assigning the multichannel RIR data to an arbitrary setup of virtual secondary sources distributed around a binaural sensor. Finally, binaural signals are calculated as a weighted (with HRTF data) sum over all contributions of the virtual secondary sources. The SDM routines offer multiple parameters (temporal resolution of the incident angle estimation, quantity and distribution of the secondary sources, length of the smoothing window applied to the estimated incident angles of individual reflections, HRTF data used in the binaural summation process etc.) which can be set by the user. For a list of SDM parameters applied in the present evaluation, refer to Kokabi et al. (2018). The so calculated binaural signals are referred to as *SDM-BRIRs* in the remainder of this document.

2.2.4 Statistical evaluation

The assessment of the prediction accuracy with the binaural model for both fixed and room-/ receiver dependent U/D-limits was performed by calculating the mean absolute error (MAE) in dB over all conditions between measured and predicted SRTs.

3 Results

3.1 Prediction accuracy with fixed U/D limits

The residual room noise in the hemi-anechoic chamber during both listening tests can be considered uncritical (LAeq $<$ 25 dB), especially as the closed circumaural headphones further damp any

acoustic signal penetrating to the ear. All participants can be considered as normal hearing with measured dBHL DIN EN ISO 8253-1 (2011) values between −10 dB and +20 dB.
Measured mean SRTs across all participants for all four test conditions are shown in Figure 2. Additionally, predicted mean SRTs across all participants with two fixed U/D-limits (50 ms and 100 ms) with the method described above are depicted. Standard error both for measured and predicted SRTs are shown as vertical bars.

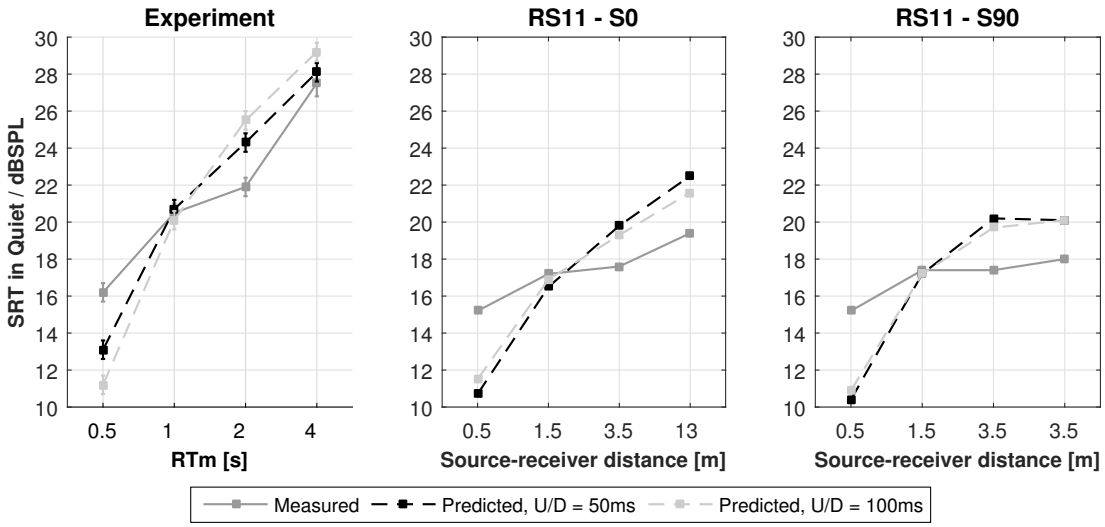

Figure 2: Measured and predicted SRTs with fixed U/D limits

To test for systematic differences in measured SRT data between conditions, a one-factorial repeated measures ANOVA was applied with Post hoc Bonferroni correction with a significance level of 0.05. Mauchly's test indicated that the assumption of sphericity had been violated, $\chi^2(5) = 27.9, p < .01$, therefore degrees of freedom were corrected using Greenhoise-Geisser estimates of sphericity ($\epsilon = .48$). Results reveal, that there was a significant effect of level of reverberation on the measured SRTs, $F(1.4, 24.4) = 206.3, p < .001$.
As can be seen for the experimental data in the left panel of Figure 2, both measured (solid line) and predicted SRTs (dashed lines) with fixed U/D-limits increase with increasing level of reverberation. Note: For interpreting the figure, please keep in mind that only relative SRT differences between conditions can be calculated with the prediction method depicted above, which can be deduced from the gradient of the line connecting two test conditions.
Comparing the prediction accuracy of both fixed U/D-limits for the experimental data (left panel in Figure 2) it can be seen, that the prediction of the experimental data with U/D = 50 ms (MAE = 1.9 dB) is slightly better than with U/D = 100 ms (MAE = 2.9 dB). For the RS11 data (middle and right panel in Figure 2), it can be seen, that the prediction of the RS11 data with U/D = 50 ms (MAE = 2.6 dB) is slightly worse than with U/D = 100 ms (MAE = 2.1 dB).
Comparing measurement and prediction for the experimental data (left panel in Figure 2), the increase in SRT with increasing reverberation is generally overestimated (predicted increase in SRT between conditions one and two and two and three higher than the measured SRT increase) by the prediction model in the low and medium reverberant conditions with both fixed U/D limits. Between condition three and four (medium to high level of reverberation), the model slightly underestimates the measured increase in SRT (predicted increase in SRT between conditions three and four lower than the measured SRT increase).
For the RS11 data, the increase in SRT with increasing distance for medium and short distances (increase in SRT between conditions one and two and conditions two and three) is overestimated by the prediction model with both fixed U/D limits. Between conditions three and four, the model pre-

dicts the SRT increase quite good (under-/ overestimation < 1dB). For condition S90, the measured and predicted SRTs between condition three and four are quite constant, as the distance between source and receiver – contrary to condition S0 – is also constant (d = 3.5 m).
The under-/ overestimation of SRT increase with the experimental data and RS11 data and the prediction model with the fixed U/D-limits is depicted in Figure 3.

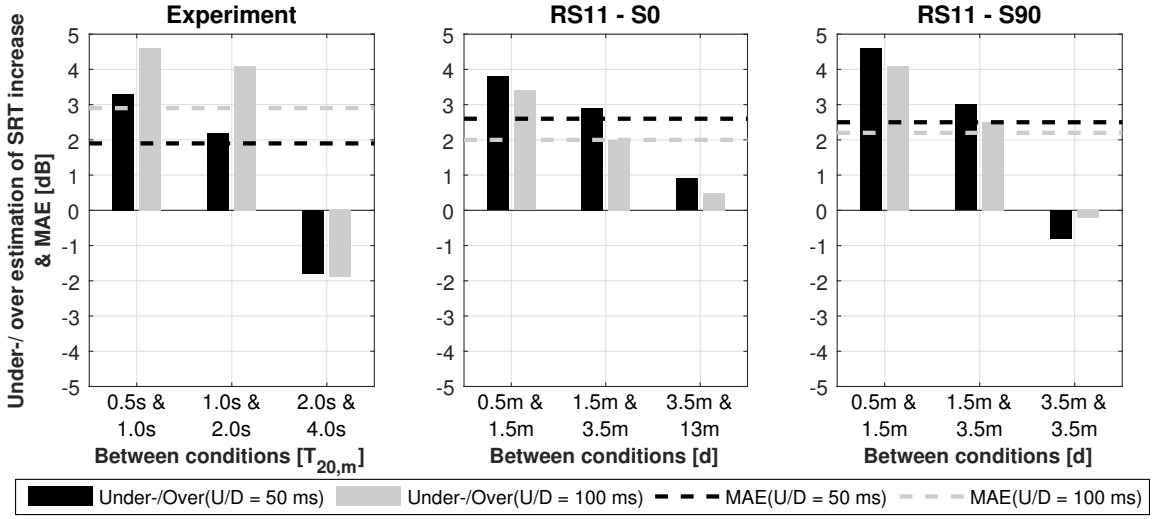

Figure 3: Under-/over estimation of SRT increase with fixed U/D limits

3.2 Prediction accuracy with room-/ receiver dependent U/D limits

3.2.1 Calculated optimum, room dependant U/D limits

With the experimental and RS11 data, the optimum room-/ receiver dependent U/D-limits (with standard deviation) averaged across all participants as shown in the left column of Table 1 were calculated with the method depicted above.

	Condition	Optimum U/D limits, mean (Std dev) [ms]	Room acoustic parameters			U/D limits predicted by		
			D/R [dB]	$C80_m$ [dB]	$IACC_m$	D/R [dB]	$C80_m$ [dB]	$IACC_m$
Experiment	$T_{20,m}$ = 0.5 s	59 (8)	-1.6	6.4	0.43	62	61	57
	$T_{20,m}$ = 1.0 s	90 (14)	-7.4	-0.7	0.22	97	97	99
	$T_{20,m}$ = 2.0 s	142 (11)	-10.9	-4.9	0.08	118	118	127
	$T_{20,m}$ = 4.0 s	122 (24)	-14.0	-8.7	0.06	136	137	131
RS11 - S0	d = 0.5 m	48 (25)	3.2	4.6	0.65	33	70	12
	d = 1.5 m	122 (37)	-3.4	-0.5	0.29	73	96	85
	d = 3.5 m	162 (28)	-9.5	-2.1	0.13	109	104	117
	d = 13.0 m	162 (26)	-20.8	-3.7	0.10	177	112	123
RS11 - S90	d = 0.5 m	35 (15)	2.8	3.5	0.54	35	76	34
	d = 1.5 m	125 (34)	-5.5	-2.2	0.28	85	104	87
	d = 3.5 m	171 (22)	-12.3	-2.5	0.21	126	106	101
	d = 3.5 m	169 (24)	-15.2	-2.5	0.26	143	106	91

Table 1: Optimum U/D limits, room acoustic parameters and predicted U/D limits

As can be seen from the left column of Table 1, the optimum room-/ receiver dependent U/D-limits leading to a MAE $<$ 1dB increase with increasing level of reverberation (experimental data) and with increasing distance from the source (RS11 data), respectively.

3.2.2 Prediction of optimum room/ receiver dependant U/D limits from room acoustical parameters

For the experimental and RS11 data, the room acoustic parameters as shown in the middle three columns of Table 1 were calculated with MATLAB and routines provided within the itaToolbox (Dietrich et al., 2010). The monaural parameters D/R and $C80_m$ for the experimental data are calculated from monaural RIRs with omnidirectional source and receiver directivities at the exact same positions as the source and receiver in the binaural case, but with the binaural receiver absent. $IACC_m$ is calculated from the same BRIRs used for the auralisations. For the RS11 data, The monaural parameters D/R and $C80_m$ are mean values across both ears, $IACC_m$ was calculated from the provided BRIRs.
For the experimental data, a regression analysis with the calculated optimum room-/ receiver dependent U/D limits per participant as the dependent variable and monaural/ binaural room acoustic parameters as independent variables was performed.
For both monaural parameters D/R and $C80_m$ significant regression equations could be found with ($F(1, 70) = 121.6, p < .000$), with an adjusted r^2 of .62 for D/R and ($F(1, 70) = 120.7, p < .000$), with an adjusted r^2 of .62 for $C80_m$. The optimum room-/ receiver dependent U/D-limits are equal to $52.1 - 6.0$ (D/R) ms and $93.4 - 5.0$ ($C80_m$) ms, both with a standard error of 21 ms.
In the binaural case, a significant regression equation was found ($F(1, 70) = 186.7, p < .000$), with an adjusted r^2 of .72. The optimum room-/ receiver dependent U/D-limit is equal to $143 - 201$ ($IACC_m$) ms with a standard error of 19 ms.
Due to the fact, that $IACC_m$ is a better predictor than the monaural parameters, it can be assumed, that the underlying perceptional mechanism is somehow related to a binaural phenomenon. This will be further discussed below.

3.2.3 Resulting prediction error with room/ receiver dependent U/D limits

Using the results from the regression analysis, the room-/ receiver dependent U/D limits as shown in the right three columns of Table 1 were found for the different predictors for the experimental data as well as the RS11 data. For a final comparison of the prediction accuracy between fixed and room-/ receiver dependent U/D limits, the resulting MAEs are depicted in Table 2. Further, the MAEs with the optimum U/D limits (which served as the dependent variable in the regression analysis) are depicted.

MAE [dB] with...	Fixed U/D limits [ms]		Optimum room-/ receiver dependent U/D limits	Room-/ receiver dependent U/D limits predicted by		
	50	100		D/R ($r^2_{adj.} = .62$)	$C80_m$ ($r^2_{adj.} = .62$)	$IACC_m$ ($r^2_{adj.} = .72$)
Experiment	1.9	2.9	0.2	1.3	1.3	1.2
RS11 data (S0/S90)	2.6/2.5	2.0/2.2	0.3/0.5	1.2/1.1	1.5/2.0	0.4/1.5
Ø	2.3	2.4	0.3	1.2	1.6	1.0

Table 2: Mean absolute errors (MAEs) with fixed and room dependent U/D limits

From the results depicted in Table 2 it can be concluded, that the mean prediction error can be reduced by means of a room-/ receiver dependent U/D by about 0.7 dB ($C80_m$) to 1.3 dB ($IACC_m$). The predicted SRTs with both fixed and room-/ receiver dependent U/D limits over condition are plotted in Figure 4 for both the experimental and the RS11 data.

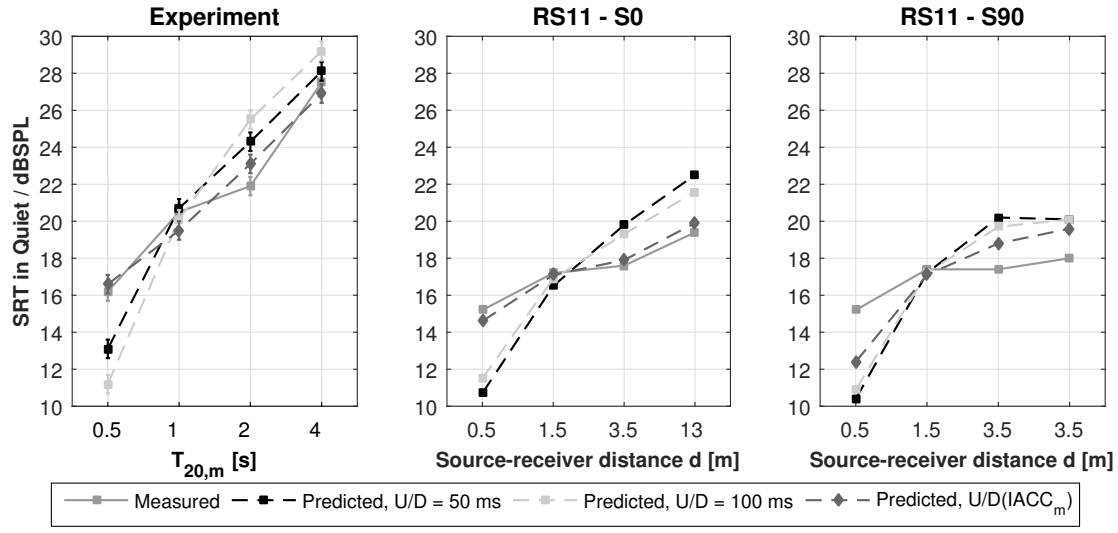

Figure 4: Measured and predicted SRTs with fixed and room dependent U/D limits

3.3 Prediction accuracy with pseudo-BRIRs

With the pseudo-binaural signals applied to the procedure outlined above, SRT predictions with fixed and room-/ receiver dependent U/D-limits can be observed as shown in Figure 5. Measured SRTs and predicted SRTs with BRIRs are shown for comparison.

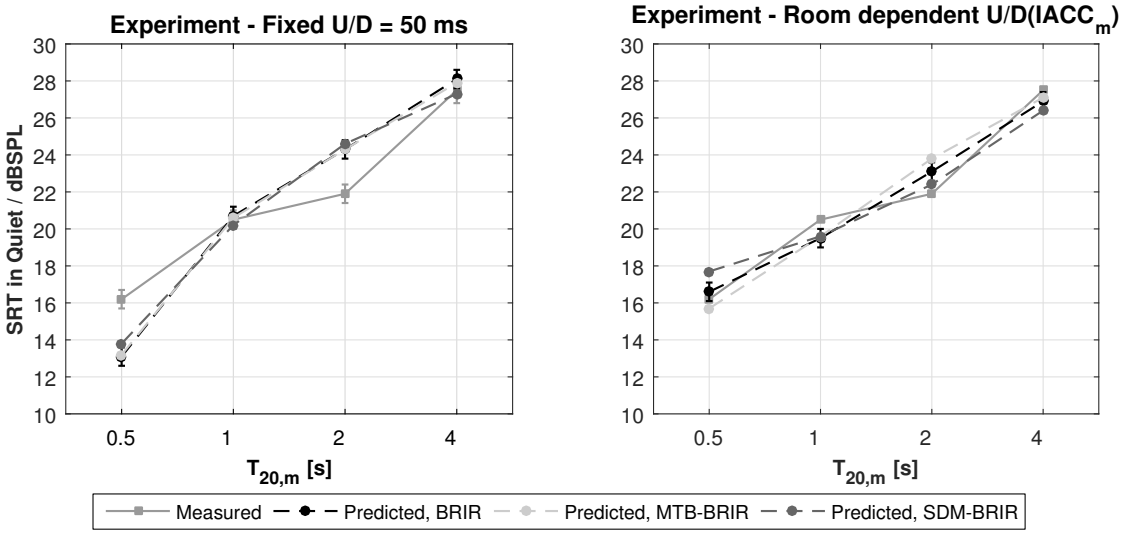

Figure 5: Measured and predicted SRTs with pseudo-BRIRs and fixed (left panel) and room dependent U/D limits (right panel)

The resulting MAEs with the pseudo-binaural signals as input to the prediction stage are depicted in Table 3. Results with BRIRs are depicted for comparison. Looking at the results in Figure 5 and Table 3, it can be concluded, that with the pseudo-BRIRs, similar trends as well as a similar prediction accuracy can be calculated as with the BRIRs. Again, with room-/ receiver dependent U/D limits, the MAE can be reduced by similar amounts as with BRIRs. Hence, both approaches seem to provide a sufficient accuracy in predicting SRT differences.

MAE [dB] with...	Fixed U/D limits [ms]		Room-/ receiver dependent U/D limits predicted by		
	50	100	D/R ($r^2_{adj.} = .62$)	$C80_m$ ($r^2_{adj.} = .62$)	$IACC_m$ ($r^2_{adj.} = .72$)
BRIR	1.9	2.9	1.3	1.3	1.2
MTB-BRIR	1.9	2.5	1.2	1.4	1.3
SDM-BRIR	1.9	2.7	1.2	1.2	1.3

Table 3: Mean absolute errors (MAEs) with pseudo-BRIRs

4 Discussion

4.1 Prediction accuracy with fixed U/D limits

Looking at the prediction accuracy over condition for the experimental as well as the RS11 data with the fixed U/D-limits (Figure 2 and 3), the following can be observed:
a) For small to medium receiver distances (RS11 data) and low to medium levels of reverberation (experimental data), the model with the fixed U/D limit fails to predict the influence of the room correctly. The deteriorating effect of reverberation on SRTs is overestimated by the model.
b) For larger source-receiver distances and high levels of reverberation, the model with the fixed U/D limit is generally able to predict the influence of reverberation on SRTs. Apparently, there is some room dependence in the prediction accuracy with the applied model with the fixed U/D limits, otherwise a general offset (i.e. constant under-/ overestimation of SRT increase) over all conditions between measured and modelled data would have been observed.
The main deviation between measured and predicted data is an overestimation of SRT increase (= degredation of speech perception) calculated by the model. Listeners are obviously able to make use of aspects of the reverberant signal which are not considered as useful by the model in its current implementation, or, to put in another way: There are perceptual mechanisms which had a positive effect on intelligibility in medium and low reverberation and short to medium receiver distances (without competing sources) which are ignored by the model.
For the underlying perceptual mechanisms leading to this lower increase in measured SRT with increasing level of reverberation compared to the model predictions, it can thus be deduced, that a) they must be room-/ receiver dependent by its nature and b) they are currently not captured by the model. In the context of speech perception in reverberation, two potential candidates fulfilling these requirements are binaural de-reverberation and room adaptation.
Binaural de-reverberation refers to the partial suppression/ squelch of reverberation when listening binaurally compared to monaurally. This effect is typically addressed by some sort of identification problem, i.e. the correct detection of test signals in a reverberant context with monaural and binaural presentation. Studies of the squelch of room effects date back to the early 60s with e.g. Koenig (1950) observing higher identification performance when listening binaurally to a signal played back in a reverberant room compared to monaural listening. Studies by Gelfand and Hochberg (1976), Moncur and Dirks (1967) and Nábělek and Robinson (1982) showed, that binaural de-reverberation is further subject to the absolute level of reverberation apparent in the room. They measured the identification performance of a test word with a preceding carrier phrase with three and four varying levels of reverberation respectively, ranging from reverberation time (RT) = 0 s (anechoic) to about RT = 3 s. The largest benefits due to binaural listening could be observed with an increased identification performance by about 10 % – 25 % for medium reverberant rooms, i.e. reverberation times of $1 - 2$ seconds. For lower and higher levels of reverberation, this benefit vanished. The strength of the effect of binaural de-reverberation is obviously room dependent.
Room adaptation refers to the partial suppression/ squelch of reverberation with prior exposure to the reverberant context compared to no prior exposure. Zahorik and Brandewie (2016) measured SRTs with and without prior exposure to the room context by a preceding carrier phrase played back before the actual test sentence. This was done for five different simulated sound fields with the

level of reverberation ranging from RT = 0 s (anechoic) to RT = 3 s. It was observed, that room adaptation had its largest influence at medium levels of reverberation of RT = 1 s with a decrease in SRT (= better intelligibility) of about 3 dB, vanishing to lower and higher levels of reverberation. This is in line with the results by Watkins (2005a,b) and Beeston et al. (2014) showing on the one hand a lower consonant identification performance with increasing level of reverberation on the test word alone but on the other hand an increasing performance, when the context (= preceding words) features the same level of reverberation as the test word. Further, it was shown, that the identification performance increased with increasing length of the reverberant context. The strength of the effect of room adaptation is obviously also room dependent.

To account for the room dependent effects of room adaptation, the model would need some knowledge about prior exposure to the rooms context. In its current implementation, there is no option to provide the model with such information. Further, to the authors best knowledge, there is still 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 room dependent effects of binaural de-reverberation, some sort of binaural processing stage is required. In the applied model, the only candidate therefore would be the implemented EC-stage. The original EC-model has been initially developed based on observations of masking level threshold of pure tones by broadband gaussian noise as a function of ITD and ILD and has been incorporated into the applied model to account for the unmasking of spatially distributed, localized target and masker sources, that is, the modelling of binaural unmasking. The current EC-implementation is driven by the interaural phase differences of the speech target, the masker and weighted by the interaural coherence of the masker. In a fixed spatial configuration, where target and masker are not co-located (= target interaural phase difference $\neq$ masker interaural phase difference) 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 cancelled more effectively.

With an increasing level of reverberation, the interaural coherence of the masking signal decreases, hence the binaural advantage according to the EC theory decreases. This can be seen e.g. in the unmasking study by Lavandier and Culling (2010). Here, the binaural advantage was calculated with the same model as in the present evaluation for a speech target and a spatially separated, localized masking source, where only the latter comprised reverberant components. In their figure 5, the decrease in binaural advantage with decreasing interaural masker coherence (implemented by the applied wall absorption coefficient α used for calculating the reverberant masker) for different spatial separations can be seen.

As can be concluded from the studies on binaural de-reverberation however, the binaural advantage due to de-reverberation rises with increasing level of reverberation (up to a certain maximum), the latter being correlated with a decrease in interaural coherence both for the speech target and any masking source. To model de-reverberation, the calculated binaural benefit would thus need to be high when the masker coherence is low which is the case at high medium and high levels of reverberation. This is contrary to EC-theory, hence the binaural model in its applied form does not/ cannot account for the effects of binaural de-reverberation. It can thus not be concluded, that binaural de-reverberation is unmasking from the late, diffuse masking source, as stated by Leclère et al. (2015), as unmasking und binaural de-reverberation are obviously inversely correlated with diffuse reverberation.

The suppressive effect of binaural de-reverberation might be correlated with the same perceptual mechanism causing the binaural echo suppression observed by Zurek (1979). Here, echo detection with a frontal source and a single delayed echo as a function of delay time and reflection amplitude was measured with the test reflection having either an ITD of zero (same as the source) or an ITD of 0.5 msec, roughly corresponding to a lateral sound incidence for the test reflection.

According to EC theory, the perception of the test reflection should be the better, i.e. suppression should be less when sound source and reflection are spatially separated, i.e. their ITDs differ. However, the contrary was observed, that is, the perception of the test reflection for very short reflection delays between 7 – 15 ms was better for an ITD of zero (no spatial separation between sound source and reflection). This finding might be correlated with the well-known precedence effect Wallach et al. (1949).

It was reported by the participants, that the main cue facilitating the detection of the test reflection was related to the perceived level of coloration. For the underlying mechanism, Zurek followed the concept of the "central spectrum" by Bilsen (1977), which assumes, that the auditory system calculates and evaluates some kind of inner representation combining both ear signals. Due to the presence of the direct sound and one or multiple reflections comb-filtering occurs in this central spectrum, causing more prominent dips and peaks, if both ears signals are identical i.e. with an ITD/ ILD = 0 compared to a spatially separated reflection, i.e. ITD/ ILD $\neq$ 0.
Buchholz (2007) confirmed this binaural echo suppression for very short reflection delays up to 10 ms. For larger reflection delays, a binaural enhancement of reflection detection was observed, i.e. the reflection was easier to detect when featuring ITD/ ILD $\neq$ 0, which is analogous to the basic idea of EC.
Obviously, suppressive effects such as binaural echo suppression (and presumably binaural de-reverberation) and enhancement effects such as binaural unmasking are opposite effects in the auditory system and thus cannot be modelled by EC alone.
Buchholz was able to (at least qualitatively) model both effects correctly by combining the concept of the central spectrum for modelling binaural suppression for very short reflection delays and an EC-based stage for modelling the binaural unmasking for larger delays. Open questions remained regarding a spectral importance weighting in the central spectrum concept and how both stages are combined in the auditory system. However, a similar approach might be applicable as a possible extension to the model used in the present evaluation, to account for both binaural echo suppression, presumably binaural de-reverberation and binaural unmasking.
Recently, a model framework featuring efferent (= Top-down) processing was presented (Beeston, 2015), which could at least qualitatively model binaural de-reverberation. However, it is still speculative if binaural suppression of reverberation is actually driven by efferent mechanisms.

4.2 Rationale for a room dependent U/D limit

With the room-/ receiver dependent U/D approach as presented in the evaluation at hand, the respective U/D limit is adapted to each room/ receiver configuration according to the respective similarity between both ear signals of the entire room response by means of the IACC_m.
For a low IACC_m, 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. This is to partially account for the benefits due to the perceptual mechanisms of binaural de-reverberation and room adaptation which also rise with rising level of reverberation (up to a certain limit) as observed in the present experiment. Vice versa, for a high IACC_m, correlated with a low level of diffuse reverberation, the U/D limit is decreased resulting in a reduced energy ratio between early useful and late detrimental components, i.e. the better-ear SNR calculated by the model. This decrease in SNR is analogous to the reduced benefits of both addressed perceptual mechanisms with low levels of reverberation.
By doing to, the overall model output can partially be compensated for the missing consideration of the room-/ receiver dependent effects of binaural de-reverberation and room adaptation in the context of speech perception, which are currently not handled by the model with a fixed U/D limit. The resulting mean prediction error with the applied model is reduced by about 1 dB compared to the model with fixed U/D limits. The presented method serves as a functional model minimizing the prediction error rather than constituting a correct psychophysical implementation of both perceptual mechanisms (room adaptation and binaural de-reverberation).

4.3 Prediction accuracy with pseudo-BRIRs

Pseudo-BRIRs have been calculated and applied to the binaural intelligibility model based on two different approaches, a straight forward MTB-based approach where the ear signals are taken from the two sensors placed on opposite sides of the spherical, rigid array and a spatial encoding method where the ear signals are calculated based on microphone array (here: B-Format) responses.
With both approaches, pseudo-binaural signals could be calculated leading to a similar prediction

accuracy when fed into the applied binaural prediction model with the fixed U/D limits as with BRIRs. Further, for both approaches, similar improvements in intelligibility by means of a room-/ receiver dependent U/D limit could be calculated with the presented methods.
As the sensors in the MTB microphone array are flush mounted on the equatorial line of a rigid sphere with a diameter of d = 17.6 cm (in analogy to the interaural distance of an average human head), the pseudo-binaural signals taken from sensors located on opposite sides of the sphere feature similar ILDs/ITDs as the BRIRs with the FABIAN HRTF data. The SDM-BRIRs are calculated as an HRTF-weighted sum of spatially distributed (by means of a virtual secondary source distribution) RIRs with the identical FABIAN HRTF data as with the BRIRs. Deviations to the BRIRs are primarily introduced by the limited spatial resolution in both the analysis (estimation of the incident angle of individual reflections) and synthesis (assignment of individual reflections to the nearest virtual secondary sound source) stage of the SDM routine with maximum localization errors of about 15° (Amengual Garí et al., 2017), maintaining similar ILDs/ITDs as the BRIRs with the FABIAN HRTF data. Besides the similarity of ILDs/ITDs introduced by all three types of receivers (FABIAN HRTF, spherical MTB array, SDM routine with FABIAN HRTF) identical sound fields were employed in the calculations for all three approaches by keeping all of Ravens' stochastic processes fixed. Hence, the similarity in the predicted SRTs with fixed and room-/ receiver dependent U/D limits with the three different types of receivers seems plausible. Both approaches can already be applied in practical room acoustic evaluation using existing measurement hard- and software. While the MTB based approach is generally also applicable to real-time applications, both approaches allow the re-orientation of the pseudo-binaural listener in postprocessing by adjusting the calculation/ selection of the respective multichannel signals used for calculating the respective ear signals.
Thus, with both methods arbitrary head orientations per position can be evaluated from a single set of multichannel microphone signals captured with a single impulse response measurement, making a physical re-orientation of the HATS superfluous stating the typical method for capturing BRIRs with different head orientations.
Improvements in intelligibility due to an optimized head orientation can thus easily be assessed in a less time-consuming manner both for assessing the suitability of existing room acoustic evaluation for speech reproduction as well as for prediction in a room acoustic design stage.

4.4 Validity of the simulated sound field representations

All binaural signals for both the perceptual experiments and intelligibility predictions have been simulated based on the concept of geometrical acoustics with a combination of image-source method, raytracing and stochastic processes for modelling the fine temporal structure of the late diffuse part, allowing the simulation of both specular and diffuse reflections. Realistic room acoustic absorption and scattering data have been incorporated in all simulations. Measured, full spherical directivity data for all sound sources and receivers have further been used with the only exception being the B-Format response used for calculating the pseudo-binaural signals based on the spatial encoding method which used a theoretically ideal directivity pattern. Thus, apart from the latter, deviations between simulated and measured binaural room responses cannot be excluded, general deviations affecting the validity of the presented methods are not expected.

5 Conclusion

As has been shown in the present evaluation, the applied 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 late reflections/ diffuse reverberation cannot fully account for the room-/ receiver dependent perceptual mechanisms affecting speech perception (without competing sources) in low and medium reverberation. Two room-/ receiver dependent effects, namely room adaptation and binaural de-reverberation are suspected to having affected the measured SRTs. The applied model cannot account for both effects, hence, deviations between measured and modelled data can be observed. The room-/ receiver dependence of the observed deviations is caused by the room-/ receiver dependent nature of both perceptual mechanisms addressed.

With the methodical implementation of a room-/ receiver dependent U/D classification coupled to the room acoustic parameters of the respective environment, a functional extension is presented which serves to reduce the mean prediction error by about 1 dB in the first place rather than constituting a correct psychophysical implementation of the two perceptual mechanisms improving the binaural models applicability for room acoustical studies.
Further, the general suitability of pseudo-binaural signals based on two different approaches for the assessment of speech perception in rooms has been shown in simulations, allowing the evaluation of arbitrary head orientations of the pseudo-binaural listener in a post-processing stage. All sound field simulations performed in the course of the present evaluation were generated with realistic (measured) boundary conditions (source/ receiver directivities, room geometry and surface properties). Hence, general deviations introduced by the simulation approach affecting the validity of the presented methods are not expected.

References

Ackermann, David; Felicitas Fiedler; Fabian Brinkmann; and Stefan Weinzierl (2016): "A high-quality microphone array for motion-tracked binaural (MTB) recording and reproducing of spatial sound." URL http://www.ak.tu-berlin.de/fileadmin/a0135/images/Mitarbeiter/AES_MTB.pdf.

Ahveninen, Jyrki; et al. (2006): "Task-modulated "what" and "where" pathways in human auditory cortex." In: *Proceedings of the National Academy of Sciences*, **103**(39), pp. 14608–14613.

Algazi, V. Ralph; Richard O. Duda; and Dennis M. Thompson (2004): "Motion-tracked binaural sound." In: *Journal of the Audio Engineering Society*, **52**(11), pp. 1142–1156.

Amengual Garí, Sebastià V.; Winfried Lachenmayr; and Eckard Mommertz (2017): "Spatial analysis and auralization of room acoustics using a tetrahedral microphone." In: *The Journal of the Acoustical Society of America*, **141**(4), pp. EL369–EL374. URL http://asa.scitation.org/doi/abs/10.1121/1.4979851.

ANSI, ANSI (1997): "S3. 5-1997, Methods for the calculation of the speech intelligibility index." In: *New York: American National Standards Institute*, **19**, pp. 90–119.

Beeston, Amy V. (2015): *Perceptual compensation for reverberation in human listeners and machines*. PhD Thesis, University of Sheffield.

Beeston, Amy V.; Guy J. Brown; and Anthony J. Watkins (2014): "Perceptual compensation for the effects of reverberation on consonant identification: Evidence from studies with monaural stimuli a." In: *The Journal of the Acoustical Society of America*, **136**(6), pp. 3072–3084.

Bernschütz, Benjamin (2016): *Microphone arrays and sound field decomposition for dynamic binaural recording*. Doctoral thesis, Fakultät I - Geisteswissenschaften, Technische Universität Berlin, https://depositonce.tu-berlin.de/handle/11303/5407. URL https://depositonce.tu-berlin.de/handle/11303/5407.

Beutelmann, Rainer and Thomas Brand (2006): "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners." In: *The Journal of the Acoustical Society of America*, **120**(1), pp. 331–342. URL http://scitation.aip.org/content/asa/journal/jasa/120/1/10.1121/1.2202888.

Beutelmann, Rainer; Thomas Brand; and Birger Kollmeier (2010): "Revision, extension, and evaluation of a binaural speech intelligibility model." In: *The Journal of the Acoustical Society of America*, **127**(4), pp. 2479–2497. URL http://scitation.aip.org/content/asa/journal/jasa/127/4/10.1121/1.3295575.

Bilsen, Frans A. (1977): "Pitch of noise signals: Evidence for a"central spectrum"." In: *The Journal of the Acoustical Society of America*, **61**(1), pp. 150–161.

Bradley, J. S.; Hiroshi Sato; and M. Picard (2003): "On the importance of early reflections for speech in rooms." In: *The Journal of the Acoustical Society of America*, **113**(6), pp. 3233–3244. URL http://scitation.aip.org/content/asa/journal/jasa/113/6/10.1121/1.1570439.

Bradley, John S. (1986): "Predictors of speech intelligibility in rooms." In: *The Journal of the Acoustical Society of America*, **80**(3), pp. 837–845.

Brinkmann, Fabian; et al. (2017a): "The FABIAN head-related transfer function data base." URL https://depositonce.tu-berlin.de//handle/11303/6153.

Brinkmann, Fabian; et al. (2017b): "A High Resolution and Full-Spherical Head-Related Transfer Function Database for Different Head-Above-Torso Orientations." In: *Journal of the Audio Engineering Society*, **65**(10), pp. 841–848.

Buchholz, Jörg M. (2007): "Characterizing the monaural and binaural processes underlying reflection masking." In: *Hearing research*, **232**(1), pp. 52–66. URL http://www.sciencedirect.com/science/article/pii/S0378595507001608.

Burkhard, M. D. and R. M. Sachs (1975): "Anthropometric manikin for acoustic research." In: *The Journal of the Acoustical Society of America*, **58**(1), pp. 214–222.

Chabot-Leclerc, Alexandre; Ewen N. MacDonald; and Torsten Dau (2016): "Predicting binaural speech intelligibility using the signal-to-noise ratio in the envelope power spectrum domain." In: *The Journal of the Acoustical Society of America*, **140**(1), pp. 192–205.

Culling, J. F.; M. Lavandier; and S. Jelfs (2013): "Predicting binaural speech intelligibility in architectural acoustics." In: *The technology of binaural listening*. Springer, pp. 427–447. URL http://link.springer.com/chapter/10.1007/978-3-642-37762-4_16.

Culling, John F.; Monica L. Hawley; and Ruth Y. Litovsky (2005): "Erratum: The role head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources [J. Acoust. Soc. Am. 116, 1057 (2004)]." In: *The Journal of the Acoustical Society of America*, **118**(1), pp. 552–552.

Dalenbäck, Bengt-Inge (2008): "Catt-acoustic v. 8."

Dietrich, Pascal; Bruno Masiero; Martin Pollow; Roman Scharrer; and M. Muller-Trapet (2010): "Matlab toolbox for the comprehension of acoustic measurement and signal processing." In: *Fortschritte der Akustik–DAGA, Berlin.*

DIN 18041 (2016): "Hörsamkeit in Räumen - Anforderungen, Empfehlungen und Hinweise für die Planung." URL https://www.beuth.de/de/norm/din-18041/245356770.

DIN EN 60645-1 (2015): "Akustik - Audiometer - Teil 1: Reinton-Audiometer (IEC 60645-1:2012); Deutsche Fassung EN 60645-1:2015." URL https://www.beuth.de/de/norm/din-en-60645-1/237774964.

DIN EN ISO 8253-1 (2011): "Akustik - Audiometrische Prüfverfahren - Teil 1: Grundlegende Verfahren der Luft- und Knochenleitungs-Schwellenaudiometrie mit reinen Tönen (ISO 8253-1:2010); Deutsche Fassung EN ISO 8253-1:2010." URL https://www.beuth.de/de/norm/din-en-iso-8253-1/132501782.

Durlach, Nathaniel I. (1963): "Equalization and Cancellation Theory of Binaural Masking-Level Differences." In: *The Journal of the Acoustical Society of America*, **35**(8), pp. 1206–1218. URL http://scitation.aip.org/content/asa/journal/jasa/35/8/10.1121/1.1918675.

Edmonds, Barrie A. and John F. Culling (2006): "The spatial unmasking of speech: Evidence for better-ear listening." In: *The journal of the Acoustical Society of America*, **120**(3), pp. 1539–1545. URL http://scitation.aip.org/content/asa/journal/jasa/120/3/10.1121/1.2228573.

Ellis, Gregory M.; Pavel Zahorik; and William M. Hartmann (2015): "Using multidimensional scaling techniques to quantify binaural squelch." In: *Proceedings of Meetings on Acoustics*, **23**(1), p. 050007. doi:10.1121/2.0000164. URL http://asa.scitation.org/doi/abs/10.1121/2.0000164.

Ewert, Stephan D. and Torsten Dau (2000): "Characterizing frequency selectivity for envelope fluctuations." In: *The Journal of the Acoustical Society of America*, **108**(3), pp. 1181–1196.

Fuß, Alexander; Fabian Brinkmann; Thomas Jürgensohn; and Stefan Weinzierl (2015): "Ein vollsphärisches Multikanalmesssystem zur schnellen Erfassung räumlich hochaufgelöster, individueller kopfbezogener Ubertragungsfunktionen." In: *Fortschritte der Akustik–DAGA Nürnberg*, pp. 1114–1117.

Gelfand, S. A. and I. Hochberg (1976): "Binaural and monaural speech discrimination under reverberation." In: *Audiology*, **15**(1), pp. 72–84.

Grange, Jacques (2016): "The benefit of head orientation to speech intelligibility in noise." In: *The Journal of the Acoustical Society of America*, **139**(2), pp. 703–712. doi:10.1121/1.4941655. URL http://asa.scitation.org/doi/full/10.1121/1.4941655.

Grange, Jacques A. and John F. Culling (2016): "Head orientation benefit to speech intelligibility in noise for cochlear implant users and in realistic listening conditions." In: *The Journal of the Acoustical Society of America*, **140**(6), pp. 4061–4072. URL http://asa.scitation.org/doi/abs/10.1121/1.4968515.

HörTech gGmbH (2011): "Oldenburger Satztest Rev01.0 - Bedienungsanleitung für den manuellen Test auf Audio-CD." URL http://www.hoertech.de/images/hoertech/pdf/mp/produkte/olsa/HT.OLSA_Handbuch_Rev01.0_mitUmschlag.pdf.

IEC 60268-16 (2011): "Sound system equipment-Part 16: Objective rating of speech intelligibility by speech transmission index." In: *International Electrotechnical Commission (IEC).*

Jelfs, Sam; John F. Culling; and Mathieu Lavandier (2011): "Revision and validation of a binaural model for speech intelligibility in noise." In: *Hearing research*, **275**(1), pp. 96–104. URL http://www.sciencedirect.com/science/article/pii/S0378595510004387.

Jørgensen, Søren and Torsten Dau (2011): "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing." In: *The Journal of the Acoustical Society of America*, **130**(3), pp. 1475–1487.

Jørgensen, Søren; Stephan D. Ewert; and Torsten Dau (2013): "A multi-resolution envelope-power based model for speech intelligibility." In: *The Journal of the Acoustical Society of America*, **134**(1), pp. 436–446.

Kock, W. E. (1950): "Binaural localization and masking." In: *The Journal of the Acoustical Society of America*, **22**(6), pp. 801–804.

Koenig, W. (1950): "Subjective effects in binaural hearing." In: *The Journal of the Acoustical Society of America*, **22**(1), pp. 61–62.

Kokabi, Omid; Fabian Brinkmann; and Stefan Weinzierl (2018): "Assessment of speech perception based on binaural room impulse responses." In: . doi:http://dx.doi.org/10.14279/depositonce-6725. URL https://depositonce.tu-berlin.de//handle/11303/7505.

Kuehnel, Volker; Birger Kollmeier; and Kirsten Wagener (1999): "Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests." In: *Zeitschrift für Audiologie*, **38**, pp. 4–15.

Lavandier, Mathieu and John F. Culling (2008): "Speech segregation in rooms: Monaural, binaural, and interacting effects of reverberation on target and interferer." In: *The Journal of the Acoustical Society of America*, **123**(4), pp. 2237–2248. URL http://asa.scitation.org/doi/abs/10.1121/1.2871943.

Lavandier, Mathieu and John F. Culling (2010): "Prediction of binaural speech intelligibility against noise in rooms." In: *The Journal of the Acoustical Society of America*, **127**(1), pp. 387–399. URL http://scitation.aip.org/content/asa/journal/jasa/127/1/10.1121/1.3268612.

Lavandier, Mathieu; et al. (2012): "Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources." In: *The Journal of the Acoustical Society of America*, **131**(1), pp. 218–231. URL http://scitation.aip.org/content/asa/journal/jasa/131/1/10.1121/1.3662075.

Leclère, Thibaud; Mathieu Lavandier; and John F. Culling (2015): "Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking, and binaural de-reverberation." In: *The Journal of the Acoustical Society of America*, **137**(6), pp. 3335–3345. URL http://scitation.aip.org/content/asa/journal/jasa/137/6/10.1121/1.4921028.

Lindau, Alexander; Stefan Weinzierl; and H. J. Maempel (2006): "FABIAN-An instrument for software-based measurement of binaural room impulse responses in multiple degrees of freedom." In: *24. Tonmeistertagung*.

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

McShefferty, David; William M. Whitmer; and Michael A. Akeroyd (2015): "The just-noticeable difference in speech-to-noise ratio." In: *Trends in hearing*, **19**, p. 2331216515572316.

Middlebrooks, John; Jonathan Z. Simon; Arthur N. Popper; and Richard R. Fay (2017): *The auditory system at the cocktail party*. Springer.

Moncur, John P. and Donald Dirks (1967): "Binaural and monaural speech intelligibility in reverberation." In: *Journal of speech and hearing research*, **10**(2), pp. 186–195.

Nábělek, Anna K. and Pauline K. Robinson (1982): "Monaural and binaural speech perception in reverberation for listeners of various ages." In: *The Journal of the Acoustical Society of America*, **71**(5), pp. 1242–1248.

Paquier, Mathieu and Vincent Koehl (2015): "Discriminability of the placement of supra-aural and circumaural headphones." In: *Applied Acoustics*, **93**, pp. 130–139.

Paquier, Mathieu; Vincent Koehl; and Brice Jantzem (2012): "Influence of headphone position in pure-tone audiometry." In: *Acoustics 2012 joint congress (11ème Congrès Français d'Acoustique-2012 Annual IOA Meeting)*. pp. 3925–3930.

Rennies, Jan (2014): "Modeling the effects of a single reflection on binaural speech intelligibility." In: *The Journal of the Acoustical Society of America*, **135**(3), pp. 1556–1567. doi:10.1121/1.4863197. URL http://asa.scitation.org/doi/full/10.1121/1.4863197.

Rennies, Jan; Thomas Brand; and Birger Kollmeier (2011): "Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet." In: *The Journal of the Acoustical Society of America*, **130**(5), pp. 2999–3012. URL http://scitation.aip.org/content/asa/journal/jasa/130/5/10.1121/1.3641368.

Schröder, Dirk and Michael Vorländer (2011): "RAVEN: A real-time framework for the auralization of interactive virtual environments." In: *Forum Acusticum*. URL https://www2.ak.tu-berlin.de/~akgroup/ak_pub/seacen/2011/Schroeder_2011b_P2_RAVEN_A_Real_Time_Framework.pdf.

Søndergaard, P. and P. Majdak (2013): "The Auditory Modeling Toolbox." In: *The Technology of Binaural Listening*. Berlin, Heidelberg: Springer, pp. 33–56.

Tervo, Sakari (2016): "SDMtoolbox." URL http://de.mathworks.com/matlabcentral/fileexchange/56663-sdmtoolbox.

Tervo, Sakari; Jukka Pätynen; Antti Kuusinen; and Tapio Lokki (2013): "Spatial decomposition method for room impulse responses." In: *Journal of the Audio Engineering Society*, **61**(1/2), pp. 17–28. URL http://www.aes.org/e-lib/browse.cfm?elib=16664.

The MathWorks, Inc. (2013): "MATLAB 2015b." URL http://www.walkingrandomly.com/?p=4767.

van Wijngaarden, Sander J. and Rob Drullman (2008): "Binaural intelligibility prediction based on the speech transmission index." In: *The Journal of the Acoustical Society of America*, **123**(6), pp. 4514–4523. URL http://scitation.aip.org/content/asa/journal/jasa/123/6/10.1121/1.2905245.

Wagener, K.; T. Brand; and B. Kollmeier (1999a): "Entwicklung und Evaluation eines Satztests für die deutsche Sprache III: Evaluation des Oldenburger Satztests." In: *Zeitschrift fr Audiologie/Audiological Acoustics*, **38**, p. 8695.

Wagener, Kirsten; Thomas Brand; and Birger Kollmeier (1999b): "Entwicklung und Evaluation eines Satztests für die deutsche Sprache II: Optimierung des Oldenburger Satztests." In: *Zeitschrift für Audiologie/Audiological Acoustics*, **38**, pp. 44–56.

Wallach, Hans; Edwin B. Newman; and Mark R. Rosenzweig (1949): "A precedence effect in sound localization." In: *The Journal of the Acoustical Society of America*, **21**(4), pp. 468–468.

Warzybok, Anna; Jan Rennies; Thomas Brand; Simon Doclo; and Birger Kollmeier (2013): "Effects of spatial and temporal integration of a single early reflection on speech intelligibilitya)." In: *The Journal of the Acoustical Society of America*, **133**(1), pp. 269–282. URL http://scitation.aip.org/content/asa/journal/jasa/133/1/10.1121/1.4768880.

Watkins, Anthony J. (2005a): "Perceptual compensation for effects of echo and of reverberation on speech identification." In: *Acta acustica united with Acustica*, **91**(5), pp. 892–901.

Watkins, Anthony J. (2005b): "Perceptual compensation for effects of reverberation in speech identification." In: *The Journal of the Acoustical Society of America*, **118**(1), pp. 249–262.

Wefers, Frank (2010): "A free, open-source software package for directional audio data." In: *Proceedings of the 36th German Annual Conference on Acoustics (DAGA 2010)*.

Zahorik, Pavel and Eugene J. Brandewie (2016): "Speech intelligibility in rooms: Effect of prior listening exposure interacts with room acoustics." In: *The Journal of the Acoustical Society of America*, **140**(1), pp. 74–86.

Zurek, Patrick M. (1979): "Measurements of binaural echo suppression." In: *The Journal of the Acoustical Society of America*, **66**(6), pp. 1750–1757.