Speech Recognition Predictions for Measured and Simulated Binaural Room Impulse Responses

Merle Gerken[1], Christoph Kirsch[1], Julia Schütze[1], Stephan D. Ewert[1], Anna Warzybok[1]
[1] *Medizinische Physik and Cluster of Excellence Hearing4all, Carl von Ossietzky Universität Oldenburg, Germany*
E-mail: {merle.gerken, christoph.kirsch, julia.schuetze, stephan.ewert, a.warzybok}@uol.de

Introduction

For many applications, such as hearing device benefit assessment or hearing diagnostics, speech recognition performance is typically evaluated under simple conditions. This leads to discrepancies compared to performance in real-life scenarios, which involve more complex configurations of sound sources [1, 2]. In contrast to laboratory setups with a single noise source, the use of complex auditory environments allows for more ecologically valid investigations of speech recognition. Such complex auditory environments can be reproduced under controlled conditions using room acoustic simulations (e.g., [3]).

Various models can be used to simulate the acoustic environment, such as the Perceptually Plausible Room Acoustics Simulator (RAZR) [4] and the Toolbox for Acoustic Scene Creation and Rendering (TASCAR) [5]. Within the models, various parameters can be compared, e.g. characterizing the source and receiver model. So far, it is unknown to what extent these different simulation algorithms and approaches have an impact on the resulting speech recognition predictions. Therefore, the aim of this study is to simulate speech recognition performance for measured and simulated binaural room impulse responses (BRIRs) of a living room. For the simulated BRIRs, the two models RAZR and TASCAR are compared considering the influence of the parameters source directivity and head model. The two models are compared to each other as well as to the measured BRIRs in terms of predicted speech recognition performance.

Methods

Generation of binaural room impulse responses

A real living room with the dimensions 4.97 x 3.78 x 2.71 m and an adjacent kitchen with the dimensions 4.97 x 2.00 x 2.71 m was used as the acoustic environment. The rooms were connected by an open door, and the living room was furnished with a couch, two armchairs, a coffee table, a TV on a board, a bookshelf, a cabinet and curtains. The kitchen contained a table with two chairs. A floor plan is shown in Figure 1. A geometric model of the living room environment as defined by [6] was implemented in RAZR [4] and TASCAR [5] to create virtual replications of the real room.

BRIRs were measured in the real room, using a G.R.A.S. KEMAR 45BB head and torso simulator (HATS) and a Genelec 8030 loudspeaker. These measurements were conducted for the HATS placed in position R and oriented towards position STV, for sources in positions S4, S5, and STV, respectively. With RAZR, sets of BRIRs were generated using either measured head-related transfer functions (HRTF) [7] or an extended version of the spherical head model (referred to as extSHM [8, 9]). In the HRTF head model, for each desired sound direction, the three closest HRTF measurement directions were selected and interpolated, and the interpolated HRTF was convolved with the sound signal. The extSHM head model was an extended version of the spherical head model (SHM) [8] with additional filters to model shadow effects and reflections [9]. The receiver was rotated to either STV or S4. As an additional permutation of the rendering parameters, the virtual sound sources were either rendered as omnidirectional sources or the source directivity and spectral characteristics of the Genelec 8030 loudspeaker were recreated based on a set of impulse responses measured on a spherical surface around the loudspeaker in anechoic conditions. The same interpolation procedure as described for the HRTF was used. Reverberation was simulated by a third-order image source model (ISM) and a physically based feedback delay network (FDN) [4, 10]. In TASCAR, the receiver at position R was simulated by a modified version of extSHM (referred to as extSHM*) [11]. In this version, the filter parameters of the head model were optimized to match the resulting speech recognition scores obtained with measured HRTFs [12]. The receiver was again rotated towards STV or towards S4, and the sources were rendered omnidirectionally. Reverberation was generated using an ISM for early reflections and a simple FDN based on [13, 14] for late reverberation. For both RAZR and TASCAR, additional sets of BRIRs were generated for an anechoic case to study the effect of reverberation.

The measured and simulated sets of BRIRs were used to evaluate different room configurations. The target loudspeaker was placed at position STV or S4, and the maskers were placed as single sources at positions STV, S4, S5, or simultaneously at S4 and S5.

Speech recognition predictions

The Binaural Speech Intelligibility Model (BSIM) [15] was used to predict the speech recognition threshold (SRT), i.e. the signal-to-noise ratio at which 50 % speech recognition is achieved. The measured and simulated BRIRs of the target and masker sources were convolved with the stationary, test-specific noise of the German matrix sentence test (Oldenburger Satztest) OLSA [16]. These separate target and masker signals were provided as inputs to the model. The input signals were filtered by a gammatone filterbank [17] in the range of 146 to

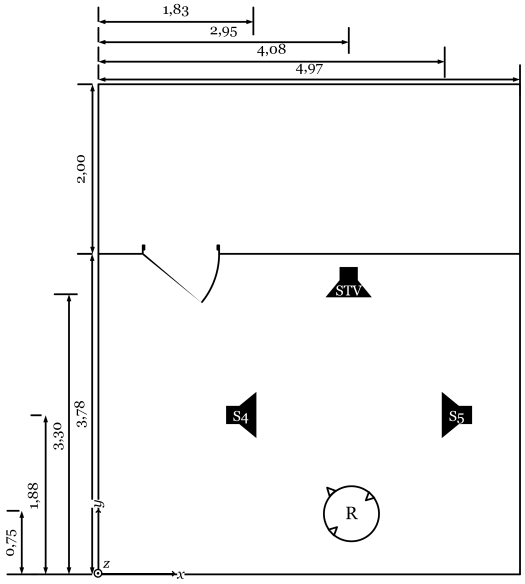

Figure 1: Floor plan of the measured and simulated living room (bottom) and the adjacent kitchen (top). The receiver was located at position R, the sound sources were located at positions S4, S5 and STV.

8346 Hz, before an equalization-cancellation (EC) process [18] was applied in each frequency band. In this process, the interaural time difference (ITD) is estimated in each frequency band before calculating an attenuation factor. By applying this attenuation factor in the cancellation step, the SNR is maximized. In the next step, this maximum SNR is used to compute the speech intelligibility index (SII) [19], which is then mapped to an SRT. For this mapping, a reference SII of 0.2 was used, which refers to an SRT of -7.1 dB SNR [15, 16].

Results

Speech recognition performance

In Figure 2, the predicted SRTs are visualized for all tested conditions. In anechoic conditions, the variations between spatial conditions and between BRIR simulation methods (room model, head model, source directivity) show a larger impact than in echoic conditions. In all model configurations, there is a beneficial effect when target and masker are spatially separated. Source directivity has an effect of up to 3 dB, and the head model affects the SRTs by up to 2 dB.

In echoic conditions with head direction towards source STV, the SRTs based on simulated BRIRs can be compared to those obtained with measured BRIRs. They appear in a similar range of -13 dB to -8 dB, with deviations up to about 2 dB depending on the spatial configuration of target and masker(s).

Influence of room model

To investigate the influence of the room model, Figure 3 shows the correlation of SRTs predicted with RAZR and with TASCAR for anechoic and echoic conditions. For comparability, conditions with omnidirectional sources and extSHM / extSHM* head model are visualized, including all spatial conditions. There is a high similarity between the SRTs obtained with both models, quantified by a correlation of $R^2 = 0.94$, a bias of -0.4 dB and a root-mean-square error (RMSE) of 0.9 dB.

Comparison of measured and simulated BRIRs

Table 1 provides an overview of the bias, RMSE and R^2 between SRTs based on measured BRIRs and SRTs based on simulated BRIRs for different parameters. In terms of correlation with SRTs from measured BRIRs, signals generated with RAZR benefit from having source directivity turned on. The correlation increases from 0.55 to 0.77 and from 0.59 to 0.84 for HRTF and extSHM head model, respectively. The RMSE also benefits from the source directivity and decreases from 1.0 dB to 0.8 dB and from 1.1 dB to 0.7 dB for HRTF and extSHM head model, respectively. For the bias, the picture is less clear, but the differences between the bias values have a magnitude of only 0.2 dB. The type of receiver model also influences the simulated SRTs, a better agreement between measured and simulated BRIRs was observed for extSHM head model ($R^2 = 0.84$ and $R^2 = 0.59$) than for HRTF ($R^2 = 0.77$ and $R^2 = 0.55$). The bias also benefits from the extSHM head model instead of the HRTF head model, decreasing from 0.5 dB to 0 dB and from 0.3 dB to -0.2 dB. The RMSE is not influenced by the type of the head model used (the differences do not exceed 0.1 dB). Comparing TASCAR and RAZR (directivity off , head model extSHM / extSHM*), a slightly better agreement in terms of R^2 with the SRTs for measured BRIRs can be observed for TASCAR ($R^2 = 0.62$ versus $R^2 = 0.55$), but at the same time RAZR results in a smaller bias. The RMSE seems to be almost the same (1.1 dB and 1.0 dB, respectively). The highest agreement between SRTs with measured and simulated BRIRs was obtained for the RAZR configuration with source directivity and extSHM head model, resulting in $R^2 = 0.84$.

Table 1: Comparison of predicted SRTs for target positions STV and S4 with measured and simulated BRIRs, for echoic conditions with head direction towards STV. The bias, root-mean-square error (RMSE) and correlation R^2 between SRTs predicted for measured BRIRs and SRTs predicted for simulated BRIRs with different parameters (room model, source directivity, head model) are shown.

Room model	Source directivity	Head model	Bias / dB	RMSE / dB	R^2
RAZR	on	HRTF	-0.2	0.8	0.77
RAZR	on	extSHM	0.3	0.7	0.84
RAZR	off	HRTF	0.0	1.0	0.55
RAZR	off	extSHM	0.5	1.1	0.59
TASCAR	off	extSHM*	-0.3	1.1	0.62

Discussion

The SRT predictions have shown the effect of the BRIR simulation parameters on the predicted speech recognition performance compared to the measured BRIRs. In general, source directivity seems to have an important influence, especially in conditions where binaural aspects should be correctly reproduced, such as in anechoic con-

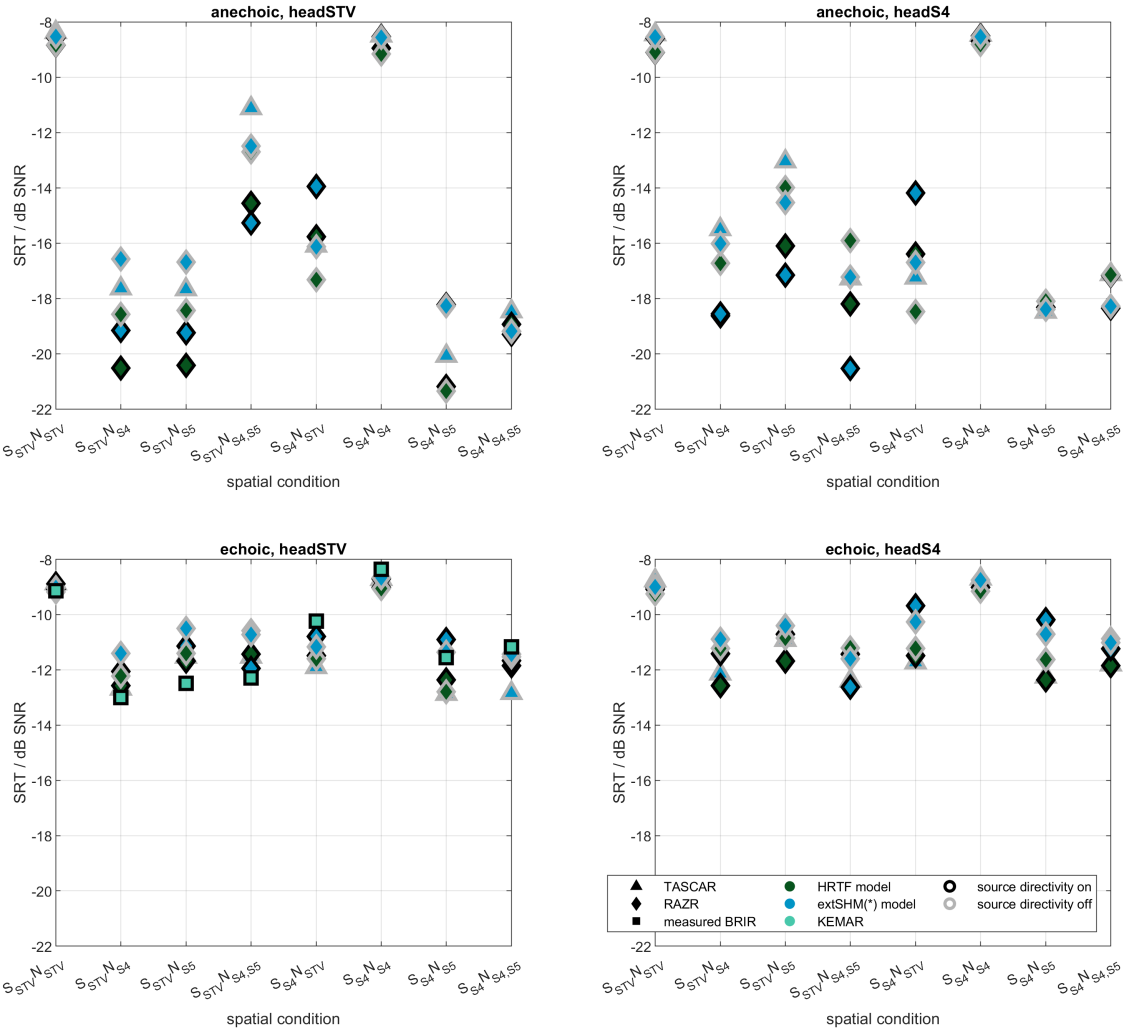

Figure 2: Overview of all predicted SRTs in dB SNR as a function of the spatial condition, the reverberation, head direction, room model (marker shape), head model (marker color), and source directivity (marker edge).

ditions with spatially separated target and masker(s). The influence is marginal in co-located and echoic conditions. The influence of the head model was less pronounced than that of the source directivity, and seemed to be important under conditions similar to those of the source directivity. Due to the positive effect of source directivity on prediction accuracy for RAZR, it would be interesting to investigate this effect for TASCAR in future studies.

The applied speech recognition model is based on the SII, which performs an index calculation based on weighted frequency-dependent SNRs. This may limit the validity of the results with respect to room acoustics. In the model, the whole energy of the target signal (including the reverberant part) is treated as useful and by that is assumed to contribute to speech recognition. This does not take into account the detrimental effect of reverberation in the target signal. However, this aspect is not crucial for short distances between receiver and target source [20], as in the present study. For a more detailed evaluation, the predictions may need to be verified with human listeners.

Conclusions

SRTs predicted for measured and simulated BRIRs show a generally high agreement. Comparing the SRTs predicted for RAZR and TASCAR, there is a high correlation ($R^2 = 0.95$) and small deviations. The SRTs based on recorded and simulated BRIRs show a bias of 0.0-0.5 dB and correlations of $R^2 = 0.55 - 0.84$, depending on the room model. Source directivity affects the predicted SRTs by up to about 3 dB. The influence is more pronounced under anechoic conditions than under echoic conditions. Depending on the spatial configuration of speech source, masker source, and head direction, the SRTs are affected up to about 2 dB by the applied head model. The highest agreement between SRTs with measured BRIRs and simulated BRIRs was achieved with active source directivity and the extSHM head model.

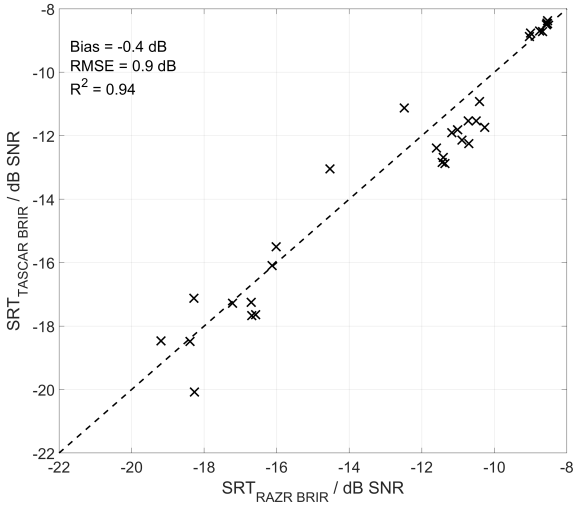

Figure 3: Comparison of SRTs for target positions STV and S4 based on BRIRs generated with TASCAR and RAZR, for anechoic and echoic conditions. The BRIRs were obtained with omnidirectional sources and head model extSHM / extSHM*. The dashed line visualizes the case of perfect correlation.

Acknowledgement

The project is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 352015383 – SFB 1330 A5 / C4 / C5.

References

[1] Bentler, R. A. (2005). Effectiveness of Directional Microphones and Noise Reduction Schemes in Hearing Aids: A Systematic Review of the Evidence. Journal of the American Academy of Audiology, 16(07), 473–484. https://doi.org/10.3766/jaaa.16.7.7

[2] Cord, M. T., Surr, R. K., Walden, B. E., & Dyrlund, O. (2004). Relationship between Laboratory Measures of Directional Advantage and Everyday Success with Directional Microphone Hearing Aids. Journal of the American Academy of Audiology, 15(05), 353–364. https://doi.org/10.3766/jaaa.15.5.3

[3] Aspöck, L., Pausch, F., Stienen, J., Berzborn, M., Kohnen, M., Fels, J., & Vorländer, M. (2018). Application of virtual acoustic environments in the scope of auditory research. Proceedings of XXVIII Encontro da Sociedade Brasileira de Acústica.

[4] Wendt, T., Van De Par, S., & Ewert, S. D. (2014). A computationally-efficient and perceptually-plausible algorithm for binaural room impulse response simulation. Journal of the Audio Engineering Society, 62(11), 748-766.

[5] Grimm, G., Luberadzka, J., & Hohmann, V. (2019). A toolbox for rendering virtual acoustic environments in the context of audiology. Acta acustica united with acustica, 105(3), 566-578.

[6] Schütze, J., Kirsch, C., Wagener, K. C., Kollmeier, B., & Ewert, S. D. (2021). Living room environment (1.1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5747753

[7] Braren, H., & Fels, J. (2020). A High-Resolution Head-Related Transfer Function Data Set and 3D-Scan of KEMAR. Institute for Hearing Technology and Acoustics, RWTH Aachen University, Technical report.

[8] Brown, C. P., & Duda, R. O. (1998). A structural model for binaural sound synthesis. IEEE transactions on speech and audio processing, 6(5), 476-488.

[9] Ewert, S. D., Buttler, O., & Hu, H. (2021). Computationally Efficient Parametric Filter Approximations for Sound-Source Directivity and Head-Related Impulse Responses. 2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA), 1–6. https://doi.org/10.1109/I3DA48870.2021.9610923

[10] Kirsch, C., Wendt, T., Van De Par, S., Hu, H., & Ewert, S. D. (2023). Computationally-efficient simulation of late reverberation for inhomogeneous boundary conditions and coupled rooms. Journal of the Audio Engineering Society, 71(4), 186-201.

[11] Schwark, F., Schädler, M. R., & Grimm, G. (2022). Data-driven optimization of parametric filters for simulating head-related transfer functions in real-time rendering systems. EUROREGIO BNAM2022, 1–10.

[12] Denk, F., Ernst, S. M., Ewert, S. D., & Kollmeier, B. (2018). Adapting hearing devices to the individual ear acoustics: Database and target response correction functions for various device styles. Trends in hearing, 22, 2331216518779313.

[13] Schroeder, M. R. (1961, October). Natural sounding artificial reverberation. In Audio Engineering Society Convention 13. Audio Engineering Society.

[14] Rocchesso, D., & Smith, J. O. (1997). Circulant and elliptic feedback delay networks for artificial reverberation. IEEE Transactions on Speech and Audio Processing, 5(1), 51-63.

[15] Beutelmann, R., Brand, T., & Kollmeier, B. (2010). Revision, extension, and evaluation of a binaural speech intelligibility model. The Journal of the Acoustical Society of America 127(4), 2479–2497.

[16] Wagener, K., Brand, T., Kuehnel, V., & Kollmeier, B. (1999). Entwicklung und Evaluation eines Satztests für die deutsche Sprache I-III: Design, Optimierung und Evaluation des Oldenburger Satztest. Zeitschrift für Audiologie, 38.

[17] Hohmann, V. (2002). Frequency analysis and synthesis using a Gammatone filterbank. Acta Acustica united with Acustica, 88(3), 433–442.

[18] Durlach, N. I. (1963). Equalization and Cancellation Theory of Binaural Masking-Level Differences. The Journal of the Acoustical Society of America, 35(8), 1206–1218. https://doi.org/10.1121/1.1918675

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

[20] Minelli, G., Puglisi, G. E., Astolfi, A., Hauth, C., & Warzybok, A. (2023). Objective Assessment of Binaural Benefit from Acoustical Treatment in Real Primary School Classrooms. International Journal of Environmental Research and Public Health, 20(10), Article 10. https://doi.org/10.3390/ijerph20105848