

Using interaural mistuning for modelling binaural processing inaccuracies in human speech recognition

Simon Weihe^{1,3}, Jan Rennies-Hochmuth^{1,2,3}, Thomas Brand^{1,3}

¹ *CvO Universität Oldenburg, 26111 Oldenburg, Germany, Email: simon.weihe@uol.de, thomas.brand@uol.de*

² *Fraunhofer IDMT, HSA, 26129 Oldenburg, Germany, Email: jan.rennies-hochmuth@idmt.fraunhofer.de*

³ *Exzellenzcluster "Hearing4All"*

Introduction

The objective of the collaborative research center (Sonderforschungsbereich, SFB) 1330 "Hörakustik: Perzeptive Prinzipien, Algorithmen und Anwendungen" (HAPPAA) is to improve human communication in difficult conditions. The project takes into account that communication takes place in a loop: A sound field is generated and then processed by a device (for instance, a pair of hearing aids), then it is perceived by a listener (subject) who in turn influences the sound field by his or her behavior (Fig. 1). In HAPPAA we try to model all of these different stages. Furthermore, we try to integrate such models into the device in a device-internal model loop (green in Fig. 1), in order to control speech enhancement algorithms in the device.

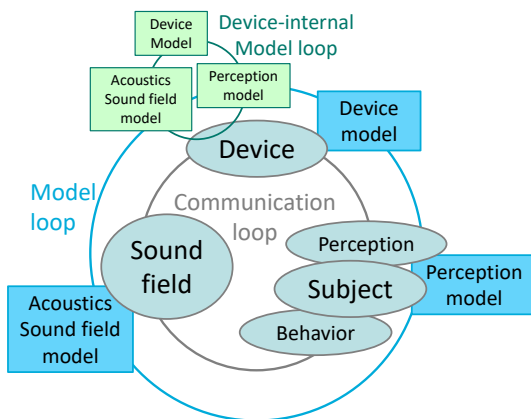


Figure 1: Communication loop, models of all stages (blue), and device-internal model loop (green). [1]

The focus of this study is on a binaural perception model for predicting speech intelligibility in spatial scenes. This model has to be fast enough for online monitoring of speech intelligibility inside a device-internal model loop. The model also has to be non-intrusive (also called "blind"), which means that the model does not need the speech or noise signals separately nor any other auxiliary information about the auditory scene.

The blind binaural speech intelligibility model (bBSIM) as described by [2] was the starting point for this study. In this model, the left and right input signals are divided into frequency bands using a gammatone filterbank. Furthermore, noise that simulates the listener's hearing threshold is added. This threshold simulating noise is uncorrelated between left and right ear, in or-

der to avoid, that it can be canceled out by the following binaural processing of the model. For the upper frequency bands (> 1.5 kHz) there is a blind selection of the better ear, which is based on the speech-likeness of the amplitude modulations. For the lower frequency bands (< 1.5 kHz), an equalization and cancellation (EC) process is involved. This means that the interaural time and level differences (ITD, ILD) are equalized and that the left and right ear signals are subtracted from each other for a cancellation or suppression of the strongest noise source. Alternatively, the left and right ear signal are added for amplifying the dominant speech source. The selection between the alternative subtracting or adding processes is again based on a blind maximization of the speech-likeness of the amplitude modulations of the resulting signal. Note, that the resulting mono signal is a speech enhanced signal which considers human better ear listening and which considers binaural unmasking at frequencies below 1.5 kHz. Subsequently, this resulting signal is fed into a backend which determines speech intelligibility.

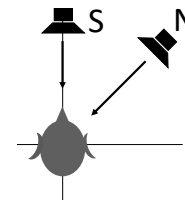


Figure 2: A subject receives a speech signal (S) coming from the front and a noise signal (N) coming from an angle in the horizontal plane.

The benefit that is achieved by the cancellation stage (which performs the subtraction or summation) strongly depends on the interaural gains and interaural time shifts that are applied in the equalization stage. Note, that the interaural time shifts can be replaced by interaural phase shifts, which make virtually no difference as the shifts are applied to the band limited outputs of the gammatone filterbank. Fig. 2 shows an example with speech coming from the front and noise coming from an angle in the horizontal plane. Without any equalization, in an anechoic situation, the subtraction would cancel out the speech signal perfectly, leading to the worst possible signal-to-noise ratio (SNR). A perfect equalization of the interaural time and level differences of the noise, on the other hand, would lead to a cancellation of the noise and a residual speech signal leading to an unrealistic high SNR. Since

human binaural processing is not perfect, interaural inaccuracies according to [3] are assumed by the model.

Interaural Mistuning

In the past, these inaccuracies were realized as Monte Carlo simulations in a stochastic manner and the entire model had to be used in several instances to average results [4, 2]. For an application in the device-internal model loop, there is not enough time for this approach. Therefore, we have replaced the repeated stochastic inaccuracies of the model with one representative deterministic inaccuracy, which we called ‘mistuning’, as the theoretically optimal set of interaural equalization parameters is replaced by a slightly mistuned set. The direction of the mistuning is chosen always towards an overcompensation of the ITD and ILD to preserve more of speech if it is coming from the front.

Evaluation

We evaluated the mistuning approach using speech recognition threshold (SRT) data in spatial auditory scenes by [4]. We used two alternative backends to predict the speech intelligibility:

Firstly, we used the Speech Intelligibility Index (SII) [5], which is an SNR-based model. As the SII is not blind, it is not suitable for the in-the-device approach. However, we used this approach for a first evaluation as we know from previous studies that the original Monte Carlo method of realizing the binaural processing inaccuracies works well with the SII.

Secondly, we used a blind phoneme classifier, called the Mean Temporal Distance (MTD) [6], which is in principle applicable in real-time on a hearing device and which is described in more detail in the next paragraph.

Mean Temporal Distance Backend (MTD)

The MTD backend calculates a phoneme posteriorgram, which is a measure of certainty of the automatically classified phonemes in each time frame. Two phoneme probability vectors are compared at a distance of 350 to 800 ms. This is done for comparing the phoneme probabilities at different points in time. This enables to quantify the certainty as well as the change in classified phonemes over time, which is important because distortion of speech due to noise or reverberation may lead to less pronounced contrasts and smaller temporal changes in the posteriorgram. This results in a lower MTD value. The MTD value is sometimes also referred to as the \bar{M} -Measure. Based on this measure, it is possible to predict speech intelligibility as well as listening effort using different mapping functions. [7]

Prediction of Speech Recognition Thresholds

To predict SRT using the SII or the MTD backend, a reference SII or MTD value is needed, respectively. In the following, we describe the procedure for the MTD. Since there is a certain variance in MTD values across different sentences at a certain SNR (vertical stacks of colored dots in Fig. 3), the mean value was taken over ten sentences, in order to obtain more stable results in this study. From these mean values, a smoothing interpolation was derived to obtain a mapping function between SNR and MTD. As reference for calibrating the model, the SRT from [4]

for collocated speech and noise (S_0N_0) in an anechoic condition was used and transformed to a reference MTD value with this function (big green dot in Fig. 3).

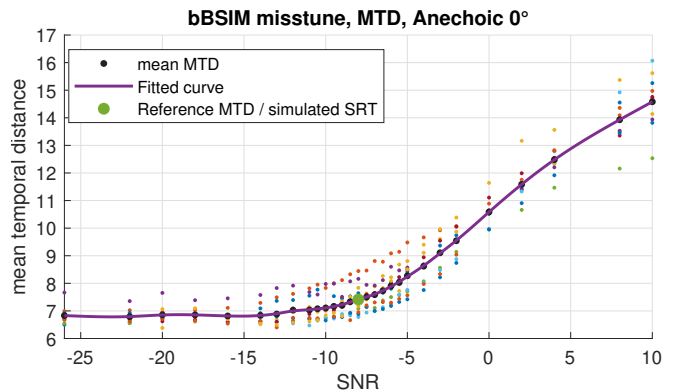


Figure 3: Calculated MTD for single sentences (colored dots) at different SNR values (horizontal axis), mean across ten sentences (black dots), and smoothed interpolation (fitted purple curve). The big green dot indicates the reference point, which corresponds to the SNR at which listeners with normal hearing can correctly recognize 50% of the words (SRT).

Mapping functions for all spatial conditions of [4] were simulated in the same way and used to obtain the SNRs corresponding to the simulated SRTs.

Results

Clean speech and noise of the German matrix sentence test were convolved with head related impulse responses measured with an artificial head from [4] and fed into the model’s inputs for a direct comparison. The acoustic scenarios were: speech virtual always from the front and noise from different angles in the horizontal plane.

SRTs predicted using SII

In the anechoic environment (upper panel in Fig. 4) a very large binaural benefit can be seen in the measurement, which the model does predict as well. In the more reverberant office room (middle panel in Fig. 4) the binaural benefit is much smaller and the model underestimates it even further. In the larger cafeteria (lower panel in Fig. 4) the binaural benefit is again larger due to the more direct sound and asymmetric due to a window wall on one side. The asymmetric benefit is also predicted by the model despite a small underestimation.

Two different manipulations of the standard SII were tested, which both led to better predictions:

- Raising the lower limit of the -15 to 15 dB SNR range which is considered by the SII within each frequency band. This led to a more glimpse-like approach.
- Weighted summation of the SNR of the frequency bands not in the logarithmic (dB) domain, but in the linear domain. This method takes the higher SNRs into greater consideration.

Both optimizations are not discussed here in detail, since the SII does not fulfill the requirement of blindness. Nevertheless, these results suggest that the mistuning in principle works well.

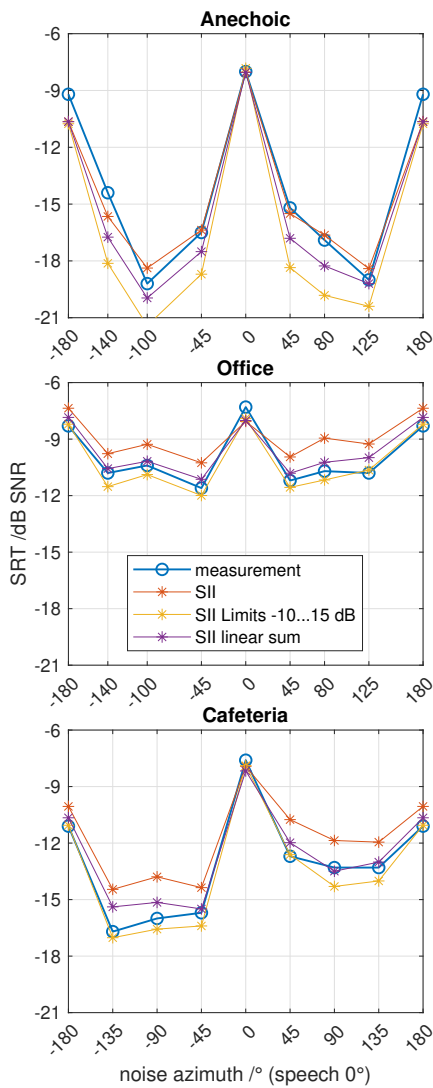


Figure 4: SRTs with noise from different directions in the horizontal plane (azimuth) and speech from the front in three acoustic environments. Mean measurement data from [4] (blue circles), predictions with the standard SII backend (orange stars), a modified SII with limited SNR range (yellow stars), and a modified SII with linear summation of SNR (purple stars).

SRTs predicted using MTD

The completely blind model with mistuning and MTD backend (green triangles in Fig. 5) is compared to the measurements (blue circles), the best model version from above with modified SII backend (purple stars) and the original model predictions from [4] with intrusive front- and backend and with Monte-Carlo simulations.

In the anechoic environment (upper panel in Fig. 5) all models tend to slightly underestimate the measured SRT and perform similarly well. In the office (middle panel in Fig. 5), however, five out of eight predictions of the completely blind model (purple stars) are further away from the measured values (maximum 2.7 dB) than those of the non-blind references. In the cafeteria (lower panel in Fig. 5) most predictions of the blind model are close to the measured SRT and three show a deviation of at maximum 2.6 dB.

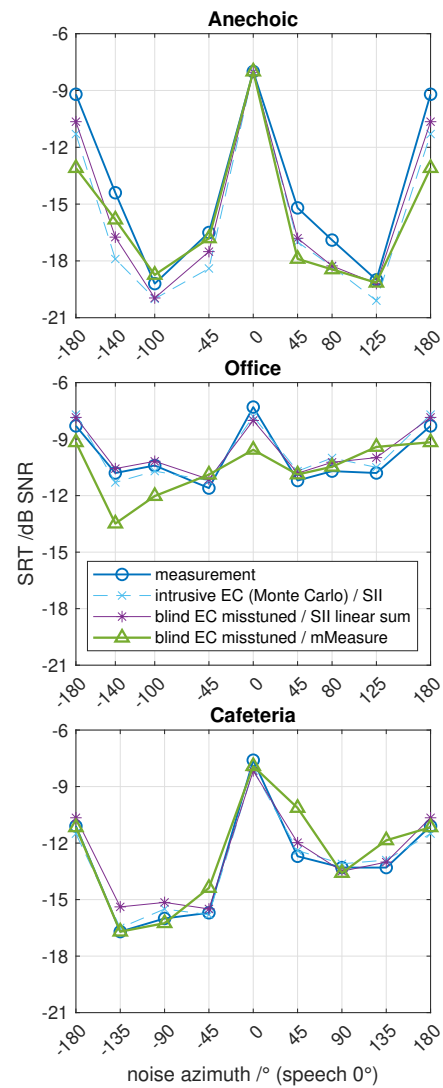


Figure 5: SRTs with noise from different directions in the horizontal plane (azimuth) and speech from the front in three acoustic environments. Mean measurement data from [4] (blue circles), predictions of the complete intrusive EC/SII model from [4] (light blue crosses), predictions of the blind mistuned frontend with the modified SII with linear summation of SNR (purple stars), and predictions of the complete blind model with mistuned frontend and the MTD backend.

Conclusion

In this study, we replaced the Monte Carlo simulations that have so far been used to model human inaccuracies in our Binaural Speech Intelligibility Model (BSIM) by a defined mistuning of the model’s interaural equalization parameters. This eliminates the need of iterating over several stochastic realizations as performed in [8] of processing errors, which makes the novel real-time capable. In combination with the MTD backend, this model is fully blind so that it can be integrated in the device-internal model loop.

One reason why some predictions using the MTD backend show larger deviations to the measured SRT than with using the SII backend may be due to the fact that the mapping function between MTD and SNR is quite shallow at 50% intelligibility. Consequently, the

SNR value related to this threshold (see green point in Fig. 3) is not very well defined and difficult to find. However, in real-life applications, intelligibility values larger than 50% are much more relevant and typically hearing aids will be controlled to enable higher intelligibility values. Consequently, this problem will probably be much smaller in real-life conditions. Probably, the MTD would give more accurate results at higher, more realistic speech intelligibilities. Note, that this approach of non-intrusive modelling of binaural processing inaccuracies does also work for the estimation of listening effort, which is related to higher MTD values.

Acknowledgement

HAPPAA is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project ID 352015383 – SFB 1330 A1

Contact: sfb1330@uni-oldenburg.de

Website: www.hearing-acoustics.de / uol.de/sfb1330

Simulations were conducted on the HPC cluster ROSA funded by the DFG under INST 184/225-1 FUGG.

References

- [1] Hohmann, V., Paluch, R., Krueger, M., Meis, M. and Grimm, G.: The Virtual Reality Lab: Realization and Application of Virtual Sound Environments. *Ear and Hearing* 41: 31S-38S, November/December 2020. doi: 10.1097/AUD.0000000000000945
- [2] Hauth, C. F., Berning, S. C., Kollmeier, B., and Brand, T.: Modeling binaural unmasking of speech using a blind binaural processing stage. *Trends in Hearing*, 24, (2020) doi: 10.1177/2331216520975630
- [3] vom Hövel, H.: Zur Bedeutung der Übertragungseigenschaften des Außenohrs sowie des binauralen Hörsystems bei gestörter Sprachübertragung (1984) Dissertation, Fakultät für Elektrotechnik, RTWH Aachen.
- [4] Beutelmann, R., and Brand, T.: Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners, *J. Acoust. Soc. Am.* vol. 120, (2006) pp. 331–342.
- [5] ANSI S3.5.: American National Standard - Methods for the calculation of the Speech Intelligibility Index (1997)
- [6] Castro Martinez, A. M., Spille, C., Roßbach, J., Kollmeier, K. and Meyer, B. T.: Prediction of speech intelligibility with DNN-based performance measures. *Computer Speech & Language* 74 (2022) doi: 10.1016/j.csl.2021.101329
- [7] Rennie, J., Röttges, S., Huber, R., Hauth, C. F., and Brand, T.: A joint framework for blind prediction of binaural speech intelligibility and perceived listening effort. *Hearing Research* 426 (2022), 1-13. doi: 10.1016/j.heares.2022.108598
- [8] Roßbach, J., Röttges, S., Hauth, C. F., Brand, T., and Meyer, B. T.: Non-intrusive binaural prediction of speech intelligibility based on phoneme classification. (2021) In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 396-400). IEEE.