Performance Prediction of the Binaural MVDR Beamformer with Partial Noise Estimation using a Binaural Speech Intelligibility Model

Christopher F. Hauth [1], *Nico Gößling* [2], *Thomas Brand* [1]

[1] Medizinische Physik and Cluster of Excellence Hearing4All, University of Oldenburg, Oldenburg, Germany
[2] University of Oldenburg, Department of Medical Physics and Cluster of Excellence Hearing4All, Oldenburg, Germany
Email: {christopher.hauth,nico.goessling,thomas.brand}@uni-oldenburg.de

Abstract

An objective evaluation of binaural noise reduction algorithms allows for directly comparing the performance of different algorithm realizations. Here, a binaural speech intelligibility model (BSIM), which mimics the effective binaural processing of a human listeners, is used to predict the performance of the binaural minimum-variance distortionless response beamformer with partial noise estimation (BMVDR-N), which aims at preserving the speech component in a reference microphone and a scaled version of the noise component. The BMVDR-N beamformer is evaluated with respect to a predicted change in SRT depending on the parameter η, which controls a trade-off between noise reduction and binaural cue preservation of the noise component. The results show that BSIM benefits from the preserved binaural cues suggesting that the BMVDR-N beamformer can improve the spatial quality of a scene without affecting speech intelligibility.

1 Introduction

In everyday life, human listeners have to deal with complex acoustic scenarios, with multiple interfering sources located at different places in the surrounding. Binaural information is especially useful in these situations, because interaural time differences (ITD) and interaural level differences (ILD) can be used by the human binaural auditory system to separate the target from interfering sources. Two mechanism are thought to play a role in binaural processing: first, listening with the ear, which has a favorable signal-to-noise ratio (SNR), which is often referred to as better-ear listening and second, binaural unmasking, where the binaural system uses interaural disparities between target and interfering source to effectively enhance the SNR.

Hearing impaired listeners are often provided with hearing aids, which use the interaural differences in the signals picked up by the microphones to perform spatial filtering (i.e. beamforming) in order to improve the SNR. In this study, the binaural minimum-variance distortion-less response (BMVDR) [1] beamformer is considered, which minimizes the output power, while the target-speech is not distorted. In this case, the interaural parameters of the background noise are discarded and it is perceived as coming from the target-speech direction. An extension of the BMVDR beamformer is the BMVDR-N beamformer, which combines the classical approach with a partial noise estimation [2], which combines the classical BMVDR approach with a partial noise estimation. Here, a portion of the noisy input signal is added to the processed signal in order to partially preserve the binaural cues of the background noise at the cost of noise reduction performance and thereby improve the spatial impression of the scene. However, the SNR improvement is limited by the amount of added noise and, therefore, objective measures, like the intelligibility-weighted SNR (iSNR), predict decreased performance. If human binaural processing is taken into account, the added noise not necessarily leads to a worse performance. The binaural system might be very well able to benefit from the noise portion: On the one hand, it improves the spatial impression of the processed signal, which might improve the perceptual segregation of target from the interfering signal; on the other hand, the additional noise might be suppressed by the human binaural noise reduction by exploiting differences in ITD and ILD between speech and noise [3, 4].

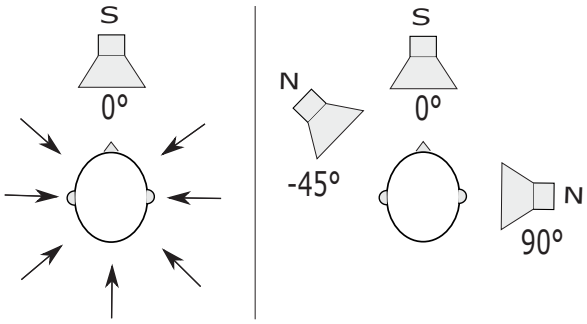

Figure 1: Speech intelligibility predictions were obtained for speech coming from 0° in the horizontal plane in non-stationary diffuse noise (left panel) and for speech in coherent stationary noise located either at −45° or 90° in the horizontal plane (right panel).

In order to investigate interaction between noise reduction and partial noise estimation performed by the BMVDR-N beamformer and the human auditory processing on speech reception thresholds (SRTs), a binaural speech intelligibility model (BSIM) [5] is used, where effective binaural processing using the equalization-cancellation (EC) mechanism [6] is combined with the speech intelligibility index (SII) [7]. It is hypothesized that human binaural processing benefits from the cues provided by the partial noise estimate in the BMVDR-N beamformer and thus the reduced SNR is partially compensated for. Moreover, BSIM is used to estimate the best fitting trade-off parameter for both coherent as well as diffuse noise sources. The results obtained via model predictions can also be used to condense the parameters, which should be tested in listening experiments.

2 BMVDR-N beamformer and acoustic scenario

The BMVDR beamformer aims at minimizing the noise output power (minimum variance), while leaving the target direction undistorted (distortionless response). As a consequence of this processing, the spatial characteristics of the noise are destroyed, such that the binaural cues of the

ISBN 978-3-8007-4767-2

signal, which is presented to a listener, are the same for target signal and noise. Therefore, differences in the binaural cues of speech and noise are not available to the listener and thus, the noise component is perceived as coming from the same direction as the target source. In order to overcome this restriction, the BMVDR beamformer with partial noise estimation (BMVDR-N) was proposed, where a portion of the unprocessed input signal is mixed with the BMVDR beamformer output. By doing so, the spatial impression of the scenario can be better preserved than in the BMVDR beamformer. This trade off between noise reduction and binaural cue preservation is controlled by the parameter $0 \leq \eta \leq 1$, where for $\eta = 0$ the output is the BMVDR beamformer and for $\eta = 1$ the output equals the unprocessed input signal. In this study, the effect of the trade-off parameter η on predicted SRTs is investigated for two scenarios, which are schematically shown in 1: In the first scenario, a spatially diffuse and temporally non-stationary noise was considered, which was the recording of the ambient noise in a cafeteria with a $T_{60} \approx 1250$ms [8]. In the second scenario, a spatially coherent and temporally stationary noise with the same long-term spectrum as the speech material was used. The spatially coherent noise was either located at $-45°$ or $90°$ in the horizontal plane. In both tested scenarios, sentences from the Oldenburg sentence test in noise (OlSa) [9–11] were used as target speech material coming from the frontal direction. The BMVDR and BMVDR-N were implemented assuming a cylindric isotropic (diffuse) noise field, where the needed spatial coherence matrix was calculated as in [2] using the measured anechoic impulse responses from hearing aids mounted to a dummy head [8]. To steer the beamformers towards the frontal direction, the corresponding anechoic impulse responses from the same database were used. All signals were processed in the short-time frequency domain using a frame length of 16 ms with 50% overlap and a square-root Hann window. The sampling rate was 16 kHz. For speech intelligibility predictions, all signals were re-sampled to 44.1 kHz.

3 Binaural speech intelligibility model - BSIM

The general framework of the binaural speech intelligibility model (BSIM) including the tested scenario is shown in Figure 2. After the device signal processing (cf. Section 2), the left and right ear signal are fed to the BSIM model. In a first stage, a gammatone filterbank [12] consisting of 30, one ERB [13] wide filters, ranging from 146 to 8346 Hz is applied in order to simulate the frequency selectivity of the basilar membrane. The binaural processing stage is realized as equalization-cancellation (EC) process [6], which is applied in each frequency band independently. This EC process equalizes the left and right ear signal in time and level (equalization step). Then, the left ear signal is subtracted from the right ear signal (or vice versa). The parameters for equalization are optimized in order to maximize the SNR of the EC processed signal. It can be interpreted as beamformer steering a null into the direction of the interfering noise. However, the EC-process cannot be assumed as perfect operation and, therefore, uncertainties in level ($\epsilon_{L,R}$) and time ($\delta_{L,R}$) for the left and right ear are incorporated mirroring inherent processing inaccuracies of the human binaural system. These processing errors are assumed to be normally distributed random variables, which are defined by zero mean and standard deviations σ_δ and σ_ϵ. The standard deviation of the normally distributed processing errors is given by

$$\sigma_\epsilon = \sigma_{\epsilon 0}\left[1+(\frac{|\alpha|}{\alpha_0})^p\right], \tag{1}$$

and

$$\sigma_\delta = \sigma_{\delta 0}\left[1+\frac{|\Delta|}{\Delta_0}\right], \tag{2}$$

with $\sigma_{\epsilon 0} = 1.5$dB, $\alpha_0 = 15$dB, $p = 1.6$, $\sigma_{\delta 0} = 65\mu$s and $\Delta_0 = 1.6$ms [14]. These equations state that the standard deviation of level and delay errors increases with increasing time or level differences between both ears. The intensity of the EC processed speech in a single gammatone filter is obtained according to

$$I(S_{EC}(\Omega)) = \int_{\Omega-\beta/2}^{\Omega+\beta/2} |e^{\gamma/2+\epsilon_L+j\omega(\tau/2+\delta_L)}S_L(\omega) - e^{-\gamma/2+\epsilon_R-j\omega(\tau/2-\delta_R)}S_R(\omega)|^2 \mathrm{d}\omega, \tag{3}$$

where $I(S_{EC})$ is the band limited intensity of the speech signal after EC processing, Ω denotes the center frequency of a gammatone filter and β its bandwidth. $S_{\{L/R\}}(\omega)$ denotes the frequency domain representation of the speech signal of the left and right ear, respectively. γ and τ are the EC parameters in level and time in order to maximize the SNR, which are jittered by the binaural inaccuracies ϵ and δ and applied symmetrically to the left and right ear. By applying

$$|x-y|^2 = |x|^2+|y|^2-2\Re(xy^*), \tag{4}$$

the EC output is obtained with respect to the expected values of the uncertainties in level and time according to

$$\begin{aligned} &< I(S_{EC}(\Omega)) >_{\epsilon_L,\epsilon_R,\delta_L,\delta_R} = \\ &e^{2\sigma_\epsilon^2}e^{\gamma} I(S_L(\Omega)) \\ &+e^{2\sigma_\epsilon^2}e^{-\gamma} I(S_R(\Omega)) \\ &-2e^{\sigma_\epsilon^2}\cdot\Re\left(\int_{\Omega-\beta/2}^{\Omega+\beta/2} S_L(\omega)S_R^*(\omega)e^{-\omega^2\sigma_\delta^2}e^{j\omega\tau}d\omega\right), \end{aligned} \tag{5}$$

where $I(S_{\{L,R\}})$ denotes the intensity of the speech in the left and right gammatone filter, * denotes the complex conjugate and $e^{2\sigma_\epsilon^2}$ and $e^{-\omega^2\sigma^2}$ denote the expected values of the binaural processing inaccuracies. The same equation is applied to the noise component in order to obtain the band limited intensity of the noise after EC processing. For frequencies above 1500 Hz, the ear providing the higher SNR is considered for further processing, because better-ear listening dominates binaural hearing. This is also in line with the duplex theory of sound localization, which states that ITD information dominates localization below 1500 Hz, while ILD information provides a stronger cue at frequencies above 1500 Hz due to head shadow effects [15]. For each frequency band, the SNR is computed for the left, right and the binaurally processed signals. The best SNR is then fed to the SII [7], which converts frequency dependent SNRs into a value between 0 and 1. This value needs to be fitted to a reference condition, which usually is a scenario with co-located speech and noise sources. In the tested scenarios, the BMVDR processed signal served as reference condition, because the relative change in SRT

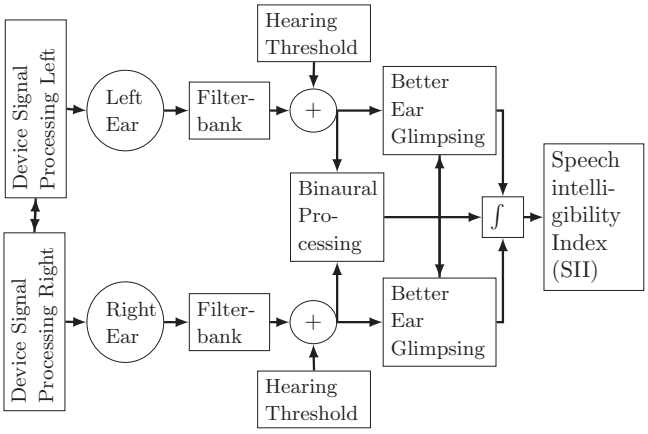

Figure 2: General processing scheme of the signal processing of BSIM. First, the speech and noise signals are processed with a signal processing device, which is the binaural MVDR-N here. Then the model processing begins: The incoming signal is decomposed into 30 frequency bands ranging from 146 to 8346 Hz. In each frequency band, the EC mechanism is applied to model the effective binaural processing below 1500 Hz. As a consequence of the binaural processing inaccuracies, better ear listening plays a dominant role above 1500 Hz. In the end, the SII is applied, which converts frequency dependent SNRs into a value between 0 and 1.

with increasing η was of interest. In the SII, the SNRs are weighted according to the human speech perception. The BSIM can be applied in two modes: a long-term mode, where the whole signal is considered as a single time frame. This method can be used for temporally stationary maskers in spatially stationary scenarios. The short-term mode uses time frames of 23 ms with 50% overlap leading to an effective window length of approx. 12 ms, which is similar to the frequency independent time window used in short-time SII proposed by [16]. This version should be used for non-stationary maskers. The BSIM model has been shown to provide very good predictions for SRTs obtained in various acoustic conditions and for listeners with normal hearing and hearing impairment [5, 17, 18].

4 Results

In this section, BSIM predictions of the BMVDR beamformer processed signals are presented for the spatially incoherent, temporally instationary noise (Section 4.1) and for the spatially coherent, temporally stationary noise (Section 4.2).

4.1 Predictions for BMVDR-N processed speech in diffuse noise

In Figure 3, the predicted change in SRT is shown as a function of η ranging from 0 (BMVDR) to 1 (unprocessed) in steps of 0.05. Additionally, the results obtained for the $BMVDR_{opt}$ is shown, which artificially combines the BMVDR solution with a perfect preservation of the binaural cues. The $BMVDR_{opt}$ processed signal is obtained by adjusting the global SNR of the unprocessed signal to the SNR at the output of the BMVDR beamformer. Moreover, predictions for a frequency dependent and psychoacoustically motivated boundary of the magnitude-squared coherence (MSC) are shown, which
were proposed in [19]. With increasing η, the predicted SRT is slightly improved up to an η of 0.2, but then decreases linearly (on a dB axis) with increasing η until approximately -3 dB for the unprocessed signal (or an η equal to 1). Note, that the short-time version of BSIM was used. The results of the simulations suggest that the trade-off parameter η can be increased up to 0.6 without loosing performance in predicted SRTs, even though the output SNR gets worse. The largest improvement in the range of 3 dB in SRT was obtained with the $\text{BMVDR}_{\text{opt}}$. For the MSC boundaries of MSC0.36 almost the same SRT as for the BMVDR processed signal is predicted. For a MSC boundary of MSC0.04, SRT get worse by approximately 1.5 dB. As these relative changes are only simulations, the result was compared to data collected by [20], where listening experiments were performed for a subset of the tested parameters, namely the BMVDR, the $\text{BMVDR}_{\text{opt}}$, the unprocessed signal and for the fixed MSC boundaries of 0.04 and 0.36. In Figure 4, the predicted SRTs are shown along with the SRTs measured for 15 listeners with normal hearing. The unprocessed signal was used for model calibration and fits the median across the 15 listeners. Predictions of BSIM were compared to measured data based on the coefficient of determination

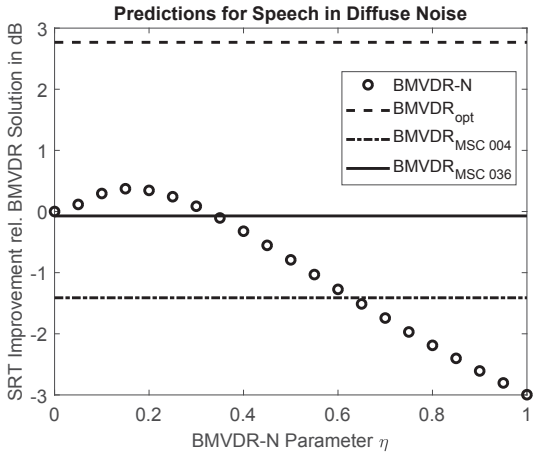

Figure 3: Predicted change in speech reception threshold (SRT) relative to the BMVDR solution for values of η ranging from 0 to 1. All results were obtained for speech in ambient cafeteria noise, which can be considered as temporally non-stationary and diffuse.

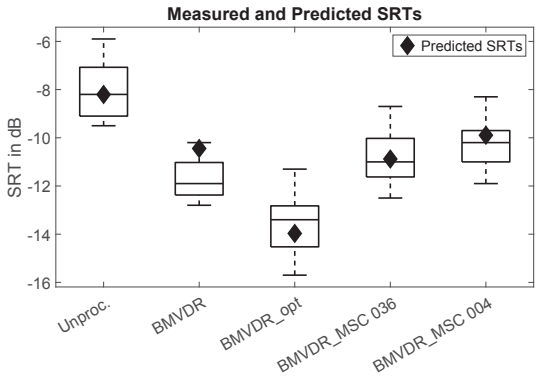

Figure 4: Measured and predicted SRTs for the different version of the BMVDR-N beamformer for OlSa sentences in ambient cafeteria noise. The boxplots depict data for 15 listeners with normal hearing, black diamonds indicate predicted SRTs by BSIM.

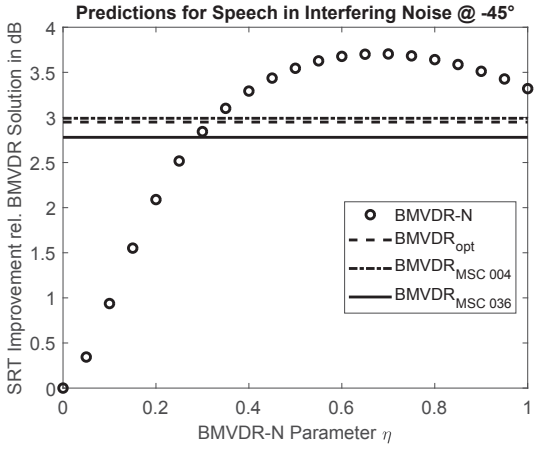

Figure 5: Predicted change in speech reception threshold (SRT) relative to the BMVDR solution for values of η ranging from 0 to 1. All results were obtained for speech in speech-shaped noise, which can be considered as spatially coherent and stationary. The interferer was located at 90° in the horizontal plane.

R^2 and the root mean square error (RMSE) between the median SRTs and predicted SRTs. An $R^2 = 0.81$ was obtained, while the root mean square error (RMSE) was below 1 dB (RMSE = 0.79 dB). In summary, the predictions obtained for those scenarios, which were also used for listening experiments, suggest that BSIM can be used to predict SRTs for processed speech and, therefore, for comparing different binaural beamforming algorithms.

4.2 Predictions for BMVDR-N processed speech in spatially coherent noise

In Figure 5, the relative change of the SRT in dB is shown as a function of the trade-off parameter η for a coherent noise source at 90°, while speech again is coming from the frontal direction. In this condition, the SRT achieved using the unprocessed signal or the BMVDR processed signal are nearly identical. This was expected, as the EC process or Null steering beamformer of BSIM can effectively cancel the coherent noise source. The SRT is improved up to 2 dB for an $\eta = 0.4$. If η is further increased, SRTs tend to decrease. For a spatially coherent and temporally stationary noise source at 90°, the BMVDR-N beamformer with an η of 0.4 provides the best predicted SRT. Besides the interferer at 90° also an interfering noise at −45° was tested. The predictions are shown in Figure 6. In this scenario, the predicted

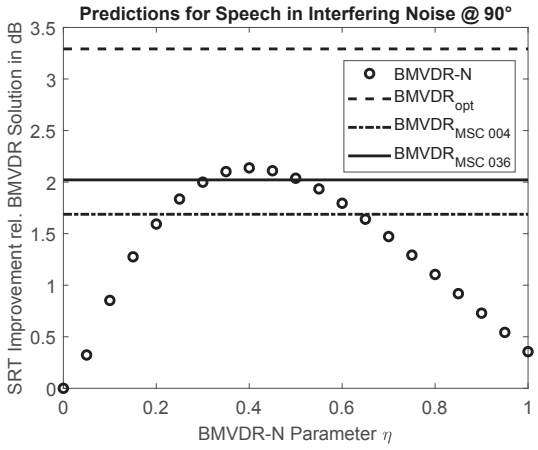

Figure 6: Predicted change in speech reception threshold (SRT) relative to the BMVDR solution for values of η ranging from 0 to 1. All results were obtained for speech in speech-shaped noise, which can be considered as spatially coherent and stationary. The interferer was located at −45° in the horizontal plane.

SRT for the unprocessed signal is approx. 3 dB lower (i.e. better) than for the BMVDR processed signal. The effects observed in Figure 5 and Figure 6 are a result of the different restrictions of the EC process and the BMVDR beamformer. While the BMVDR is restricted to the distortionless response, this is not the case for the EC process, which allows for a distortion of the speech component. If the BMVDR-N is applied, the predicted SRTs are decreased up to an $\eta = 0.65$, where the lowest predicted SRT is obtained. In this special case, the $BMVDR_{opt}$ is not the best, as the predicted SRT for the BMVDR processed signal is worse than for the unprocessed signal. Also note that the BMVDR beamformer is explicitly implemented for diffuse noise fields. Therefore, binaural beamformer designed for coherent noise sources, for example the binaural linear constraint minimum variance (BLCMV) beamformer [21] can be expected to achieve a higher performance.

5 Conclusion

In this study it was shown that a binaural speech intelligibility model can be used to evaluate the performance of a binaural beamformer. It has the advantage of taking the effective human binaural processing into account. The effect of different values of the trade-off parameter η on predicted SRTs was investigated for two acoustic scenarios, where an temporally non-stationary and diffuse noise was used in the first scenario and a spatially coherent and temporally stationary noise in the second scenario. In summary, η can be chosen to be in the range of 0.2 to 0.6 without substantially increasing SRTs in the diffuse noise scenario. In the coherent noise scenario, predictions suggest to apply an η of 0.6, because predicted SRTs are substantially reduced compared to the BMVDR processed signal. However, it is important to note that the BMVDR is explicitly implemented for diffuse noise fields.

ISBN 978-3-8007-4767-2

References

[1] S. Doclo, S. Gannot, D. Marquardt, and E. Hadad, "Binaural Speech Processing with Application to Hearing Devices," in *Audio Source Separation and Speech Enhancement*, ch. 18, Wiley, 2018.

[2] D. Marquardt and S. Doclo, "Interaural Coherence Preservation for Binaural Noise Reduction Using Partial Noise Estimation and Spectral Postfiltering," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 26, no. 7, pp. 1257–1270, 2018.

[3] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[4] A. W. Bronkhorst and R. Plomp, "The effect of head-induced interaural time and level differences on speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 83, no. 4, pp. 1508–1516, 1988.

[5] R. Beutelmann, T. Brand, and B. Kollmeier, "Revision, extension, and evaluation of a binaural speech intelligibility model," *The Journal of the Acoustical Society of America*, vol. 127, no. 4, pp. 2479–2497, 2010.

[6] N. I. Durlach, "Equalization and cancellation theory of binaural masking level differences," *The Journal of the Acoustical Society of America*, vol. 35, no. 8, pp. 1206–1218, 1963.

[7] ANSI S3.5-1997, *Methods for Calculation of the Speech Intelligibility Index*. American National Standard (ANSI), 1997.

[8] H. Kayser, S. Ewert, J. Annemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel In-Ear and Behind-The-Ear Head-Related and Binaural Room Impulse Responses," *Eurasip Journal on Advances in Signal Processing*, vol. 2009, p. 10 pages, 2009.

[9] K. Wagener, T. Brand, V. Kühnel, and B. Kollmeier, "Entwicklung und Evaluation eines Satztests für die Deutsche Sprache I: Design des Oldenburger Satztests (Development and Evaluation of a Sentence Test for the German Language I: Design of the Oldenburg Sentence Test)," *Z. Für Audiologie, Audiological Acoust.*, vol. 38, pp. 4–15, 1999a.

[10] K. Wagener, T. Brand, V. Kühnel, and B. Kollmeier, "Entwicklung und Evaluation eines Satztests für die Deutsche Sprache II: Optimierung des Oldenburger Satztests (Development and Evaluation of a Sentence Test for the German Language II: Optimization of the Oldenburg Sentence Test)," *Z. Für Audiologie, Audiological Acoust.*, vol. 38, pp. 44–56, 1999b.

[11] K. Wagener, T. Brand, V. Kühnel, and B. Kollmeier, "Entwicklung und Evaluation eines Satztests für die Deutsche Sprache III: Evaluation des oldenburger Satztests (Development and evaluation of a sentence test for the german language III: Evaluation of the oldenburg sentence test)," *Z. Für Audiologie, Audiological Acoust.*, vol. 38, pp. 86–95, 1999c.

[12] V. Hohmann, "Frequency analysis and synthesis using a gammatone filterbank," *Acta Acustica united with Acustica*, vol. 88, pp. 433–442, 2002.

[13] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, pp. 103 – 138, 1990.

[14] H. vom Hövel, *Zur Bedeutung der Übertragungseigenschaften des Aussenohrs sowie des binauralen Hörsystems bei gestörter Sprachübertragung*. PdD Dissertation, RWTH Aachen, 1984.

[15] E. A. Macpherson and J. C. Middlebrooks, "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited," *The Journal of the Acoustical Society of America*, vol. 111, no. 5, pp. 2219–2236, 2002.

[16] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2181–2192, 2005.

[17] J. Rennies, A. Warzybok, T. Brand, and B. Kollmeier, "Modeling the effects of a single reflection on binaural speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 135, no. 3, pp. 1556–1567, 2014.

[18] J. Rennies, T. Brand, and B. Kollmeier, "Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet," *The Journal of the Acoustical Society of America*, vol. 130, no. 5, pp. 2999–3012, 2011.

[19] D. Marquardt, V. Hohmann, and S. Doclo, "Interaural Coherence Preservation in Multi-channel Wiener Filtering Based Noise Reduction for Binaural Hearing Aids," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, pp. 2162–2176, Dec. 2015.

[20] D. Marquardt, *Development and Evaluation of Psychoacoustically Motivated Binaural Noise Reduction and Cue Preservation Techniques*. PhD Dissertation,Fakultät für Medizin und Gesundheitswissenschaften, Carl von Ossietzky Universität Oldenburg, 2015.

[21] E. Hadad, S. Doclo, and S. Gannot, "The Binaural LCMV Beamformer and its Performance Analysis," *IEEE/ACM Trans. on Audio, Speech, and Language Proc.*, vol. 24, no. 3, pp. 543–558, 2016.

ISBN 978-3-8007-4767-2