Trends in Hearing
Volume 24: 1–16
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2331216520975630
journals.sagepub.com/home/tia
SAGE

Modeling Binaural Unmasking of Speech Using a Blind Binaural Processing Stage

Christopher F. Hauth, Simon C. Berning, Birger Kollmeier and Thomas Brand

Abstract

The equalization cancellation model is often used to predict the binaural masking level difference. Previously its application to speech in noise has required separate knowledge about the speech and noise signals to maximize the signal-to-noise ratio (SNR). Here, a novel, *blind* equalization cancellation model is introduced that can use the mixed signals. This approach does not require any assumptions about particular sound source directions. It uses different strategies for positive and negative SNRs, with the switching between the two steered by a blind decision stage utilizing modulation cues. The output of the model is a single-channel signal with enhanced SNR, which we analyzed using the speech intelligibility index to compare speech intelligibility predictions. In a first experiment, the model was tested on experimental data obtained in a scenario with spatially separated target and masker signals. Predicted speech recognition thresholds were in good agreement with measured speech recognition thresholds with a root mean square error less than 1 dB. A second experiment investigated signals at positive SNRs, which was achieved using time compressed and low-pass filtered speech. The results demonstrated that binaural unmasking of speech occurs at positive SNRs and that the modulation-based switching strategy can predict the experimental results.

Keywords

speech recognition thresholds, binaural processing, auditory model, speech intelligibility prediction

Received 26 April 2019; Revised 29 October 2020; accepted 2 November 2020

In everyday life, human listeners deal with complex acoustic scenarios, in which target speech and interfering sound sources arise at different locations. This phenomenon is termed the “cocktail party problem” (Cherry, 1953, p. 976). When target speech and interfering sound sources are spatially separated, they differ in their interaural level differences (ILDs), interaural time differences (ITDs), and interaural phase differences (IPDs; Bronkhorst, 2000). Two mechanisms are thought to play a primary role in such adverse spatial acoustic conditions: *better-ear listening* (i.e., listening with the ear that receives the better signal-to-noise ratio [SNR]) and *binaural unmasking* (i.e., using ITD and ILD differences for separating the target from interfering signals). Binaural unmasking is often modeled by the equalization cancellation (EC) mechanism (Durlach, 1963), which equalizes ITDs and ILDs, then calculates the difference between left ear channel and right ear channels. In many computational models, the EC mechanism is combined with a band pass filter bank, which mimics auditory frequency selectivity on the basilar membrane, and some metric quantifying the amount of speech information in the signal to predict the speech recognition threshold (SRT). This metric is often referred to as the *back end*.

Particular back ends include the speech intelligibility index (SII; ANSI S3.5-1997, 1997; (Beutelmann & Brand, 2006; Beutelmann et al., 2010; Jelfs et al., 2011; Lavandier & Culling, 2010; Lavandier et al., 2012; Wan et al., 2010), analyzing the SNR in the modulation frequency domain (Chabot-Leclerc et al., 2016), and the correlation between the clean speech and the noisy

Medizinische Physik and Cluster of Excellence Hearing4All Carl-von-Ossietzky Universität Oldenburg, Oldenburg, Germany

Corresponding author:
Christopher F. Hauth, Medizinische Physik and Cluster of Excellence Hearing4All Carl-von-Ossietzky Universität Oldenburg, 26111 Oldenburg, Germany.
Email: christopher.hauth@uni-oldenburg.de

speech (Andersen et al., 2016). All of these models are based on assumptions about the EC process made by Durlach (1963), namely that the equalization process has inherent processing errors in level and time, leading to an imperfect alignment of the left and right ear signals and, therefore, to an imperfect cancellation of the masker signal. Durlach's original specifications of the error parameters were subsequently updated by vom Hövel (1984) to better agree with data by Langford and Jeffress (1964) and (Egan, 1964). vom Hövel's results have since been incorporated into the binaural speech intelligibility model by Beutelmann and Brand (2006) and its revised version (Beutelmann et al., 2010; termed here *BSIM2010*).

BSIM2010 and the other models mentioned earlier require a top-down process that optimizes the equalization parameters of the EC model to achieve optimum unmasking. These parameters are found by analyzing separately the clean speech and noise signals. This has the drawback that the model can only by applied if these clean signals are available (i.e., the model cannot be applied to the mixed speech and noise signals directly) or if assumptions are made about the direction from which one signal originates. To overcome this drawback, we propose here a bottom-up EC front end that is driven only by the mixed input signals. Our EC process can be regarded as *blind*, as it does not require target speech and interfering signals in isolation nor knowledge of their positions.

Our approach has similarities to an earlier blind model by Cosentino et al. (2014) which uses the binaural localization model proposed by Dietz et al. (2011) to estimate the IPD of both the target source and the interfering source. They differentiate between signal and interferer by assuming that the target speech source is located directly in front of the listeners, that is, at 0° in the horizontal plane. Therefore, the other IPD, different to 0°, is associated with the position of the interferer. The estimated IPDs of both target and interferer are then used to calculate the binaural masking level difference (for details, see Culling et al., 2004, 2005). Note that this model assumes that the sound sources can be localized to perform binaural unmasking. Localization certainly plays a role in our everyday communication, but binaural unmasking does not necessarily require a correct localization of target and interferer, because it works best for stimuli with an IPD of π (e.g., Licklider, 1948). Such stimuli have a frequency-dependent ITD which prevents the perception of a clear direction or lateralization.

A similar but more technical approach was presented by Tang et al. (2018), who proposed a blind model for speech intelligibility prediction, which combines a blind source separation algorithm with a nonblind speech intelligibility back end. Like Cosentino et al. (2014), they assumed that the target source is directly in front of two microphones and, additionally, that only one masker source is present. Using these assumptions, the blind source separation algorithm was able to extract estimates of the speech signal and the noise signal from the mixed signals, which were then further analyzed using different speech intelligibility back ends.

A third binaural speech intelligibility model that works blindly was presented by Geravanchizadeh and Fallah (2015). They combined a model of the auditory periphery (Dau et al., 1996) and an EC mechanism to the mixed signals. The back end was a dynamic time warp speech recognizer (Sakoe & Chiba, 1978), which compares the processed mixture of speech and noise with an internally stored reference of the target speech. However, this approach is limited to negative SNRs: Positive SNRs are explicitly excluded as for those normal EC processing would cancel the signal, not the noise. In the real world, SNRs vary over a wide range, and thus any model has to deal with both negative and positive SNRs. This leads to a problem, in that a blind binaural SNR improvement needs to do two opposing things: the *cancellation* of the dominant source if the SNR is *negative* and the *enhancement* of the dominant source if the SNR is *positive*. The first can be done by power minimization of the EC output, the latter by power maximization of the EC output. It is unclear if there is binaural unmasking of speech at positive SNRs for human listeners. At positive SNRs, a properly articulated speech signal is usually fully intelligible for listeners with normal hearing, even when it is presented monaurally and consequently binaural release from masking is not detectable due to ceiling effects. However, listeners with impaired hearing and elderly listeners with cognitive processing deficits show SRTs in the range of 0 dB SNR (e.g., Kollmeier et al., 2016). Moreover, in realistic environments, most SNRs are positive (Smeds et al., 2015), and many hearing aid algorithms require positive SNRs (e.g., Brons et al., 2013; Fredelake et al., 2012). Rennies and Kidd (2018) showed that binaural hearing decreases listening effort at positive SNRs when target and interferer are spatially separated, indicating that binaural unmasking might be relevant at positive SNRs. Therefore, investigating binaural hearing at positive SNRs is ecologically very relevant and critical for understanding binaural listening.

Aim of This Study

Our aim was to develop a blind, signal-driven binaural processing stage that considers both negative and positive SNRs. To do this, we needed to address two questions: (a) Can a blind EC process be realized without assuming a certain direction of the target or interferer? (b) How should positive SNRs be considered?

Blind Model Proposed in This Study

According to Lord Rayleigh's Duplex theory (1907) of sound source localization, the binaural auditory system uses ITDs (and thus IPDs) at frequencies below 1500 Hz to localize sound sources, but above 1500 Hz, ILDs are used. Motivated by this theory, in our model, we apply the EC model of binaural unmasking for frequencies up to 1500 Hz, while the better ear is used at frequencies above 1500 Hz.

The modeling of binaural unmasking below 1500 Hz is achieved in two steps: First, in the equalization step, the left and right ear signals are equalized in level (by amplification and attenuation) and in phase (by delaying the band pass filtered signals). Second, in the cancellation step, two alternative strategies are used. If the SNR is negative, and thus the noise is dominant, the signal from one ear is subtracted from the other which cancels the noise by destructive interference causing a minimization of the model's output level and therefore an improvement of the SNR. If instead the SNR is positive, then the two ear signals are added, which enhances the signal by constructive interference causing a maximization of the model's output level. Including both strategies in the binaural processing stage requires a slight modification of the EC mechanism to allow for constructive interference. The concept of allowing a summation of left and right ear signal was previously proposed by Green (1966). But he considered only situations where subtraction was beneficial. Moreover, the aim of his article was neither to propose a binaural model that can be applied to arbitrary SNRs, nor to propose a method to differentiate addition and subtraction as the better strategy.

To determine whichever of level minimization or level maximization is required, we use a modulation analysis based on the speech-to-reverberation modulation ratio (SRMR; Santos et al., 2014). The output of the SRMR is the ratio between modulation energy in low and high modulation frequency channels. A high value is associated with the presence of speech-like modulation. A low value is associated with the presence of reverberation or noise as both decrease the modulation depth at low modulation frequencies. The SRMR approach is conceptually similar to the classification of signals based on modulation spectra, which was proposed by Ostendorf et al. (1998).

Whichever of the two EC paths yields the higher SRMR value is then used for further processing in the model. We call this *modulation-based selection*. It is done independently for each frequency channel of the EC mechanism below 1500 Hz (in Beutelmann et al., 2009, it was shown that independent EC processing can be assumed across frequency channels). The SRMR is also used for selecting the better ear above 1500 Hz, that is, the ear yielding the higher SRMR value is selected. The blindly selected outputs of the EC processed low-frequency channels (below 1500 Hz) are combined with the blindly selected better ear, high-frequency channels (above 1500 Hz).

In this study, we focus on the binaural processing stage of the model and do not modify the SII back end from BSIM2010, allowing for maximum comparability to that. In addition, the comparability to other nonblind back ends is basically conserved: As the whole model processing is linear (apart from the nonlinear control strategy of selecting the EC parameters and level minimization vs. maximization), it is in principle possible to process target speech and interfering signals separately (see later). We call our new model *BSIM2020*.

Note that the blind model front end is independent of the model back end, as the speech intelligibility measure is not required for optimizing the EC parameters. The front end can be combined with arbitrary back ends, for example, with intrusive back ends like SII (ANSI S3.5-1997, 1997), or with blind back ends, which use principles of automatic speech recognition (e.g., Schädler et al., 2016; Spille et al., 2018).

Organization of This Article

In Experiment I, we evaluated whether the new BSIM2020 model predicts SRT data in stationary noise, located at different azimuths in the horizontal plane, with the same accuracy as the earlier, nonblind BSIM2010 model by Beutelmann et al. (2010).

In Experiment II, we collected new data to test whether the proposed maximization strategy is actually required to model human performance in speech-in-noise experiments at positive and negative SNRs or whether simply switching off EC processing at positive SNRs and using the ear with the better SNR was sufficient. To provoke binaural release from masking at positive SNRs, we low-pass filtered speech and noise (which can be regarded as a simple model of high-frequency hearing loss) and time compressed them (which can be regarded as a simple model of reduced central processing speed). In the model analyses of both experiments, we compared the new modulation-based selection between level minimization and maximization with using either level minimization only or level maximization only to determine if the new selection process was indeed necessary.

Methods

Front-end Binaural Processing

The BISM2020 model is schematically shown in Figure 1. It uses mixed speech and noise as input signals to determine the EC parameters.

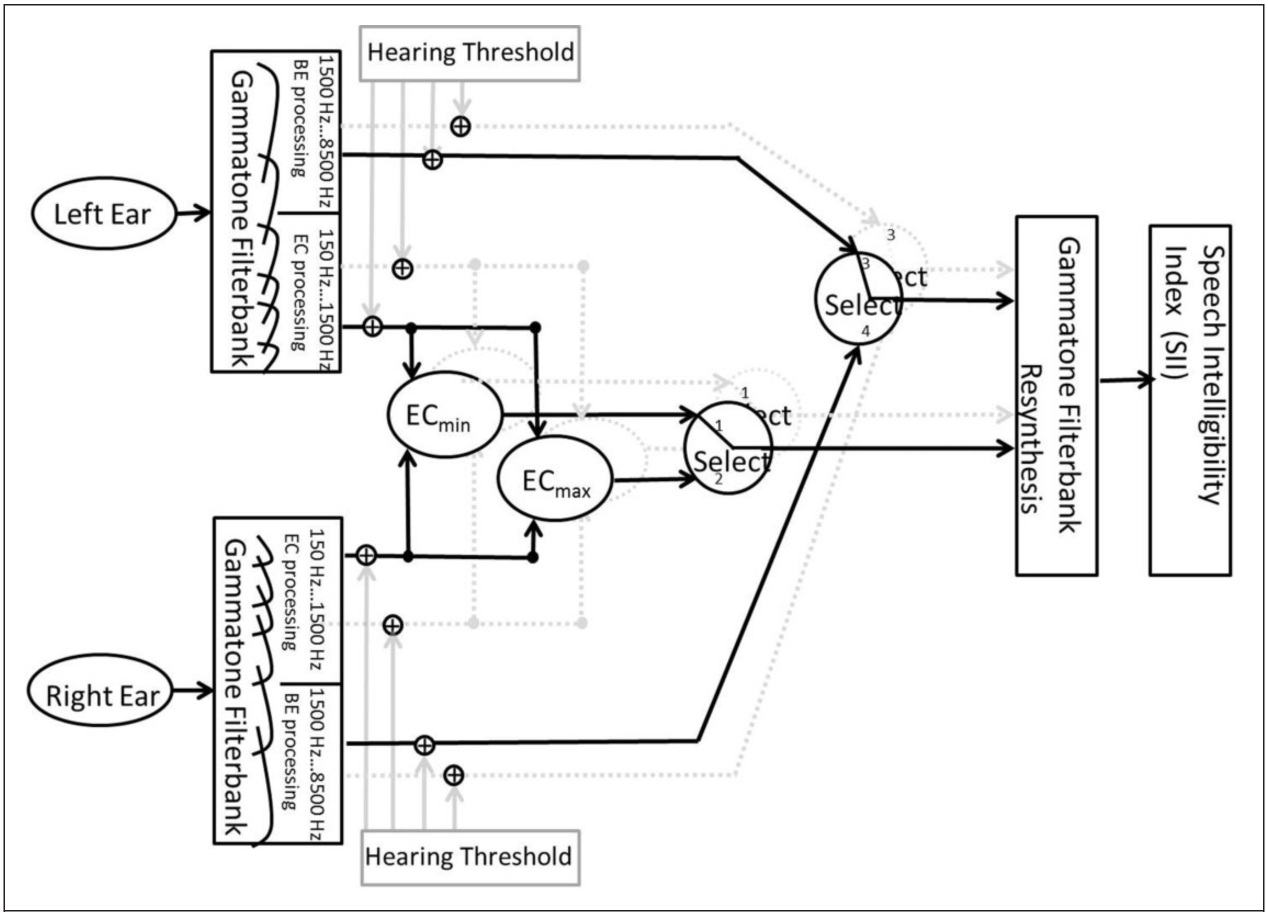
Hearing Threshold
Left Ear
Gammatone Filterbank
1500 Hz...8500 Hz BE processing
150 Hz...1500 Hz EC processing
EC$_{min}$
EC$_{max}$
Select
1
2
Select
3
4
Gammatone Filterbank Resynthesis
Speech Intelligibility Index (SII)
Right Ear
Gammatone Filterbank
150 Hz...1500 Hz EC processing
1500 Hz...8500 Hz BE processing
Hearing Threshold

Figure 1. Block Diagram of the General Processing Performed in BSIM 2020. The mixed signals on the left and right ear are divided into 30 frequency bands ranging from 150 Hz to 8500 Hz using a gammatone filter bank (Hohmann, 2002). Afterward, frequency bands below 1500 Hz are fed to the EC stage, where both a level minimization and level maximization is performed in parallel, denoted as EC_min and EC_max. The speech-to-reverberation modulation ratio (SRMR; Santos et al., 2014), denoted as select-stage is used (a) for selecting if the level-minimization or the level-maximization produces the best SNR improvement and (b) for determining the better ear. This is indicated by the numbers in the selection stage. For low frequencies, either the EC-Min (1) or the EC-Max (2) path is selected. For high frequencies (above 1500 Hz), either the left ear channel (3) or the right ear channel is selected. Both, binaurally processed channels and the better ear channels are combined, and a single-channel output is resynthesized using a gammatone synthesis filter bank. The output can then be analyzed by an arbitrary back end, which is the SII in this study.
EC = equalization cancellation; SII = speech intelligibility index.

First, the mixed signal is divided into 30 ERB (Moore & Glasberg, 1983) spaced frequency bands ranging from 150 Hz to 8500 Hz using a gammatone filter bank (Hohmann, 2002), simulating the frequency selectivity of the auditory system. Next, EC processing is then performed in frequency channels up to 1500 Hz (i.e., in the lower 15 frequency channels). The ILD is estimated in each frequency band by calculating the power in each frequency band, and then the power of the right ear channel is subtracted from the left ear channel leading to a signed ILD, where negative values correspond to a higher level at the right ear and positive values to a higher level at the left ear. The signals in the left ear and right ear channel are then amplified or attenuated such that the levels are equalized between the left and right ear channel. We assume that the equalization is imperfect, which we implemented via a jitter in the interaural level equalization process that prevents perfect level equalization between the left and right ear channel. The jitter was a normally distributed random variable (vom Hövel, 1984; see supplementary material for more details). The jitter is applied to the signals directly, and thus, a Monte-Carlo simulation is required to model the statistics of the uncertainties. This procedure is similar to the method used by Beutelmann and Brand (2006) and Wan et al. (2010), but different to the method used by Beutelmann et al. (2010), where the uncertainties were incorporated analytically with respect to their expectation values and variances. Afterward, the ITD is estimated in each frequency channel from the phase information of the cross-power spectral density between left and right ear channel. After the estimation, the ITD is compensated for. This equalization process in time again includes binaural processing inaccuracies, which are assumed to be independent realizations of another normally distributed random variable (vom Hövel, 1984).

The cancellation step is performed by either subtracting the left ear channel from the right ear channel (as the left and right ear signals are equalized, the subtraction operation is symmetric and can be performed either way) or adding, depending on the SRMR (see later). More details about the processing in BSIM2020 can be found in the supplementary material.

Selection of EC Path and Better Ear Based on Modulation Analysis

In the next step, the better of the EC processing strategies and the better ear are selected blindly to produce a binaurally processed mono signal, which can be analyzed by an arbitrary speech intelligibility back end. To do this, we use the SRMR measure (Santos et al., 2014) applied independently in each of the 30 frequency bands. The envelope of the EC processed signals and the envelope of the two ear signals are extracted by taking the absolute value of the analytical signal (i.e., the Hilbert envelope).

These are then analyzed by a modulation filter bank with eight logarithmically spaced filters ranging from 4 Hz to 128 Hz. In each modulation filter, the power (i.e., energy per block) is computed by taking the squared magnitude of the Fourier transformed envelope (in this study, each sentence was treated as one block). The SRMR is the ratio of the power in the four lowest modulation filters to the power of the four highest modulation filters. For frequency channels up to 1500 Hz, the SRMR is calculated on the outputs of both the minimizing and the maximizing EC paths. The path yielding the higher SRMR value is kept for further processing.

Above 1500 Hz, however, the SRMR measure is calculated on the left and right ear signals. If the left-right difference in SRMR exceeds a value of 0.1, then the ear providing the higher value is selected as better ear channel, but otherwise, the ear providing the lower root mean square is selected.

The selected EC channels are then combined with the selected better ear channels using a gammatone synthesis filter bank to produce a single signal, which is then analyzed by a speech intelligibility back end or listened to by a human listener.

Model Testing and SRT Calculation

The model was tested using sentences from the Oldenburg Sentence test corpus (OlSa; Wagener et al., 1999c). OlSa sentences consist of five-word sentences with a fixed grammatical structure *noun-verb-numeral-adjective-object*, where each word is randomly selected from a list of 10 words. Ten sentences were used in the simulations such that every word of the OlSa corpus appeared once.

The speech material was mixed with the noise at 41 different SNRs ranging from –20 dB to +20 dB in steps of 1 dB. For each tested SNR, 30 random sets of the jitter random variables were used in each frequency channel. In total, 12,300 simulations were conducted for each tested condition in the two experiments. We obtained SRTs via intermediate calculation of SII values. To obtain these, the speech and noise signals were processed separately using identical EC parameters and random variables as for the mixed signals.

The SII values were averaged across Monte-Carlo simulations for each of the 10 sentences used in the simulations. The SII for each of the 10 sentences leading to the SRT of –7.8 dB (the average SRT across listeners with normal hearing) obtained for the colocated condition was averaged across all sentences. The result served as reference SII value. Next, the mean SII (across the 10 sentences and 30 Monte-Carlo simulations) was calculated for each of the 41 tested SNRs, and then whichever SNR yielded an average SII closest to the reference SII was selected for the different azimuth positions of the noise. This SNR is taken as the estimate of the SRT_{50}.

We should clarify that the model is binaurally blind even though speech and noise are run separately through for the SII calculation. The reason is that the model uses the mixed signals to determine the EC parameters; hence, it is binaurally blind. But the SII calculation requires separate signals, so we used the same EC parameters for speech and noise.

Experiment I—Modeling Speech Intelligibility in Spatially Separated Stationary Noise

To compare the blind estimation process of our new BSIM2020 with the SNR optimization procedure of BSIM2010, we simulated the speech intelligibility experiments conducted by Beutelmann and Brand (2006) with both models.

In those experiments, 10 listeners with normal hearing (21–43 years; audiometric thresholds of 20 dB HL or better between 250 and 8000 Hz) participated. Speech intelligibility experiments were conducted using the OlSa sentences in noise (Wagener et al., 1999a, 1999b, 1999c). An adaptive procedure (Equation 9, Brand & Kollmeier, 2002) was used to determine the SNR at which 50% of the sentences were understood correctly. All measurements were conducted using the Oldenburg Measurement Applications (HörTech gGmbH, Oldenburg, Germany). SRTs were obtained in three acoustical environments (which were denoted as "anechoic," "office," and "cafeteria") and for different directions of the noise source, while the target speech was always presented from an azimuth of 0° (directly in front of a listener). The tested noise directions were –140°, –100°, –45°, 0°, 45°, 80°, 125°, and 180° in the anechoic and office conditions and –135°, –90°, –45°, 0°, 45°, 90°, 135°, and 180° in the cafeteria condition. In the anechoic condition, speech and noise signals were convolved with head-related transfer functions taken from Algazi et al. (2001). For the office and cafeteria conditions, Beutelmann and Brand (2006) used their own recordings of head-related transfer functions. All stimuli were presented binaurally using HD200 headphones (Sennheiser, Wedemark, Germany), which were free-field equalized using an finite impulse response (FIR) filter with 801 coefficients. The SRT was determined using test lists of 20 sentences. The test lists were randomly selected out of 45 lists. The noise level was set to 65 dB SPL, and the speech level was varied adaptively to find the individual SRT.

Results

Figure 2 shows the predicted SIIs from our model for a selection of four of the noise directions (–100°, 0°, 45°, and 180° azimuth) and for the tested SNRs ranging from –20 dB to +20 dB. These four angles were selected because they show the general characteristics representative of all of the simulations. The top-right panel shows the SII curves for colocated speech and noise sources (i.e., S_0N_0). The different EC processing strategies result in identical SII curves. This result was expected, because there are no interaural differences for the binaural processor to make any use of to enhance the SNR. The same holds for the noise located at 180° azimuth, shown in the bottom right panel. The curves are slightly shifted toward negative SNRs, which is an effect of pinna cues as they slightly improve the SNR for the S_0N_{180} condition compared with the S_0N_0 condition.

In both of the left panels, the effects of the different processing strategies for spatially separated target and masker can be observed. The masker at –100° azimuth provides a larger release from masking than the masker located at 45° azimuth. At negative SNRs, the level minimization provides the best SII, whereas at positive SNRs, the level maximization provides the best SII. The SII of modulation-based blind selection between minimization and maximization converges toward the SII of the level minimization at negative SNRs and to the SII of the level maximization at positive SNRs. For SNRs close to 0 dB, the modulation-based selection provides higher SII values than the level minimization or level maximization alone. This is probably due to the fact that the SRMR measure selects the optimal strategy independently in each frequency channel leading to a synergistic effect.

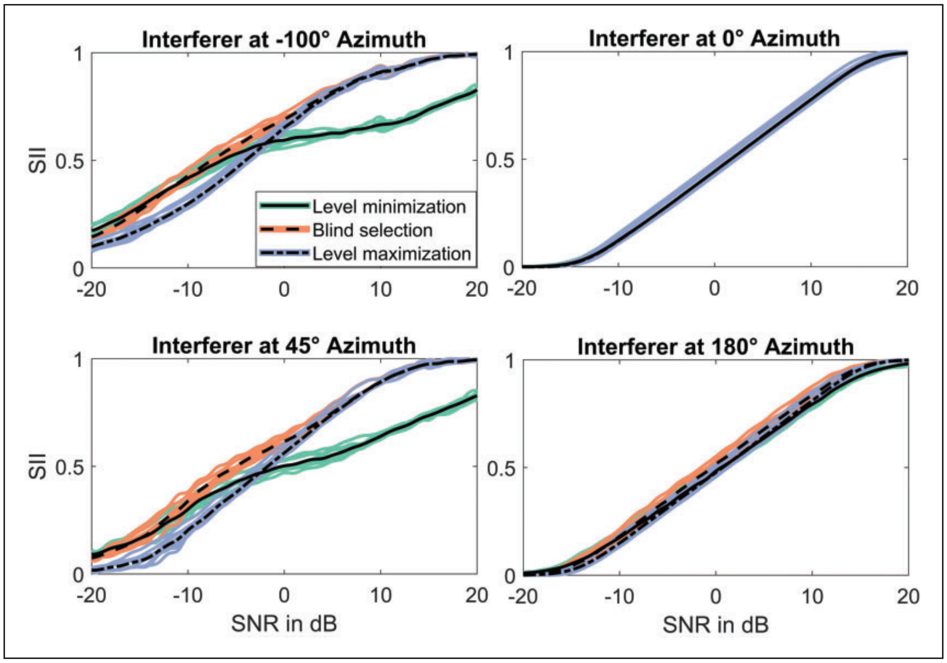

Figure 2. Exemplary SII Curves for Speech (Located at 0°) in Noise (Various Locations). The SII is shown for the three tested processing schemes in the EC mechanism, which is either a level minimization (solid green) at the output, a level maximization (dotted dashed blue line) at the output, or a modulation-based selection of level minimization and maximization (dashed red line).
SII = speech intelligibility index; SNR = signal-to-noise ratio.

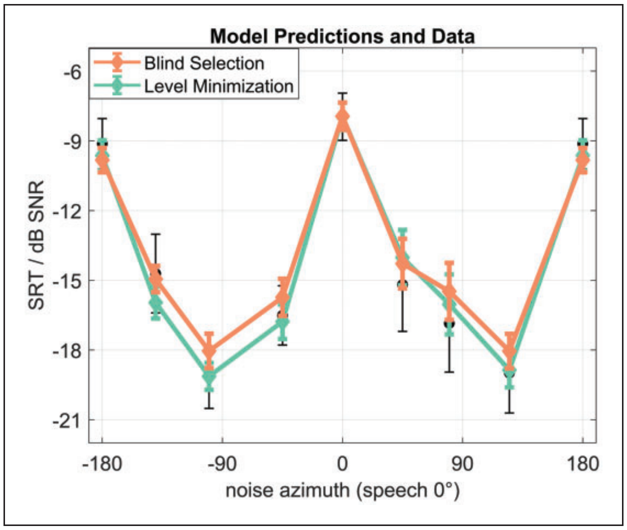

Figure 3. Data (Black Dots) and Predictions Obtained for the Anechoic Situation Using the Level Minimization as EC Processing Criterion (Green Diamonds) and Using the Modulation-Based Selection of Level Minimization and Level Maximization (Red Diamonds). Error bars of the obtained data indicate the interindividual standard deviation; error bars of the predictions show the standard deviation across sentences.
SNR = signal-to-noise ratio; SRT = speech recognition threshold.

Figure 3 shows the SRTs predicted using either the level minimizing EC mechanism on its own or the modulation-based blind selection of either the minimizing or maximizing EC processing strategy (results obtained with the maximization strategy and listening only monaurally are reported in the supplementary material). Predictions are shown along with the data by Beutelmann and Brand (2006). The predictions were essentially equally accurate. Using the level minimization, very accurate predictions in terms of the coefficient of determination $R^2 = .97$, the root mean square error (RMSE) between predicted and measured SRTs (RMSE = 0.7 dB), and the bias (–0.02 dB) were obtained. The results obtained with modulation-based selection are comparable to these: Predicted SRTs were slightly higher, but the RMSE and bias, as well as the coefficient of determination are not affected ($R^2 = .985$, RMSE = 0.9 dB, bias = –0.5 dB).

In summary, the analysis of this experiment shows that our new model, using modulation-based blind selection of the optimal binaural processing strategy (either minimizing or maximizing of the output level) and the blind selection of the better ear, is able to describe SRTs for spatially separated speech in noise at negative SNRs.

Experiment II—Binaural Speech Intelligibility at Positive SNRs

To test the model at positive SNRs, we collected new data to investigate binaural intelligibility level differences and binaural release from masking at positive SNRs. We used two sets of speech material: the OlSa speech material as in Experiment I and the Göttingen sentence test's material (GoeSa; Kollmeier & Wesselkamp, 1997), which are everyday sentences. Low-pass filtering and time compression were applied to degrade speech intelligibility and shift SRTs to positive SNRs. In Schlueter et al. (2015), it was shown that the SRT_{50} of listeners with normal hearing for OlSa sentences can be shifted to an SNR of 3 dB if the speech material is time compressed to 25% of its original length. However, in this study, such extreme compression was avoided. Instead, low-pass filtering was applied in addition to time compression. Low-pass filtering additionally lowers speech intelligibility while preserving the usable binaural ITD cues, which is necessary for achieving binaural unmasking. The binaural configuration used depended on the sentences. For the OlSa sentences, binaural release from masking was induced by imposing either an IPD of π or an ITD of 750 μs on the noise. For the GoeSa sentences, however, pilot experiments revealed that SRTs could not be determined reliably if they were compressed to 50% of their original length and low-pass filtered at 1200 Hz. Thus, they were only compressed to 66% of their original length and low-pass filtered at 1500 Hz. Moreover, only an IPD of π of the noise was tested. The IPD condition was chosen because it ought to give the maximal binaural release from masking, whereas the ITD condition was a more realistic scenario because the stimulus can be associated to a certain direction.

Listeners

A total of 13 listeners with normal hearing (6 male, 9 female, 19–28 years, mean age: 23 years) participated in the experiment. They had no previous experience with sentence test procedures, and audiometric thresholds did not exceed 20 dB HL.

Stimuli

The speech materials from the Oldenburg sentence test and the Goettingen sentence test in noise were used. The OlSa sentences have a fixed syntactical structure (e.g., "Peter sieht vier nasse Tassen."—"Peter sees four wet cups."), whereas the GoeSa sentences consist of fixed meaningful sentences with a variable syntactical structure and a variable length. These provide more semantical context information and can therefore be considered to be more similar to the sentences used in everyday conversation (e.g., "Ein kleiner Junge war der Sieger."—"A little boy was the winner." or "Jetzt wird das Fundament gelegt."—"Now the foundation is laid."). They were manipulated in three ways: (a) unprocessed, that is, standard OlSa or GoeSa measurements, (b) time compression to 66% of their original length and low-pass filtering with a cut-off frequency of 1500 Hz termed “LP 1500 Hz, TC 0.66,” and (c) time compression to 50% and low-pass filtering to 1200 Hz termed “LP 1200 Hz, TC 0.5,” which was only applied to the OlSa sentences.

Both time compression and low-pass filtering were performed using the PRAAT software (Boersma & van Heuven, 2001; Boersma & Weenink, 2018). In PRAAT, the low-pass filters are realized as a one-tailed Hann window with adjustable cut-off frequency (–6 dB) and filter slope (the width between pass and stop band) of 100 Hz, meaning that full attenuation was achieved within 100 Hz. Time compression was done using the pitch-synchronous overlap add algorithm (Moulines & Charpentier, 1990), which allows for time compression without change in pitch. This method was evaluated in Schlueter et al. (2015) and was shown to be a valid method to increase SRTs based on time compressing the speech material.

SRTs were determined as the SNR for 80% speech intelligibility and found by changing the level of the speech adaptively and by using 0.8 as the target value (Equation 9, Brand & Kollmeier, 2002). Five sentence lists with 20 sentences each were provided for training, and then the main binaural conditions with IPD of π or ITD of 750 µs (see earlier) were run.

Results

Figure 4 shows the SRT results obtained for OlSa sentences in the three tested conditions: unprocessed (left panel), time compression to 66% of the original length and low-pass filtering at 1500 Hz (LP 1500 Hz, TC 0.66; middle panel), and time compression to 50% of the original length and low-pass filtering at 1200 Hz (LP 1200 Hz, TC 0.50; right panel). The median SRT_{80} was found to be at –5.2 dB SNR in the S_0N_0 condition for the unprocessed OlSa sentences. By applying time compression and low-pass filtering to the OlSa sentences, the median SRTs in the S_0N_0 condition were shifted to –0.4 dB SNR for the “LP 1500 Hz, TC 0.66” condition and +4.6 dB SNR for the “LP 1200 Hz, TC 0.50” condition. In the S_0N_π condition, the median SRT was found to be 6 dB below the SRT in the N_0S_0 condition, which was independent of the manipulation applied to the stimuli. In the S_0N_{750} condition, the median SRT was always 1 dB higher (worse) than the median SRT in the $N_\pi S_0$ condition. However, in both of the binaural conditions, the variation across

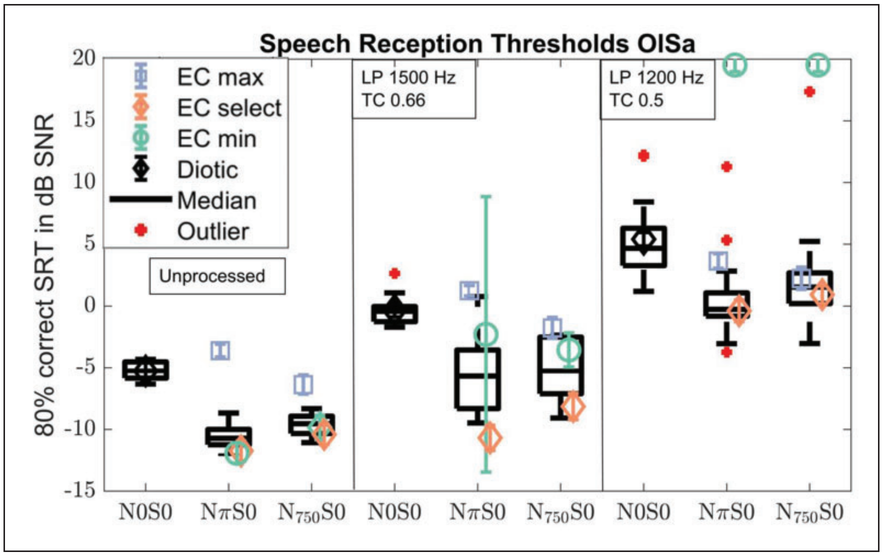

Figure 4. Boxplots (Median: 25%–75% Confidence Interval [Box], 9%–91% Confidence Interval [Whisker] in Black, and Outliers in Red) of the SRT_{80} Obtained for 13 Listeners With Normal Hearing Using the OlSa. Unprocessed denotes the original stimuli, LP denotes the cut-off frequency of the low-pass filter, and TC denotes the applied time compression. N_0S_0 denotes the diotic (same signal at both ears) presentation of speech and noise, $N_\pi S_0$ denotes that the noise was interaurally phase inverted, and $N_{750}S0$ denotes that the noise was interaurally delayed by 750 μs. Predictions obtained with the three EC outputs are shown, where blue squares show the predicted SRT for the level maximization, green circles are the predicted SRTs using the level minimization, and red diamonds denote the results obtained by combining level minimization and maximization based on modulation analysis. Black diamonds correspond to the diotic, that is, N0S0, model outcome.
OlSa = Oldenburg Sentence test corpus; EC = equalization cancellation; SNR = signal-to-noise ratio; SRT = speech recognition threshold.

listeners was increased for the time-compressed and low-pass filtered stimuli.

Figure 4 also shows the predictions from the BSIM2020 model using the three processing strategies: level minimization only, level maximization only, and modulation-based selection between level minimization and maximization. In the “unprocessed” condition, both the level minimization and the modulation-based strategies resulted in the same predicted SRT, which slightly overestimated the mean binaural release from masking and coincided with the best SRT obtained by the human listeners, which was found to be at –12 dB SNR. The level-maximization strategy was not adequate to describe the results obtained in the listening experiment because the predicted SRT was even higher than in the N_0S_0 condition.

In the “LP1500 Hz, TC 0.66” condition, the SRT_{80} was found to be at –0.4 dB SNR for the S_0N_0 condition. In the binaural conditions, the SRTs were 5–6 dB lower. The predicted SRT using the level-minimization strategy resulted in a predicted SRT which coincided with the upper quartile of the obtained data. However, the standard deviation was very large in the S_0N_π condition. The predicted SRT using the level-maximization strategy was worse than in the diotic condition. The predicted SRT obtained using the output of the modulation-based selection strategy was again in the range of the best human listener.

In the “LP1200, Hz TC 0.5” condition, the diotic SRT was found to be at +4.6 dB SNR. The SRTs in both of the binaural conditions were found to be at 1.5 dB SNR for the S_0N_{750} condition and 0.26 dB for the S_0N_π condition. In the binaural conditions, the level-minimization strategy failed to predict the obtained SRTs. The predicted SRT using the level-maximization strategy was obtained at the upper end of the figure, while the modulation-based selection resulted in predicted SRTs close to the median. The model performance was evaluated by calculating the RMSE between the predicted and measured SRT. For the modulation-based selection strategy, it was 2.5 dB.

For statistical evaluation, first the Kolmogorov–Smirnov test ($\alpha = .05$) was conducted to test if the data was normally distributed. It revealed that a normal distribution can only be assumed for the results obtained in the S_0N_0 condition (LP = 1500 Hz, TC = 0.66) and $N_\pi S_0$ condition (LP = 1200 Hz, TC = 0.5). Therefore, a Wilcoxon signed-rank test ($\alpha = .01$) was conducted for the evaluation of the SRT differences across conditions. A statistically significant effect of inverting the phase of the noise on SRTs was found for all conditions—$p(\text{Unprocessed}) = 2.4 \times 10^{-4}$, $p(LP_{1.500\,Hz}) = 4.8 \times 10^{-4}$, $p(LP_{1200Hz}) = 4.8 \times 10^{-4}$. The same was shown for the S_0N_{750} condition—$p(\text{Unprocessed}) = 2.4 \times 10^{-4}$, $p(LP_{1.500\,Hz}) = 2.4 \times 10^{-4}$—but not for the $LP_{1.200\,Hz}$ condition—$p\ (LP_{1.200\,Hz}) = 0.0105$.

Figure 5 shows the corresponding binaural intelligibility differences (BILDs), which are the differences in SRTs between the S_0N_0 conditions and S_0N_π conditions. In general, the median BILD was comparable across manipulations at about 2–7 dB. However, the variance across listeners was increased for the conditions with time compression and low-pass filtering. A Wilcoxon signed-rank test found no statistical difference between the BILDs obtained for the unprocessed stimuli and the time compressed and low-pass filtered stimuli [$p(LP_{1.500\,Hz}) = 0.68$, $p(LP_{1.200\,Hz}) = 0.58$].

Figure 6 shows the BILDs for the S_0N_{750} condition. Again, these were not significantly different for the low-pass filtered and time-compressed conditions compared with the unprocessed condition [$p(LP_{1.500\,Hz}) = 0.41$, $p(LP_{1.200\,Hz}) = 0.24$)].

Figure 7 shows the results obtained with the GoeSa sentences. Compared with the SRTs obtained with the OlSa sentences, SRTs with the GoeSa sentences were shifted to even more positive SNRs. The standard deviation was also larger. For the unprocessed GoeSa sentences, the median SRT in the S_0N_0 condition was found to be at –4.6 dB SNR. The median SRT in the S_0N_π conditions was found to be at –7.7 dB SNR, which was significantly lower (Wilcoxon signed-rank: p value = 2.44×10^{-4}). This corresponded to a median BILD (median of the difference) of 2.6 dB. For the time-compressed and low-pass filtered condition, the median SRTs were found to be at 9.9 dB SNR (S_0N_0)

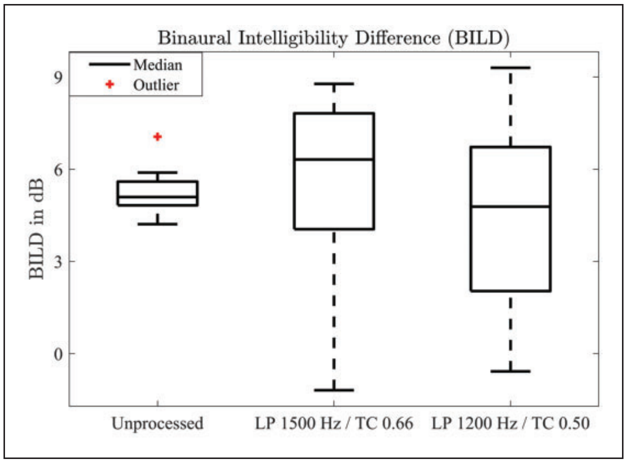

Figure 5. Boxplots (Median [Horizontal Line], 25%–75% Confidence Interval [Box], 9%–91% Confidence Interval [Whisker], and Outliers [Red Crosses]) of BILDs Obtained for 13 Listeners With Normal Hearing Using OlSa Sentences. The noise had an IPD of π. Unprocessed denotes the unmanipulated OlSa sentences, and LP and TC denote the low-pass filter and time compression applied to the OlSa material.
BILD = binaural intelligibility difference.

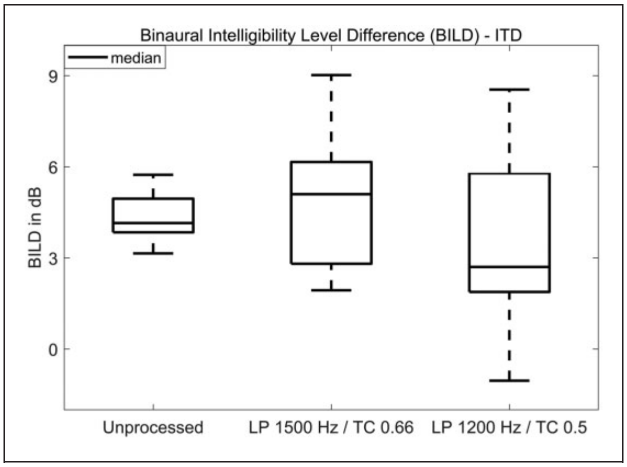

Figure 6. Boxplots (Median [Horizontal Line], 25%–75% Confidence Interval [Box], 9%–91% Confidence Interval [Whisker], and Outliers [Red Crosses]) of BILDs Obtained for 13 Listeners With Normal Hearing Using OlSa Sentences. The noise had an ITD of 750 μs. Unprocessed denotes the unmanipulated OlSa sentences, and LP and TC denote the low-pass filter and time compression applied to the OlSa material.
BILD = binaural intelligibility differences; ITD = interaural time difference.

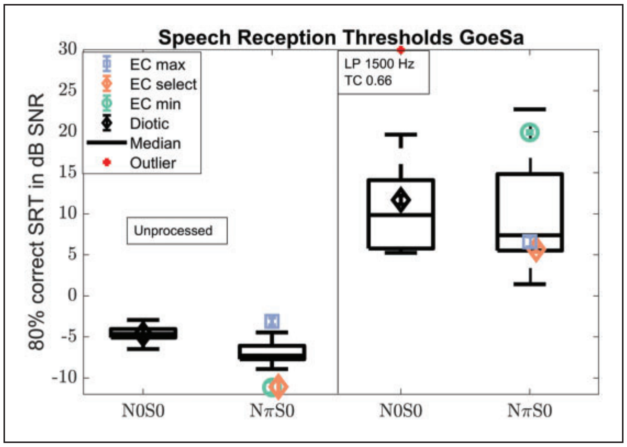

Figure 7. Boxplots (Median [Horizontal Line], 25%–75% Confidence Interval [Box], 9%–91% Confidence Interval [Whisker], and Outliers [Red Crosses]) of SRT80 Obtained for 13 Listeners With Normal Hearing Using the GoeSa. Unprocessed denotes the original stimuli, LP denotes the cut-off frequency of the low-pass filter, and TC denotes the applied time compression. Predictions obtained with the three EC outputs are shown, where blue squares show the predicted SRT for the level maximization, green circles are the predicted SRTs using the level minimization, and red diamonds denote the results obtained by combining level minimization and maximization based on modulation analysis.
GoeSa = Göttingen sentence test's material; EC = equalization cancellation; SNR = signal-to-noise ratio; SRT = speech recognition threshold.

and 7.4 dB SNR (S_0N_π). A median BILD of 3.3 dB was obtained. However, the reduced SRT in the S_0N_πcondition was not statistically different to the SRT in the S_0N_0 condition due to the large variability of SRTs across listeners (Wilcoxon signed-rank: p value = .17).

The BSIM2020 model predicted a binaural release from masking for both unprocessed and processed GoeSa sentences, respectively. The predicted SRTs using the level maximization and the modulation-based selection were close to the median SRT in the binaural condition. However, due to the very large variability across listeners, a direct comparison between measured and predicted SRTs was problematic.

In summary, Experiment II showed that binaural release from masking can be found at negative and positive SNRs. Therefore, it is also necessary to consider positive SNRs in models of binaural speech intelligibility. The SRTs predicted with the modulation-based selection strategy of our BSIM 2020 model showed good agreement with the measured SRTs at positive and negative SNRs.

Discussion

Blind Modeling of Binaural Processing

This study introduces a new *blind* model of binaural speech intelligibility, termed BSIM2020, which requires only mixed signals as input. Below 1500 Hz, the binaural unmasking is performed in two parallel paths, which either minimize or maximize the level at the output of the EC process. The interferer is attenuated at negative SNRs using destructive interferences, and the target is enhanced at positive SNRs using constructive interferences. Above 1500 Hz, the monaural left ear channel and the monaural right ear channel are considered. The decision as to which of the two EC paths is used below 1500 Hz and which of the two ear channels is used above 1500 Hz for further processing is done independently for each auditory frequency channel based on a modulation analysis based on the SRMR measure (Santos et al., 2014), which selects whichever EC path and ear channel provides most speech-like modulations. This processing has the advantage that no separate speech and noise signals are required.

The concept of minimizing and maximizing the level at the EC output is different to binaural speech intelligibility models from the literature, where the back end (speech intelligibility metric) is typically used to optimize the EC parameters in a top-down process (Andersen et al., 2016; Beutelmann et al., 2010). In such a top-down process, the EC parameters are adjusted to maximize the speech intelligibility metric. Compared with the models proposed by Cosentino et al. (2014), Tang et al. (2018), and Geravanchizadeh and Fallah (2015), our model has the advantage that it does not make assumptions about the SNR, the location of a target speaker or interfering noise source, or the number of interfering noise sources. Moreover, the model is designed in such a way that it can be combined with arbitrary speech intelligibility back ends, because it produces a binaurally processed mono signal. Therefore, it can also be used as simple binaural beamformer for signal enhancement. Geravanchizadeh and Fallah (2015) proposed a binaural speech intelligibility model with a blind EC processing stage. However, positive SNRs where explicitly excluded, because they assumed 100% speech intelligibility at positive SNRs and, consequently, that binaural processing is only relevant at negative SNRs.

The concept presented here can be regarded as a bottom-up process, where the binaurally processed signals are fed to a modulation-based selection stage, which serves as a simple "gate" that passes the channel with the best representation of speech to the back end. Moreover, the relatively simple binaural processing presented here does not require any assumption on the localization of the target and/or interfering source. Binaural unmasking is not necessarily based on localization, as it works best for stimuli which are interaurally phase inverted (e.g., Levitt & Rabiner, 1967) and, consequently, have frequency-dependent ITDs which are ambiguous with respect to location or lateralization. Nevertheless, binaural unmasking and binaural sound source localization are certainly related, for example, with respect to object formation and stream segregation (e.g., Bronkhorst, 2000). The inclusion of localization cues into the present model will be subject of future studies.

In our study, we evaluated the blind binaural processing using the SII to be able to compare it with an earlier model, BSIM2010 (Beutelmann et al., 2010), which used an SNR optimization for its EC processing, that is, a level-minimization strategy at negative SNRs. Nevertheless, the novel binaural processing stage can be combined with arbitrary speech intelligibility back ends, which may be either based on mixed signals (e.g., Andersen et al., 2017; Schädler et al., 2016; Spille et al., 2018) or on separate speech and noise signals as input (e.g., ANSI S3.5-1997, 1997; Steeneken & Houtgast, 1980; Taal et al., 2011). The latter can be achieved (like it is done in this study) by additionally processing speech and noise in isolation but using blindly estimated binaural parameters and processing strategies.

The results of Experiment I demonstrate that the level-minimization strategy of the EC output is sufficient to describe the data obtained in the binaural listening experiment by Beutelmann and Brand (2006). The predictions of the blind selection in the BSIM2020 model were comparable to predictions obtained with the earlier BSIM2010 model.

The bias and RMSE were only slightly increased, which was caused by an interaction between the modulation analysis of the EC processed output and the binaural processing inaccuracies used in the Monte-Carlo simulation. To explain, the binaural processing inaccuracies limit the SNR improvement of the EC process to mirror human performance. In some cases, where the random variables for the jitter are drawn from the tails of the normal distribution, the difference in the SRMR measure between maximized and minimized EC output is very small, because both paths are dominated by the noise. If this is the case, the selection of the theoretically better EC channel is uncertain. If the binaural processing inaccuracies are disabled, the modulation-based selection of either the minimized or maximized output works more robustly. However, as we assume that processing errors in lower binaural processing stages also affect all following processing stages including the assumed selection of the best channel, only results using binaural processing inaccuracies are presented in this study.

In further work, we also attempted to use SRMR as back end for directly predicting SRT based on the mixed speech and noise signals. However, this approach failed, because the increase of the SRMR with increasing SNR was too shallow at negative SNRs to derive SRTs for 50% speech intelligibility. Figure 8 shows the SRMR output for speech in noise based on the signal at the output of the blind binaural processing stage, where the noise is either located at 0° or 125° in the horizontal plane. The SRMR is able to select the better EC processing path and the better ear. However, the binaural benefit is overestimated as the difference in SRMR between 0° and 125° is too large. Consequently, no SRT criterion can be chosen that holds for both conditions, because both curves do not cross. Cosentino et al. (2014) did not run into this problem, because they used the SRMR only to determine the better ear by applying it to the left and right ear signal. The higher SRMR value of both ears was then converted to a dB value using a fitting function, which was derived in their study, assuming that listening with the better ear produces a benefit in the range from 0 to 6 dB. We do not use their mapping function of SRMR to SNR improvement, because our model produces a binaurally processed mono signal.

Binaural Release From Masking at Positive SNRs

In Experiment II, binaural speech intelligibility and binaural unmasking was investigated at positive SNRs. To this end, time compression and low-pass filtering of the stimuli was used to increase the SRT_{80}. For the

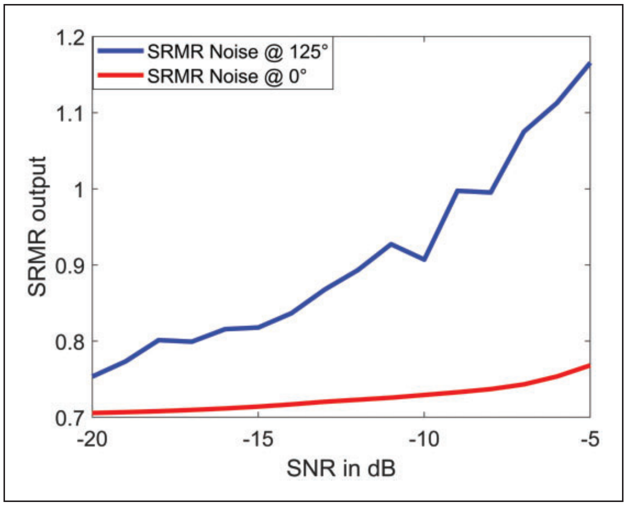

Figure 8. SRMR of Output of EC Stage as a Function of SNR for Speech in Noise. The noise was either located at 0° in the horizontal plane (red line) or at 125° in the horizontal plane (blue line). SNR = signal-to-noise ratio; SRMR = speech-to-reverberation modulation ratio.

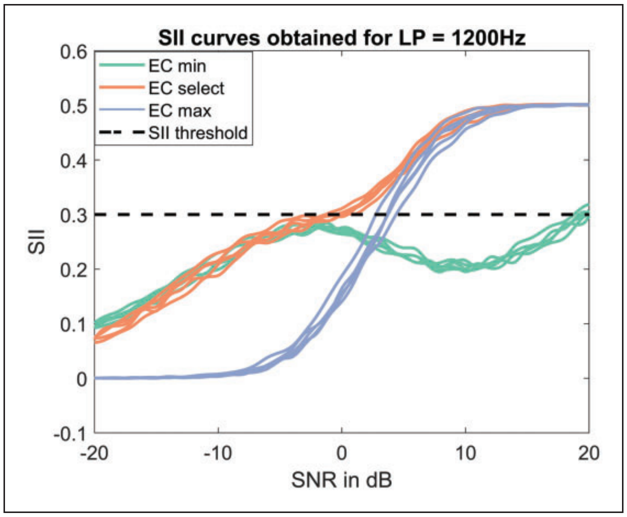

Figure 9. SII Curves Obtained for Low-Pass Filtered and Time-Compressed OlSa Sentences. The green curves denotes the SII obtained for the output of the level minimization, blue curves the output of the level maximization, and red the output of the non-intrusively selected output. The black-dotted line indicates the SII criterion for SRT_{80}.
SII = speech intelligibility index; EC = equalization cancellation; SNR = signal-to-noise ratio.

OlSa sentences, the release from masking did not differ significantly across all tested scenarios, demonstrating that the binaural auditory system produces release from masking even at positive SNRs. The BSIM2020 model is able to predict this by using the modulation-based blind selection of level minimization or level maximization at frequencies below 1500 Hz and selecting the better ear above 1500 Hz. In the "LP 1500 Hz, TC 0.66" condition, the standard deviation of the predicted SRT using level minimization was very large. In the "LP 1200 Hz, TC 0.5" condition, the level-minimization strategy failed to predict an SRT, because the target is cancelled from the mixed signals at positive SNRs. This was caused by the flattening and nonmonotonic character of the SII curves at 0 dB SNR, which was also observed in the SII curves based on the level minimization obtained for Experiment I (see Figure 2). However, the effect was larger in Experiment II, because there was no better ear in the NπS0 stimulus. Moreover, the low-pass filtering enhanced the effect of the EC processing on the obtained SII value, because only the low-frequency region was considered, where binaural processing (and EC processing) can be assumed to take place. This can also be seen in Figure 9, where the SII curves obtained for the low-pass filtered stimuli (cut-off frequency 1200 Hz) are shown. The SII based on the output of the level minimization is a nonmonotonic function, causing the large variance in the predicted SRTs with LP = 1500 Hz and TC = 0.66. The blindly selected output shows the synergistic effects of combining level maximization and level minimization for SNRs close to 0 dB, providing a better agreement between measured and predicted data. However, the RMSE between predictions using the blind binaural processing stage and measured SRTs is in the range of 2–5 dB for the OlSa sentences.

The results obtained with the GoeSa sentences were not as clear as for the OlSa sentences. For the GoeSa sentences, binaural release from masking was only observed for the unprocessed condition. In the time-compressed and low-pass filtered condition, the large standard deviation across listeners made it impossible to draw conclusions about the binaural release from masking. Moreover, by applying low-pass filtering to 1500 Hz and time compression to 0.66 of the original length, SRTs were increased by 15 dB for the open-set GoeSa sentences but only by 5 dB for the closed-set OlSa sentences. This finding is in line with observations made by Rennies et al. (2014) and Warzybok et al. (2016), where the combined effect of reverberation and SNR on speech intelligibility (and listening effort) was investigated using OlSa sentences (Rennies et al., 2014) and GoeSa sentences (Warzybok et al., 2016): GoeSa sentences were more affected by additional reverberation than OlSa sentences, meaning that the intelligibility for a certain combination of noise and reverberation is higher for OlSa sentences than for GoeSa sentences. This might indicate that the GoeSa sentences in general are more affected by manipulations (like time compression and low-pass filtering) than OlSa sentences. This might be caused by the special structure of OlSa sentences which consist of 5 words drawn from a pool of 50 words and have a fixed grammatical structure, which makes it easier for the listeners to generate an expectation.

Future Work

While Experiment II showed that it is difficult to quantify binaural unmasking at positive SNRs by measuring SRTs with listeners with normal hearing, listening effort has been shown to be affected by the SNR even at 100% speech intelligibility. In Rennies and Kidd (2018), a binaural benefit was observed in measurements of listening effort, where a spatial separation of the target source from the interfering source resulted in reduced listening effort, similar to the binaural unmasking observed for SRTs presented in the current study. In principle, the BSIM2020 model could be applied to predict binaural listening effort in its current version. This requires the modeling of speech in noise at positive SNRs, which is possible with the BSIM2020 model using a readjustment of the reference SII value to fit the SNR range where listening effort changes but speech intelligibility is already saturated. Instead of converting SII values to SRTs, a conversion of SII values to the different categorical effort scaling values as measured, for example, by the method of (Krueger et al., 2017a) would be needed. Note that even when 100% speech intelligibility has been achieved, listening effort can still be further reduced by increasing the SNR which is related to a further increase of the SII with increasing SNR above the SRT (Krueger et al., 2017b).

Conclusions

This study presents a new binaural speech intelligibility model termed *BSIM2020* which predicts SRTs for arbitrary SNRs. The model combines blind EC processing below 1500 Hz with blind better-ear listening above 1500 Hz. The optimal EC processing strategy (either minimization or maximization of the EC output level) and the better ear are selected based on the SRMR measure which maximizes speech-like modulations. This selection is performed independently in each auditory frequency channel. The output of the blind binaural processing stage is a single (mono) signal which can be used in combination with arbitrary back ends for speech intelligibility prediction.

We found that the model gave RMSEs from experimental data of less than 1 dB (Experiment I) or 2.5 dB (Experiment II). In that second experiment, the increase may have been due to the large variance in the observed data, especially for SRTs at positive SNRs. Blind-level minimization of the output of the EC process is sufficient to describe results of listening experiments at negative SNRs, which is usually the case in SRT_{50} measurements with listeners with normal hearing.

Our experimental results demonstrated that binaural release from masking also occurred at positive SNRs. We did this using time-compressed and low-pass filtered OlSa sentences. These two manipulations preserved the binaural cues and can be regarded as simple simulations of reduced cognitive processing speed and high-frequency hearing loss.

Author's Note

The BSIM2020 is freely available at http://medi.uni-oldenburg.de/BSIM_2020/

Acknowledgment

The authors would like to thank the associate editor Michael Akeroyd for his very supportive comments on an earlier version of this article, which helped to improve it a lot.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by the Cluster of Excellence "Hearing4All" and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project Number 352015383—SFB 1330 A 1.

ORCID iD

Christopher F. Hauth https://orcid.org/0000-0002-9067-5874

Supplemental material

Supplementary material for this article is available online.

References

Andersen, A. H., de Haan, J. M., Tan, Z.-H., & Jensen, J. (2016). Predicting the intelligibility of noisy and nonlinearly processed binaural speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*, 1908–1920. https://doi.org/10.1109/TASLP.2016.2588002

Andersen, A. H., de Haan, J. M., Tan, Z., & Jensen, J. "“A non-intrusive Short-Time Objective Intelligibility measure,” *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017*, pp. 5085–5089, doi: 10.1109/ICASSP.2017.7953125.

ANSI S3.5-1997. (1997). *Methods for calculation of the speech intelligibility index*.

Beutelmann, R., & Brand, T. (2006). Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. *Journal of the*

Acoustical Society of America, *120*, 331–342. https://doi.org/10.1121/1.2202888

Beutelmann, R., Brand, T., & Kollmeier, B. (2009). Prediction of binaural speech intelligibility with frequency-dependent interaural phase differences. *Journal of the Acoustical Society of America*, *126*, 1359–1368. https://doi.org/10.1121/1.3177266

Beutelmann, R., Brand, T., & Kollmeier, B. (2010). Revision, extension, and evaluation of a binaural speech intelligibility model. *Journal of the Acoustical Society of America*, *127*, 2479–2497. https://doi.org/10.1121/1.3295575

Boersma, P., & van Heuven, V. (2001). Speak and unSpeak with PRAAT 5, 8. *Glot International*, *5*(9/10), 341–347.

Boersma, P., & Weenink, D. (2018). *Praat: Doing phonetics by computer* [Computer program]. https://www.fon.hum.uva.nl/praat/

Brand, T., & Kollmeier, B. (2002). Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *Journal of the Acoustical Society of America*, *111*, 2801–2810. https://doi.org/10.1121/1.1479152

Bronkhorst, A. (2000). The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*, *86*, 117–128.

Brons, I., Houben, R., & Dreschler, W. A. (2013). Perceptual effects of noise reduction with respect to personal preference, *speech intelligibility, and listening effort. Ear and Hearing*, *34*, 29–41. https://doi.org/10.1097/AUD.0b013e31825f299f

Chabot-Leclerc, A., MacDonald, E. N., & Dau, T. (2016). Predicting binaural speech intelligibility using the signal-to-noise ratio in the envelope power spectrum domain. *Journal of the Acoustical Society of America*, *140*, 192–205. https://doi.org/10.1121/1.4954254

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, *25*, 975–979. https://doi.org/10.1121/1.1907229

Cosentino, S., Marquardt, T., McAlpine, D., Culling, J. F., & Falk, T. H. (2014). A model that predicts the binaural advantage to speech intelligibility from the mixed target and interferer signals. *Journal of the Acoustical Society of America*, *135*, 796–807. https://doi.org/10.1121/1.4861239

Culling, J. F., Hawley, M. L., & Litovsky, R. Y. (2004). The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources. *Journal of the Acoustical Society of America*, *116*, 1057–1065. https://doi.org/10.1121/1.1772396

Culling, J. F., Hawley, M. L., & Litovsky, R. Y. (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)]. Journal of the Acoustical Society of America*, *118*, 552–552. https://doi.org/10.1121/1.1925967

Dau, T., Püschel, D., & Kohlrausch, A. (1996). A quantitative model of the “effective” signal processing in the auditory system. I. Model structure. *Journal of the Acoustical Society of America*, *99*, 3615–3622. https://doi.org/10.1121/1.414959

Dietz, M., Ewert, S. D., & Hohmann, V. (2011). Auditory model based direction estimation of concurrent speakers from binaural signals. *Speech Communication Perceptual and Statistical Audition*, *53*, 592–605. https://doi.org/10.1016/j.specom.2010.05.006

Durlach, N. I. (1963). Equalization and cancellation theory of binaural masking-level differences. *Journal of the Acoustical Society of America*, *35*, 1206–1218. https://doi.org/10.1121/1.1918675

Egan, J. P. (1964). Masking-level differences as a function of interaural disparities in intensity of signal and of noise. *Journal of the Acoustical Society of America*, *36*, 1992–1992. https://doi.org/10.1121/1.1939216

Fredelake, S., Holube, I., Schlueter, A., & Hansen, M. (2012). Measurement and prediction of the acceptable noise level for single-microphone noise reduction algorithms. *International Journal of Audiology*, *51*, 299–308. https://doi.org/10.3109/14992027.2011.645075

Geravanchizadeh, M., & Fallah, A. (2015). Microscopic prediction of speech intelligibility in spatially distributed speech-shaped noise for normal-hearing listeners. *Journal of the Acoustical Society of America*, *138*, 4004–4015. https://doi.org/10.1121/1.4938230

Green, D. M. (1966). Signal-detection analysis of equalization and cancellation model. *Journal of the Acoustical Society of America*, *40*, 833–838. https://doi.org/10.1121/1.1910155

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

Jelfs, S., Lavandier, M., & Culling, J. F. (2011). Revision and validation of a binaural model for speech intelligibility in noise. *Hearing Research, 275*, 1–2, 96–104, https://doi.org/10.1016/j.heares.2010.12.005

Kollmeier, B., Schädler, M. R., Warzybok, A., Meyer, B. T., & Brand, T. (2016). Sentence recognition prediction for hearing-impaired listeners in stationary and fluctuation noise with FADE: Empowering the attenuation and distortion concept by plomp with a quantitative processing model. *Trends in Hearing, 20, 233121651665579*. https://doi.org/10.1177/2331216516655795

Kollmeier, B., & Wesselkamp, M. (1997). Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. *Journal of the Acoustical Society of America*, *102*, 2412–2421. https://doi.org/10.1121/1.419624

Krueger, M., Schulte, M., Brand, T., & Holube, I. (2017a). Development of an adaptive scaling method for subjective listening effort. *Journal of the Acoustical Society of America*, *141*, 4680–4693. https://doi.org/10.1121/1.4986938

Krueger, M., Schulte, M., Zokoll, M. A., Wagener, K. C., Meis, M., Brand, T., & Holube, I. (2017b). Relation between listening effort and speech intelligibility in noise. *American Journal of Audiology*, *26*, 378–392. https://doi.org/10.1044/2017_AJA-16-0136

Langford, T. L., & Jeffress, L. A. (1964). Effect of noise cross-correlation on binaural signal detection. *Journal of the Acoustical Society of America*, *36*, 1455–1458. https://doi.org/10.1121/1.1919224

Lavandier, M., & Culling, J. F. (2010). Prediction of binaural speech intelligibility against noise in rooms. *Journal of the Acoustical Society of America*, *127*, 387–399. https://doi.org/10.1121/1.3268612

Lavandier, M., Jelfs, S., Culling, J. F., Watkins, A. J., Raimond, A. P., & Makin, S. J. (2012). Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources. *Journal of the Acoustical Society of America*, *131*, 218–231. https://doi.org/10.1121/1.3662075

Levitt, H., & Rabiner, L. R. (1967). Binaural release from masking for speech and gain in intelligibility. *Journal of the Acoustical Society of America*, *42*, 601–608. https://doi.org/10.1121/1.1910629

Licklider, J. C. R. (1948). The Influence of interaural phase relations upon the masking of speech by white noise. *Journal of the Acoustical Society of America*, *20*, 150–159. https://doi.org/10.1121/1.1906358

Moore, B. C. J., & Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, *74*, 750–753. https://doi.org/10.1121/1.389861

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*, 453–467. https://doi.org/10.1016/0167-6393(90)90021-Z

Ostendorf, M., Hohmann, V., & Kollmeier, B. (1998). Klassifikation von akustischen Signalen basierend auf der Analyse von Modulationsspektren zur Anwendung in digitalen Hörgeräten [Classification of acoustical signals based on the analysis of modulationspectra for the application in digital hearing devices]. *Fortschritte der Akustik – DAGA, 1998*, 402–403.

Rennies, J., & Kidd, G. (2018). Benefit of binaural listening as revealed by speech intelligibility and listening effort. *Journal of the Acoustical Society of America*, *144*, 2147–2159. https://doi.org/10.1121/1.5057114

Rennies, J., Schepker, H., Holube, I., & Kollmeier, B. (2014). Listening effort and speech intelligibility in listening situations affected by noise and reverberation. *Journal of the Acoustical Society of America*, *136*, 2642–2653. https://doi.org/10.1121/1.4897398

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *26*, 43–49. https://doi.org/10.1109/TASSP.1978.1163055

Santos, J. F., Senoussaoui, M., & Falk, T. H. (2014, September). *An improved non-intrusive intelligibility metric for noisy and reverberant speech* 2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC), Juan-les-Pins, 2014, pp. 55–59, doi: 10.1109/IWAENC.2014.6953337

Schädler, M. R., Warzybok, A., Ewert, S. D., & Kollmeier, B. (2016). A simulation framework for auditory discrimination experiments: Revealing the importance of across-frequency processing in speech perception. *Journal of the Acoustical Society of America*, *139*, 2708–2722. https://doi.org/10.1121/1.4948772

Schlueter, A., Brand, T., Lemke, U., Nitzschner, S., Kollmeier, B., & Holube, I. (2015). Speech perception at positive signal-to-noise ratios using adaptive adjustment of time compression. *Journal of the Acoustical Society of America*, *138*, 3320–3331. https://doi.org/10.1121/1.4934629

Smeds, K., Wolters, F., & Rung, M. (2015). Estimation of signal-to-noise ratios in realistic sound scenarios. *Journal of the American Academy of Audiology*, *26*, 183–196. https://doi.org/10.3766/jaaa.26.2.7

Spille, C., Ewert, S. D., Kollmeier, B., & Meyer, B. T. (2018). Predicting speech intelligibility with deep neural networks. *Computer Speech & Language*, *48*, 51–66. https://doi.org/10.1016/j.csl.2017.10.004

Steeneken, H. J. M., & Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *Journal of the Acoustical Society of America*, *67*, 318–326. https://doi.org/10.1121/1.384464

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2011). An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech and Language Processing*, *19*, 2125–2136. https://doi.org/10.1109/TASL.2011.2114881

Tang, Y., Liu, Q., Wang, W., & Cox, T. J. (2018). A non-intrusive method for estimating binaural speech intelligibility from noise-corrupted signals captured by a pair of microphones. *Speech Communication*, *96*, 116–128. https://doi.org/10.1016/j.specom.2017.12.005

vom Hövel, H. (1984). *Zur Bedeutung der Übertragungseigenschaften des Aussenohrs sowie des Binauralen Hörsystems bei Gestörter Sprachübertragung* [On the importance of the transmission properties of the outer ear and the binaural auditory system in disturbed speech transmission] [PhD dissertation]. RWTH Aachen.

Algazi, V. R., Duda R. O., Thompson D. M. and Avendano C., "The CIPIC HRTF database," *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, New Platz, NY, USA, 2001, pp. 99–102, doi: 10.1109/ASPAA.2001.969552.

Wagener, K. C., Brand, T., & Kollmeier, B. (1999a). 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]. *Zeitschrift für Audiologie/Audiological Acoustics*, *38*, 44–56.

Wagener, K. C., Brand, T., & Kollmeier, B. (1999b). 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]. *Zeitschrift für Audiologie/Audiological Acoustics*, *38*, 86–95.

Wagener, K. C., Kühnel, V., & Kollmeier, B. (1999b). 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]. *Zeitschrift für Audiologie/Audiological Acoustics*, *38*, 4–15.

Wan, R., Durlach, N. I., & Colburn, H. S. (2010). Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers. *Journal of the Acoustical Society of America*, *128*, 3678–3690. https://doi.org/10.1121/1.3502458

Warzybok, A., Rennies-Hochmuth, J., & Kollmeier, B. (2016, March). Masking versus cognition during speech recognition in noise and reverberation: Can different sentence tests provide a quantitative estimate? [Paper presentation]. Proceedings of the German Annual Conference on Acoustics (DAGA), Aachen, Germany.