CARL VON OSSIETZKY universität OLDENBURG

Documentation of the binaural speech intelligibility model (BSIM)

developed by
Rainer Beutelmann

written by
Christopher Hauth and Thomas Brand

Oldenburg, February 9, 2017

1 Binaural Speech Intelligibility Model (BSIM)

In Beutelmann and Brand (2006) and Beutelmann *et al.* (2010), a computational model was introduced, which goes back to vom Hövel (1984) and is able to predict the outcome of binaural speech intelligibility experiments in various acoustic conditions. The model combines an Equalization-Cancellation (EC) mechanism, as proposed by Durlach (1963), and the Speech Intelligibility Index (ANSI, 1997), which transforms frequency dependent SNRs into a certain speech intelligibility index. The resulting index can then be transformed to a speech reception threshold, where 50% of the presented speech is intelligible.

In this binaural speech intelligibility model (`BSIM`), the spatial release from masking for normal-hearing and hearing-impaired listeners can be predicted as well as the phenomenon of binaural unmasking due to interaural time and level differences. The individual hearing is accounted for by taking the audiogram into consideration.

In the following, a detailed description of the computational model, which was presented in Beutelmann *et al.* (2010), is given. Therefore, it will be labeled `BSIM2010`. This computational model is a computer program, which was written by Rainer Beutelmann. In this program, several important processing steps are required, for example the calibration, which are not described in detail in Beutelmann *et al.* (2010) and will, therefore, described in detail in this chapter. Furthermore, an official release of the `BSIM2010` is planned. Therefore, a detailed description of its signal processing might be a help for potential users of this computer model. The major difference to `BSIM 2006` is that Monte-Carlo simulations are no longer necessary as there was found an analytical way to incorporate processing errors, which will also be addressed. Furthermore, a short-time version of the `BSIM2010` was introduced in Beutelmann *et al.* (2010), which can - in contrast to the long-time version - account for amplitude modulated interferer by analyzing short time segments of the signals. First, a general overview of the model architecture is given. Then, special attention is given to the Equalization-Cancellation process and its implementation as well as the SII. All processing steps are described and commented with respect to literature and the current development of the `BSIM` model. In the end, some drawbacks of the model are highlighted in order to point out possible improvements in terms of plausibility, which are an important topic of this thesis.

1.1 Input to BSIM and Calibration

`BSIM2010` is able to predict speech intelligibility in various acoustic conditions, but requires a priori knowledge about speech and noise to achieve this aim. Therefore, target speech and interfering noise need to be delivered to `BSIM2010`, separately, resulting in four input signals (two for each ear).

First, a calibration of speech and noise signals is required. As this calibration is not described in detail in `BSIM2010` it is described here. The signals are calibrated

to a certain level, for example 65 dB. To achieve the desired level, the mean level between left and right ear channel is calculated and its difference to the desired level is determined for a reference condition (for example speech and noise co-located at an azimuth of 0°). Afterwards, the difference is applied as gain to the left and right ear channel for each of the remaining conditions (for example the remaining azimuth angles of the noise source). By using this procedure, the ILDs between left and right ear side are preserved while the overall level is adjusted.
After this calibration, the adjusted input signals can be processed by `BSIM2010`.

```
% Example code snippet for the calibration of BSIMs input signals
%
% define filepath of speech and noise signals for the reference condition
%(for example speech and noise co-located at 0Â° azimuth (denoted as S_azim))
file_path_noise  = ['.' filesep room filesep 'noise_',S_azim,'.wav'];
file_path_signal = ['.' filesep room filesep 'speech_',S_azim,'.wav'];

% load files defined above
[N fs_N] = wavread(file_path_noise);
[S fs_S] = wavread([file_path_signal]);

% calibration of signals to same level for noise and speech
ref = 1;                            % for dB FS (acc. to R. Beutelman,
                                    % personal communication)
lev2be = 65;                        % level at which speech is presented

lev_S = 20*log10(rms(S)/ref);       % actual rms-level of speech
lev_S = mean(lev_S);                % frontal speech, should be the same at
                                    % each ear, reference is MEAN level
                                    % between the two ears
Delta_L_speech = lev2be - lev_S;

lev_N = 20*log10(rms(N)/ref);       % actual rms-level of noise
lev_N = mean(lev_N);                % reference is MEAN level between the
                                    % two ears
Delta_L_noise = lev2be - lev_N;

% Difference between the desired RMS and the measured RMS (denoted as
% Delta_L_noise and Delta_L_speech) is saved as constant gain factor and is
% applied for the remaining conditions (other noise azimuths)
```

1.2 BSIM2010

In Figure 1, the processing scheme of the computational model `BSIM2010` is shown, which has already been used in several studies (for example Beutelmann *et al.*, 2010; Rennies *et al.*, 2011, 2014). First, the incoming speech and noise signals are filtered using a gammatone filterbank (Hohmann, 2002) ranging from 146 to 8346 Hz in 30 ERB spaced frequency bands (Glasberg and Moore, 1990), mirroring the frequency selectivity on the basilar membrane. Frequencies below 146 Hz and above 8346 Hz are assumed to provide negligible information for speech intelligibility and

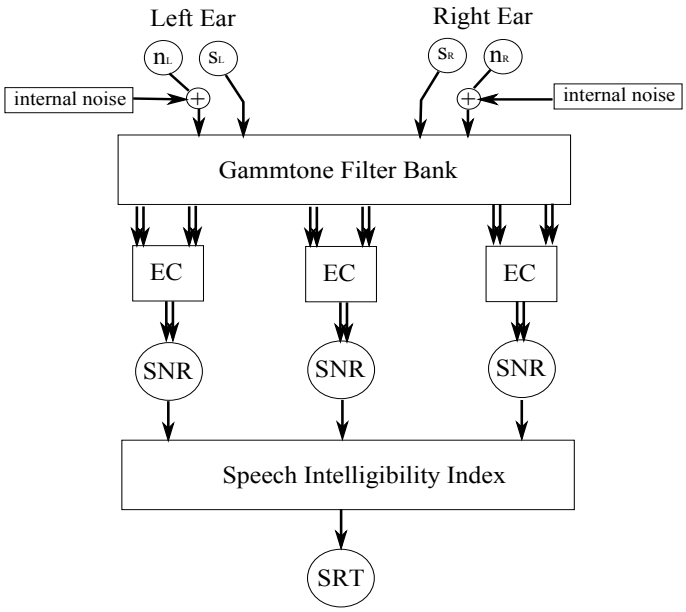

Figure 1: *Scheme of the signal processing of* `BSIM2010`*. First, a gammatone filterbank is applied to simulate the frequency selectivity of the basilar membrane. Then, the EC mechanism, which is maximizing the SNR, is applied in each frequency band, independently, and the best SNR is calculated between the monaural SNRs and the binaural SNR of the residual signal. In a last step, the frequency specific SNRs are transformed into a Speech Reception Threshold (SRT) by the SII (Adapted from Beutelmann* et al.*, 2010, Figure 1).*

are not considered for processing.

Even though it is mentioned in Beutelmann *et al.* (2010) that uncorrelated internal masking noises are added to the left and right ear signals to account for the individual hearing threshold, it is not explicitly done in the `BSIM2010` computation. That means that the EC parameters are estimated without taking the individual hearing threshold into consideration, i.e. the estimation of ITD and ILD parameters is independent of the hearing threshold.[1]

As a first step, the ITD τ is estimated in each frequency band by searching the position of the maximum of the interaural cross-correlation (ICC). The maximum possible lag is limited to be within the period of the corresponding gammatone filter, to avoid ambiguities due to the periodicity of the band limited signals. This is done for both speech and noise signals, separately, i.e. the time lag corresponding to the speech source and the noise source is estimated.

Afterwards, an attenuation factor α is computed for the speech and noise signals, which mirrors the equalization in level of the EC-process. However, the attenuation does not compensate for the level difference, but is chosen such that the SNR in each frequency band is maximized if the subtraction in the cancellation stage is applied. After calculating the equalization parameters, the subtraction operation can be applied. By choosing an optimal set of equalization parameters, the SNR

[1]This was done to reduce the computational effort for applications in audiology where many different listeners are tested in the same acoustical condition.

is maximized by applying the cancellation step, which subtracts the left ear signal from the right ear signal. Therefore, the EC stage of `BSIM2010` provides an optimal SNR improvement that is independent of the input SNR. However, the estimation of EC parameter that is implemented in `BSIM2010` differs slightly from the formulas in Beutelmann *et al.* (2010). Depending on the maximum achievable SNR, either the set of parameters corresponding to the speech signal or to the noise signal are applied to speech and noise. [2]

After the interaural parameters are estimated, such that the optimal SNR after the cancellation is achieved, the uncorrelated masking noise mirroring the individual hearing threshold is added to the external noise, before the EC-process is applied. On the one hand, the uncorrelated noise can not be attenuated by the EC mechanism as it is binaurally uncorrelated. On the other hand, the noise does not affect the estimation of the EC parameters, because it is introduced afterwards. This implementation is not realistic, because EC parameters are estimated even if the hearing threshold indicates that no acoustical information is available in a specific frequency region.

The EC-process is applied to each frequency channel, independently. For example, in Beutelmann *et al.* (2009) it was shown that binaural processing can be assumed as acting independently across frequency channels.In the experiment mentioned in Beutelmann *et al.* (2009), the interaural phase was sinusoidally modulated across frequency, i.e. a band wise EC-process is beneficial in this kind of task. The experimental results were successfully predicted by `BSIM2010`. However, best fit was achieved by increasing the bandwidth of an auditory filter from 1 ERB to 2.3 ERB, leading to the assumption that broader auditory filter are used in binaural processing. In the remainder of this thesis, an ERB spacing of 1 was chosen. Even though Beutelmann *et al.* (2010) was submitted previous to Beutelmann *et al.* (2009), it got accepted later. Furthermore, `BSIM2010` was used in several studies using an ERB spacing of 1 and in order to be consistent with Beutelmann and Brand (2006) and the other studies, here an ERB spacing of 1 is used, too. Nevertheless, the 2.3 ERB wide binaural filter bank should be used in future versions of the model.

The result of the EC-process is a binaural SNR, or more specific, the binaural SNR can be calculated from the residual signal, which is the signal after the EC processing of the monaural signals. The SNR of the residual signal is compared to the monaural SNRs of the left and right ear channel, which can account for better-ear listening (Bronkhorst and Plomp, 1988). The maximum value of the three SNRs in each frequency band is then fed to the SII, which transforms the SNR values into into a value between 0 and 1, which can then be mapped to a Speech Reception Threshold (SRT), i.e. the SNR in dB, where 50% of the presented speech is correctly understood.

Deviating from other binaural speech intelligibility models, for example introduced by Cosentino *et al.* (2014), either binaural or monaural SNRs determine the result-

[2] In general, reducing the noise source, i.e. using the EC parameters estimated from the noise, will provide the best SNR.

ing speech intelligibility. In the mentioned model by Cosentino *et al.* (2014), both binaural unmasking and better-ear listening are assumed to be cumulative effects of binaural hearing.
In the following, the underlying mathematical calculations of the EC-process and of the SII are described.

1.3 Equalization-Cancellation Process

The EC-mechanism, as proposed by Durlach (1963), is a mathematical description of a binaural processor, which is able to predict the effect of binaural unmasking. It exploits different interaural configurations of target and interferer, resulting, for example, from their different locations in space. It is assumed that an EC-process attenuates the external noise by, first, compensating for interaural disparities and second, by subtracting the in level and time adjusted left ear channel from the right ear channel or vice versa.
The time-domain representation of an EC-process can be expressed as

$$x_{EC}(t) = \alpha x_L(t + \tau) - x_R(t), \tag{1}$$

where τ is the compensation of the interaural delay (ITD) and α the compensation of interaural level difference (ILD). In principle τ and α can be both positive and negative, depending on the ear, where the sound is higher level or leading in time.
In Figure 2, a scheme of the equalization process, which is applied to a noisy sinusoidal signal, is shown. After the signals have been adjusted in level and time, as

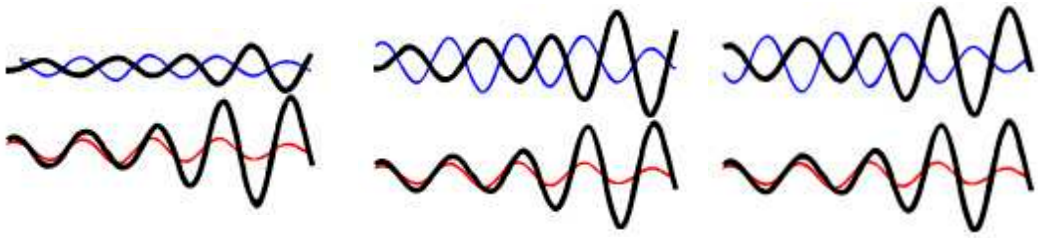

Figure 2: *In the left panel, a binaural signal is shown, where the target is interaurally in phase (red and blue lines), while the interferer is interaurally out of phase (black line). In the middle panel, the interaural level difference is compensated by amplifying the left ear channel. In the right panel, the delay of the interferer in the left ear channel is compensated for.*

it is illustrated in the middle and right panel of Figure 2, the cancellation process can be applied by subtracting the left from the right channel. The outcome of the cancellation process, which is applied to the right panel of Figure 2, can be seen in Figure 3. Equation 1 can be rewritten as symmetric expression by applying $\tau/2$ to the left channel and $-\tau/2$ to the right channel, respectively. The same can be done for the attenuation factor α: $\sqrt{\alpha}$ is applied to the left ear signal and $1/\sqrt{\alpha}$ to the right ear signal. By substituting α by e^{γ}, the time-domain EC-process can be

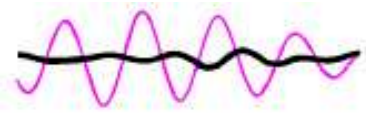

Figure 3: *Residual Signal after EC processing. The black line, indicating the noise, is attenuated by destructive interference, while the target signal (magenta) is enhanced by constructive interference.*

re-written as the symmetrical expression

$$x_{EC}(t) = e^{\gamma/2} x_L(t + \tau/2) - e^{-\gamma/2} x_R(t - \tau/2). \tag{2}$$

By applying the Fourier transform (`FT`) to Equation 2, the frequency domain representation can be obtained by

$$X_{EC}(\omega) = e^{\gamma/2 + j\omega\tau/2} X_L(\omega) - e^{-\gamma/2 - j\omega\tau/2} X_R. \tag{3}$$

The delay term τ is thereby changed into a multiplication by a linear phase term $e^{j\omega\tau/2}$, while the attenuation factor stays the same. In `BSIM2010`, the FT is realized by calculating the fast fourier transform (FFT).
However, the EC-process can not be assumed as perfect operation and, therefore, uncertainties in level (ϵ) and time (δ) are incorporated. A perfectly operating EC-mechanism heavily overestimates the effect of binaural unmasking and fails, for example, to predict the spatial release from masking (see results without processing errors in Beutelmann and Brand, 2006, Figure 2).
These processing errors are assumed to be normally distributed random variables (RVs), which are defined by zero mean and standard deviations σ_δ and σ_ϵ. Because of these assumptions, it is possible to replace the Monte-Carlo simulations used in Beutelmann and Brand (2006) by an analytical solution. The standard deviation of the normally distributed processing errors is given by

$$\sigma_\epsilon = \sigma_{\epsilon 0}[1 + (\frac{|\alpha|}{\alpha_0})^p] \tag{4}$$

and

$$\sigma_\delta = \sigma_{\delta 0}[1 + \frac{|\Delta|}{\Delta_0}], \tag{5}$$

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 (vom Hövel, 1984). These equations state that the standard deviation of level and delay errors increases as the time or level differences between both ears get larger. In Figure 4, the standard deviations of both processing errors are shown for different values of ITD and ILD. The standard deviation of the internal delay error is a linear function of the absolute value of the ITD, while the standard deviation of the internal level error is exponentially growing with increasing absolute value of the ILD.
By incorporating these processing errors into Equation 3, the frequency domain

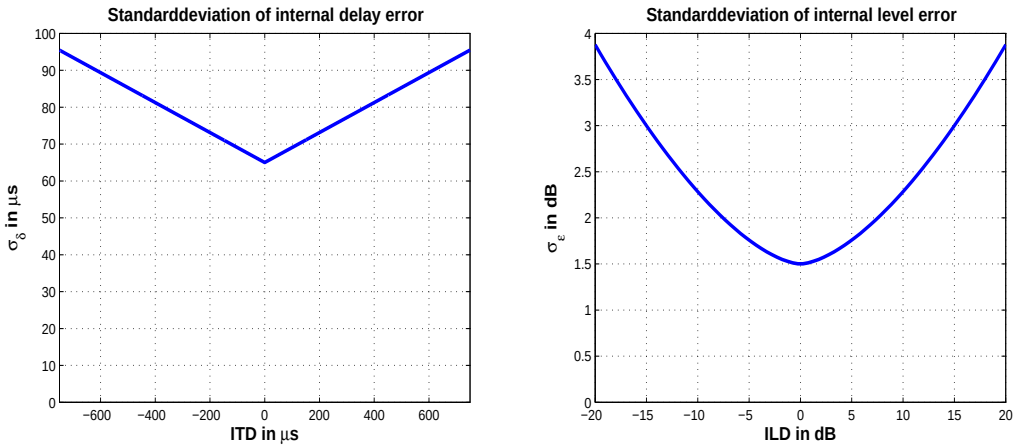

Figure 4: *In the left panel, the standard deviation of the internal delay error for ITDs ranging from -750µs to 750µs is shown. In the right panel, the standard deviation of the internal level error for ILDs ranging from -20 dB to 20 dB is shown.*

representation is expanded to

$$X_{EC}(\omega) = e^{\gamma/2+\epsilon_L+j\omega(\tau/2+\delta_L)}X_L(\omega) - e^{-\gamma/2+\epsilon_R-j\omega(\tau/2-\delta_R)}X_R. \tag{6}$$

It is necessary to compute the SNR in each frequency channel separately, which can serve as input to the SII. Therefore, the intensity of the residual speech and noise signal (after the EC-process has been applied) needs to be computed in each frequency channel separately for speech and noise.
The intensity of the speech signal in a certain gammatone filter can be obtained by

$$I(S_{EC}) = \int_{\Omega-\beta/2}^{\Omega+\beta/2} |S_{EC}|^2 \mathrm{d}\omega, \tag{7}$$

where Ω is the center frequency of a gammatone filter and β its bandwidth. The squared magnitude of the residual speech is computed and integrated over the bandwidth of each gammatone filter. By substituting the speech signal after EC processing ($|S_{EC}|^2$) in Equation 7 by the mathematical operation of the EC process, Equation 7 can be rewritten as

$$I(S_{EC}) = \int_{\Omega-\beta/2}^{\Omega+\beta/2} |e^{\gamma/2+\epsilon_L+j\omega(\tau/2+\delta_L)}S_L(\omega) - e^{-\gamma/2+\epsilon_R-j\omega(\tau/2-\delta_R)}S_R|^2 \mathrm{d}\omega. \tag{8}$$

By applying

$$|x-y|^2 = |x|^2 + |y|^2 - 2\Re(xy^*) \tag{9}$$

to Equation 8, the expression can be rewritten as

$$
\begin{aligned}
I(S_{EC}) = e^{\gamma+2\epsilon_L} \int_{\Omega-\beta/2}^{\Omega+\beta/2} |S_L(\omega)| d\omega \\
+ e^{-\gamma+2\epsilon_R} \int_{\Omega-\beta/2}^{\Omega+\beta/2} |S_R(\omega)| d\omega \\
- 2e^{\epsilon_L+\epsilon_R} \mathrm{x} \Re(\int_{\Omega-\beta/2}^{\Omega+\beta/2} S_L(\omega) S_R^*(\omega) e^{j\omega(\delta_L+\delta_R)} e^{j\omega\tau} d\omega).
\end{aligned} \tag{10}
$$

Equation 10 can be described as summation of the intensity of the left and right channel minus a phase dependent cross-correlation term. The calculation, which is performed in `BSIM2010`, differs compared to the binaural speech intelligibility model described in Lavandier and Culling (2010). There, the binaural advantage is calculated as a summation of the binaural unmasking component and the monaural channel, which has the favorable SNR. In Beutelmann *et al.* (2010), every quantity derived from the residual signals is assumed to be describable with its expected value and with normally distributed processing errors. By calculating the expected value, Monte-Carlo simulations as performed in Beutelmann and Brand (2006) can be replaced by an analytical solution, as the mean value over a large amount of Monte Carlo simulations and calculating the solution by incorporating the expectation value of each processing error, should give the same result. Note that for example Wan *et al.* (2010) still apply Monte-Carlo simulations within an EC model for speech intelligibility prediction.

Therefore, the expectation value of the intensity with respect to the processing errors needs to be computed. As mentioned before, the processing errors are normally distributed random variables. The terms e^{ϵ} and e^{δ} in Equation 10 are, therefore, logarithmic-normally distributed RVs. The expected value of a logarithmic-normally distributed random variable X can be calculated via

$$
E(X) = \frac{1}{\sqrt{2\pi}\sigma} \int_0^{\infty} e^{\frac{-(\ln x-\mu)^2}{2\sigma^2}} \tag{11}
$$

and

$$
E(X) = e^{\mu+\frac{\sigma^2}{2}}, \tag{12}
$$

where μ and σ denote mean value and standard deviation, respectively. If the exponent of a log-normal distribution is multiplied by a factor N, the expected value is calculated according to

$$
E(X^N) = e^{\mathrm{N}\mu+\frac{\sigma^2\mathrm{N}^2}{2}}. \tag{13}
$$

The expected value of e^{ϵ} and e^{δ} is calculated accordingly. The expectation value of the level processing error e^{ϵ}, which has zero mean and standard deviation σ_{ϵ}, is

$$E(e^{\epsilon}) = e^{\frac{\sigma_{\epsilon}^2}{2}} \tag{14}$$

and

$$E(e^{2\epsilon}) = e^{\sigma_{\epsilon}^2}. \tag{15}$$

The same calculation is done for the delay error $e^{j\omega\delta}$ and $e^{j\omega 2\delta}$, resulting in

$$E(e^{j\omega\delta}) = e^{-\omega^2 \frac{\sigma_{\delta}^2}{2}} \tag{16}$$

and

$$E(e^{j\omega 2\delta}) = e^{-\omega^2 \sigma_{\delta}^2}. \tag{17}$$

By incorporating the expected values of the processing errors, the expected value of the speech intensity is obtained via

$$\begin{aligned} < I(S_{EC}) >_{\epsilon_L,\epsilon_R,\delta_L,\delta_R} &= e^{2\sigma_{\epsilon}^2} e^{\gamma} I(S_L) \\ &+ e^{2\sigma_{\epsilon}^2} e^{-\gamma} I(S_R) \\ &- 2e^{\sigma_{\epsilon}^2} \mathrm{x}\Re\Big(\int_{\Omega-\beta/2}^{\Omega+\beta/2} S_L(\omega) S_R^*(\omega) e^{-\omega^2 \sigma_{\delta}^2} e^{j\omega\tau} d\omega\Big). \end{aligned} \tag{18}$$

1.4 Optimization of EC parameters

Given Equation 18 for both speech and noise, it is possible to compute the resulting SNR. However, there are two independent variables (delay and attenuation), which have to be adjusted to maximize the SNR. In `BSIM2010`, the ITD for both speech and noise is estimated using the interaural cross-correlation (ICC). The delay of the position of the maximum thereby describes the ITD. Given an ITD, only one independent variable α remains and an optimal α can be computed, such that the SNR is maximized. By deriving the equation with respect to α and computing the zeros, an α, which leads to a maximal SNR can be obtained according to

$$\frac{\partial}{\partial\alpha}\Big(\frac{< I(S_{EC}) >_{\epsilon_L,\epsilon_R,\delta_L,\delta_R}}{< I(N_{EC}) >_{\epsilon_L,\epsilon_R,\delta_L,\delta_R}}\Big) \stackrel{!}{=} 0. \tag{19}$$

In parallel to the binaural processor, monaural SNRs are calculated in each frequency band in order to account for better ear listening, which becomes more important in higher frequency regions, where a binaural benefit due to interaural phase difference is negligible. In general, the SNR calculated by the binaural processor in `BSIM2010` should be at least as good as the better monaural SNR. However, due to the processing errors, the better ear might be slightly larger in some cases. To

account for this it is necessary to compute the monaural SNRs. The best SNR, which is either the left or right monaural SNR or the SNR resulting from binaural processing, is calculated in each frequency band and fed to the SII. The resulting SII is then mapped to a speech reception threshold (SRT), where 50% of the speech is comprehensible.

1.5 Interpretation of EC processing errors as low-pass filter of the ICC

From Equation 18 it can be seen that the expected value of the random variable describing the processing error in time ($e^{-\omega^2\sigma_\delta^2}$) is an exponentially decreasing function with increasing frequency. That means a low-pass filter is applied to the interaural cross correlation term. Its cutoff frequency is thereby influenced by the aforementioned standard deviation of the delay error. In Figure 5, the transfer functions of the low-pass filter for an ITD of 0μs and the maximum anatomical possible ITD of 750μs and, therefore, different standard deviations of the delay error, are shown. Physiologically, this low-pass filter corresponds to the loss of phase-locking of the

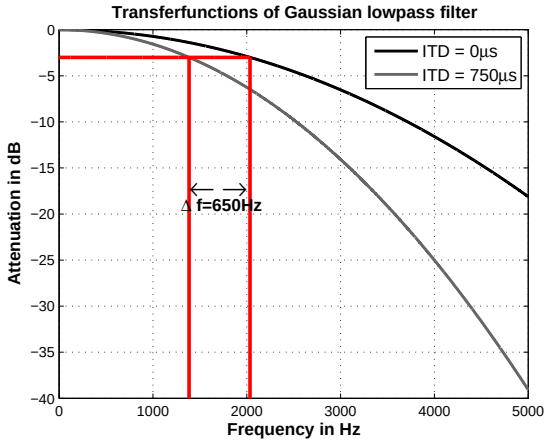

Figure 5: *Transfer function of Gaussian low-pass filter for ITDs of* 0μs *and* 750μs*. The red lines indicate the 3 dB cut-off frequency.* $\Delta f = 650$ *Hz is the difference in cutoff frequency between both ITDs.*

inner hair cells (IHC) with increasing frequency and, therefore, a loss of phase sensitivity (Palmer and Russell, 1986).

The Gaussian low-pass filter can also be used to individualize the prediction. As already mentioned before, the Gaussian low-pass filter describes the sensitivity to interaural phase differences. However, this sensitivity might differ between individual listeners and between the hearing impaired and normal hearing population. The 3 dB cutoff frequency (f_{cut}), which is determined by the standard deviation of the internal processing error, can be measured psycho-acoustically in an IPD detection task to derive an individual standard deviation and, therefore, to individualize the

low-pass filter. A possible experimental setup is for example described in Ross *et al.* (2007) and Santurette and Dau (2012), where the ability to detect IPDs as a function of the center frequency of sinusoidally amplitude modulated (SAM) pure tones is investigated. For example, the highest frequency up to which a person can detect interaural changes in phase can be determined and, afterwards, transformed into an individual standard deviation σ_{δ_0} via

$$\sigma_{\delta_0} = \frac{\sqrt{-\ln(10^{\frac{-3}{10}})}}{2\pi \mathrm{f}_{\mathrm{cut}}}, \tag{20}$$

where $\mathrm{f}_{\mathrm{cut}}$ is the 3 dB cut-off frequency in Hz (Warzybok *et al.*, 2014).
For example, a cutoff frequency of 2035 Hz (see Figure 5) corresponds to $\sigma_{\delta_0} = 65\mu$s. If the cutoff frequency is decreased to 1200 Hz, the underlying Gaussian distribution gets broader and the standard deviation is increased to $\sigma_{\delta_0} = 110\mu$s.

1.6 Speech Intelligibility Index (SII)

The Speech Intelligibility Index (ANSI, 1997) is a macroscopic model to predict speech intelligibility based on speech and interfering noise. In contrast to microscopic models, macroscopic models do not to mimic auditory processes in their signal processing. In the literature, there exist other macroscopic models, for example the Articulation Index (AI) (French and Steinberg, 1947; Fletcher and Galt, 1950), which is the predecessor of the SII , or the Speech Transmission Index (STI) (Steeneken and Houtgast, 1980). While the SII and AI are conceptually similar and predict speech intelligibility from the frequency spectra of speech and noise, the STI also takes the degrading effect of reverberation into account by analyzing the modulation transfer function of speech, which is affected by both noise and reverberation. The SII is also available as short-time version and was named extended SII (ESII) (Rhebergen and Versfeld, 2005). The ESII is able to account for lower thresholds in fluctuating noise (listening in the dips). Differing from the SII, the ESII also takes forward masking into account.
The basic concept of the SII is the analysis of frequency dependent SNRs, which are weighted according to the human speech perception. The band wise evaluation of SNR values in the SII is similar to the SNR processing of the `BSIM2010` front-end, which allows for an easy combination of both EC front-end and SII back-end. In a next step, the band wise SNRs are transformed to an index. The resulting index, which is in the range between 0 and 1, is highly correlated with the intelligibility of speech in stationary noise scenarios and can be used as predictor of speech intelligibility.
In the SII, several factors influencing speech intelligibility are incorporated. For example, the hearing threshold as well as spread of masking and distortions due to high levels are considered. In the following, the SII and its processing steps are

described. A few stages are changed due to the processing of `BSIM2010`, which will be mentioned, too.

First, speech S and noise N are divided into several frequency bands. The number of bands and the bandwidth of each filter is determined by the chosen filterbank paradigm. In ANSI (1997), four alternative filterbanks are defined: 1) critical frequency bands (21 bands), 2) one-third octave frequency bands (18 bands), 3) equally contributing critical bands (17 bands), and 4) octave frequency band (6 bands). In Beutelmann *et al.* (2010), a fifth option was added, where a gammatone filterbank with 30 frequency bands is used to be consistent to the prior EC processing. The gammatone filterbank paradigm is also used in this study.
In the next step, the equivalent speech spectrum level E is calculated in each frequency band i by summing up the speech spectrum level E and the insertion gain G, giving

$$E_i = E_i + G_i. \tag{21}$$

The insertion gain is used, when a person wears a hearing aid or a device, which attenuates or amplifies spectral components. It is applied to both speech and noise. Unless there is no hearing device used to predict speech intelligibility, the insertion gain is set to 0 in `BSIM2010`.
Afterwards, the self-speech masking level (V) is calculated by applying

$$V_i = E_i - 24, \tag{22}$$

where 24 dB are subtracted from the equivalent speech spectrum level E, which mirrors the masking of high frequency speech components by lower frequency components. The value of 24 dB was determined experimentally. The self-speech masking level V is then compared to the equivalent noise spectrum level N and the maximum

$$B_i = max(V_i, N_i) \tag{23}$$

of both is considered for further calculations, i.e. the variable providing a larger amount of masking is used.
In order to take the spread of masking into account a slope parameter C, which is the slope of the masking spectrum per octave, is calculated according to

$$C_i = -80 + 0.6 \cdot [B_i + 10\log_{10}(h_i - l_i)], \tag{24}$$

if one of the critical band methods is used. h_i and l_i denote the higher and the lower frequency band limit of the critical band i. If the (third-)octave scheme is used, C is calculated as

$$C_i = -80 + 0.6 \cdot [B_i + 10\log_{10}F_i - 6.353]. \tag{25}$$

F_i denotes the center frequency of band i. The parameter C is used to calculate the equivalent masking spectrum level Z at a later step.
One of the special features of the gammatone filterbank is its inherent consideration of spread of excitation. It enables spread of masking as well as the phenomenon of off-frequency listening. Therefore, no calculation of a sloping parameter is needed when using a gammatone filterbank. Therefore, no sloping parameter C is calculated in `BSIM2010`.
The equivalent internal noise spectrum level X' is computed, which is the reference internal noise spectrum level X, which depends on the underlying filterbank and is given in tabular form, and the individual hearing threshold T in dB HL, according to

$$X'_i = X_i + T_i. \tag{26}$$

Then the equivalent masking spectrum level Z is computed to combine the effects of spread of masking, noise, and self-speech masking in a single variable. For the lowest frequency band, Z equals the temporal variable B, which was calculated in Equation 23, because no upward spread of masking can occur. For the remaining frequency bands, the equivalent masking spectrum level is

$$Z_i = 10lg\left\{10^{0.1N_i} + \sum_{k}^{i-1} 10^{0.1[B_k + 3.32C_k lg(F_i/h_k)]}\right\} \tag{27}$$

if one of the critical band methods is used, or

$$Z_i = 10lg\left\{10^{0.1N_i} + \sum_{k}^{i-1} 10^{0.1[B_k + 3.32C_k lg(0.89F_i/F_k)]}\right\} \tag{28}$$

if the third-octave or octave band procedure is applied. The equivalent masking spectrum level Z is then compared to the equivalent internal noise spectrum level X'. By taking the maximum of both, the disturbance spectrum level D is calculated as

$$D_i = max(Z_i, X'_i). \tag{29}$$

The influence of the level distortion L on speech intelligibility is accounted for by comparing the equivalent speech spectrum level E to the standard speech spectrum level U at normal vocal effort by applying

$$L_i = 1 - (E'_i - U_i - 10)/160. \tag{30}$$

U is given in tabular form and the upper limit of the level distortion L is 1.
The temporary variable K of the band audibility function is computed according to

$$K = (E - D + 15)/30, \tag{31}$$

and restricted to be between 0 and 1. Afterwards, the temporal variable K is weighted with the level distortion factor L to get the band audibility

$$A = L \cdot K. \tag{32}$$

The resulting SII is the weighted w sum of the frequency specific audibility A according to

$$SII = \sum_{i=1}^{N} w_i A_i. \tag{33}$$

Frequency bands, which are assumed to contribute more to speech intelligibility get higher weights w than the remaining frequency bands. The weights, which are applied in the gammatone filterbank procedure are shown in Figure 6. The

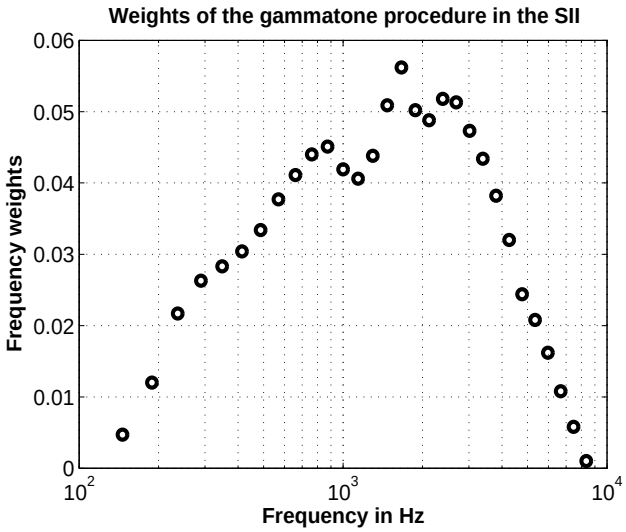

Figure 6: *Frequency weights, which are applied in the gammatone paradigm of the SII. The sum over all weights is 1.*

gammatone frequency weights are obtained by adjusting the critical band weights (ANSI, 1997, Table 1) to the new center frequencies. If, for example, the equally contributing paradigm is applied, all frequency bands receive the same weight as it is assumed that all frequency bands contribute to speech intelligibility, equally. In this case, the calculation of the SII can be simplified to

$$SII = 0.0588 \sum_{i=1}^{17} A_i \tag{34}$$

(ANSI, 1997, Table 2). In this study, the gammatone paradigm is used in the SII. The resulting SII is a number between 0 and 1, where 1 denotes a high intelligibility and 0 denotes "not intelligible". In this study, the SII is transformed to the Speech Reception Threshold, where 50% of the speech is correctly understood. The corresponding SII value, which is adjusted until the prediction equals the observation,

is 0.265 in case the SRT of 50% speech intelligibility obtained using the Oldenburg sentence test in noise (OLSA) (Wagener *et al.*, 1999a,b,c) is used as reference. This transformation can also be used to describe context related features of the speech material, because the same SII value can predict another SRT if another speech material is used. The SII only describes that the same amount of information is audible.

1.7 Short-time BSIM (stBSIM2010)

The model, which has been described above, can be used in two different versions: 1) the long-time version `BSIM2010` and 2) the short-time version `stBSIM2010`.
`stBSIM2010` was also introduced in Beutelmann *et al.* (2010). The concept was developed to account for time varying maskers, for example amplitude modulated signals. In principle, both model versions do the same processing but on different time scales. The `BSIM2010` analyzes the whole input signal to estimate the EC parameters and applies the same set of parameters to the whole time signal. However, this approach will run into trouble if the masker characteristics, for example the level, change over time. In this situation, the `stBSIM2010` has the advantage that it works on short time segments of 23 ms, such that it can adapt to changes in the stimulus. For each 23 ms segment of the signal, the EC parameters are estimated to optimize the SNR and an SRT is calculated, resulting in a vector of SRT values. In a last step, the SRT values are averaged over time to get a final SRT value for the whole signal. This procedure is deviating from the paradigm applied in the extended SII (ESII) (Rhebergen and Versfeld, 2005). In the ESII, SII values are calculated for the short-time segments and then the average over time is calculated.
In the future, the averaging of SRT values in the `stBSIM2010` should be revised. Calculating SRT values is a non-linear operation and the predictions might deviate if, for example, SNR values are averaged over time and the result is then transformed to a SRT value.

References

ANSI (**1997**), "Methods for the calculation of the speech intelligibility index," American National Standard S3.5-1997 (Standards Secreteriat, Acoustical Society of America) .

Beutelmann, R. and Brand, T. (**2006**), "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," The Journal of the Acoustical Society of America **120**(1), 331–342.

Beutelmann, R., Brand, T., and Kollmeier, B. (**2009**), "Prediction of binaural speech intelligibility with frequency-dependent interaural phase differences," The Journal of the Acoustical Society of America **126**(3), 1359–1368.

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

Bronkhorst, A. W. and Plomp, R. (**1988**), "The effect of head-induced interaural time and level differences on speech intelligibility in noise," The Journal of the Acoustical Society of America **83**(4), 1508–1516.

Cosentino, S., Marquardt, T., McAlpine, D., Culling, J. F., and Falk, T. H. (**2014**), "A model that predicts the binaural advantage to speech intelligibility from the mixed target and interferer signals," The Journal of the Acoustical Society of America **135**(2), 796–807.

Durlach, N. I. (**1963**), "Equalization and Cancellation Theory of Binaural Masking Level Differences," The Journal of the Acoustical Society of America **35**(8), 1206–1218.

Fletcher, H. and Galt, R. H. (**1950**), "The perception of speech and its relation to telephony," The Journal of the Acoustical Society of America **22**(2), 89–151.

French, N. and Steinberg, J. (**1947**), "Factors governing the intelligibility of speech sounds," The journal of the Acoustical society of America **19**(1), 90–119.

Glasberg, B. R. and Moore, B. C. (**1990**), "Derivation of auditory filter shapes from notched-noise data," Hearing Research **47**, 103 – 138.

Hohmann, V. (**2002**), "Frequency Analysis and Synthesis using a Gammatone filterbank," Acust. Acta Aust **88**, 433–442.

Lavandier, M. and Culling, J. F. (**2010**), "Prediction of binaural speech intelligibility against noise in rooms," The Journal of the Acoustical Society of America **127**(1), 387–399.

Palmer, A. and Russell, I. (**1986**), "Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells," Hearing research **24**(1), 1–15.

Rennies, J., Brand, T., and Kollmeier, B. (**2011**), "Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quieta)," The Journal of the Acoustical Society of America **130**(5), 2999–3012.

Rennies, J., Warzybok, A., Brand, T., and Kollmeier, B. (**2014**), "Modeling the effects of a single reflection on binaural speech intelligibility," The Journal of the Acoustical Society of America **135**(3), 1556–1567.

Rhebergen, K. S. and Versfeld, N. J. (**2005**), "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 **117**(4), 2181–2192.

Ross, B., Tremblay, K. L., and Picton, T. W. (**2007**), "Physiological detection of interaural phase differences," The Journal of the Acoustical Society of America **121**(2), 1017–1027.

Santurette, S. and Dau, T. (**2012**), "Relating binaural pitch perception to the individual listeners auditory profile," The Journal of the Acoustical Society of America **131**(4), 2968–2986.

Steeneken, H. J. M. and Houtgast, T. (**1980**), "A physical method for measuring speech-transmission quality," The Journal of the Acoustical Society of America **67**(1), 318–326.

vom Hövel, H. (**1984**), *Zur Bedeutung der Übertragungseigenschaften des Aussenohrs sowie des binauralen Hörsystems bei gestörter Sprachübertragung* (na).

Wagener, K., Brand, T., Kühnel, V., and Kollmeier, B. (**1999a**), "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. **38**, 4–15.

Wagener, K., Brand, T., Kühnel, V., and Kollmeier, B. (**1999b**), "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. **38**, 44–56.

Wagener, K., Brand, T., Kühnel, V., and Kollmeier, B. (**1999c**), "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. **38**, 86–95.

Wan, R., Durlach, N. I., and Colburn, H. S. (**2010**), "Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers," The Journal of the Acoustical Society of America **128**(6), 3678–3690.

Warzybok, A., Rennies, J., Brand, T., and Kollmeier, B. (**2014**), "Prediction of binaural speech intelligibility in normal-hearing and hearing-impaired listeners: a psychoacoustically motivated extension," DAGA 2014, Oldenburg, Germany .