Original Article

Modeling Sluggishness in Binaural Unmasking of Speech for Maskers With Time-Varying Interaural Phase Differences

Trends in Hearing
Volume 22: 1–10

DOI: 10.1177/2331216517753547
journals.sagepub.com/home/tia
SAGE

Christopher F. Hauth[1,2] and Thomas Brand[1,2]

Abstract

In studies investigating binaural processing in human listeners, relatively long and task-dependent time constants of a binaural window ranging from 10 ms to 250 ms have been observed. Such time constants are often thought to reflect "binaural sluggishness." In this study, the effect of binaural sluggishness on binaural unmasking of speech in stationary speech-shaped noise is investigated in 10 listeners with normal hearing. In order to design a masking signal with temporally varying binaural cues, the interaural phase difference of the noise was modulated sinusoidally with frequencies ranging from 0.25 Hz to 64 Hz. The lowest, that is the best, speech reception thresholds (SRTs) were observed for the lowest modulation frequency. SRTs increased with increasing modulation frequency up to 4 Hz. For higher modulation frequencies, SRTs remained constant in the range of 1 dB to 1.5 dB below the SRT determined in the diotic situation. The outcome of the experiment was simulated using a short-term binaural speech intelligibility model, which combines an equalization–cancellation (EC) model with the speech intelligibility index. This model segments the incoming signal into 23.2-ms time frames in order to predict release from masking in modulated noises. In order to predict the results from this study, the model required a further time constant applied to the EC mechanism representing binaural sluggishness. The best agreement with perceptual data was achieved using a temporal window of 200 ms in the EC mechanism.

Keywords

speech reception thresholds, binaural, auditory model, interaural phase difference, binaural sluggishness

Date received: 11 July 2017; accepted: 12 December 2017

Introduction

In everyday life, human listeners have to deal with complex acoustic scenarios, in which different kinds of interfering noise sources arise at different locations. In the literature, this is termed the "cocktail party problem" (Cherry, 1953, p. 76). As the noise sources are often spatially separated from the target speech, it is beneficial to have access to binaural information, such as interaural level differences (ILDs), interaural time differences (ITDs), or interaural phase differences (IPDs; Bronkhorst, 2000). For binaural speech intelligibility in adverse acoustic conditions, two mechanisms are thought to play a primary role: better-ear listening, which is listening using the ear that has the better signal-to-noise ratio (SNR), and binaural unmasking, where ITD and ILD processing enables segregation of the target from the interfering signal. In scenarios with time-varying binaural parameters, binaural temporal windows have been derived, which are task dependent and range from approximately 40 ms to 250 ms (Akeroyd & Summerfield, 1999; Culling & Summerfield, 1998; Grantham & Whightman, 1979; Holube, Kinkel, & Kollmeier, 1998). However, in some tasks the binaural system also seems to process changes in the interaural parameters on a much shorter time scale, which is in the range of 10 ms (Akeroyd and Bernstein, 2001; Bernstein, Trahiotis, Akeroyd, & Hartung, 2001). In contrast, time constants that are usually obtained in monaural psychoacoustic experiments are in the range of 4 ms to 26 ms (Holube

[1]Medizinische Physik, Carl von Ossietzky Universität, Oldenburg, Germany
[2]Cluster of Excellence Hearing4All, Carl von Ossietzky Universität, Oldenburg, Germany

Corresponding author:
Christopher F. Hauth, Medizinische Physik, Carl von Ossietzky Universität Oldenburg, Carl-von-Ossietzky-Straße 9-11, Oldenburg 26129, Germany.
Email: christopher.hauth@uni-oldenburg.de

et al., 1998). In the literature, the phenomenon of relatively long time constants involved in many binaural psychoacoustic tasks is often referred to as *binaural sluggishness*.

This study investigates the effect of ''binaural sluggishness'' on speech reception thresholds (SRTs) in stationary speech-shaped noise, where the IPD of the noise is sinusoidally modulated over time. As the binaural system reacts sluggishly to fast changes in the interaural parameters in many conditions, speech intelligibility can be expected to be affected by rapid changes in the location or IPD of the masker (e.g., Culling & Summerfield, 1998).

It was hypothesized that slow changes in the interaural phase lead to better performance in speech intelligibility, that is lower SRTs, because speech and noise can be more easily perceptually segregated and an effective binaural unmasking can be achieved by processing the IPD of the noise. If the IPD changes more rapidly, performance was expected to decline because the binaural auditory system is not able to follow fast changes in the interaural configuration of the masker and thus cannot make use of interaural disparities for achieving an effective binaural unmasking.

One mechanism that can predict quantitatively the binaural unmasking resulting from differing ITDs or ILDs of the interfering signal and the target signal, is the equalization–cancellation (EC) mechanism (Durlach, 1963). In the EC mechanism, the ILD and ITD or IPD of an interfering noise are estimated and compensated such that the interfering noise is interaurally aligned in level and phase. Later, the adjusted left- and right-ear signals are subtracted from each other, leading to an attenuation of the noise via destructive interference or to an amplification of the target signal via constructive interference. Therefore, the signal-to-noise ratio (SNR) is improved as long as the target and interferer differ in their ILDs and ITDs/IPDs.

The EC mechanism has been implemented in binaural speech intelligibility models, where it was combined with the speech intelligibility index (SII; ANSI, 1997; Beutelmann and Brand, 2006; Beutelmann, Brand, & Kollmeier, 2010; Wan, Durlach, & Colburn, 2014). A conceptually similar model has been proposed by Lavandier and Culling (2010) and Lavandier et al. (2012), where the band importance function of the SII is used to integrate SNRs across frequency. More recently, binaural speech intelligibility models have been proposed using alternative back ends to the SII. Chabot-Leclerc, MacDonald, and Dau (2016) combined the EC mechanism with the multiresolution speech-based envelope power spectrum model (mr-sEPSM; Jørgensen et al., 2013), which analyzes SNRs in the envelope domain in order to predict speech intelligibility. Furthermore, Andersen et al. (2016) combined the EC mechanism with the short-time objective intelligibility (STOI) measure (Taal, Hendriks, Heusdens, & Jensen, 2011), where the correlation between noisy speech and a clean speech reference is analyzed.

These binaural speech intelligibility models are based on assumptions about the EC process made by Durlach (1963), according to which an equalization step has inherent processing errors in level and time, which lead to an imperfect alignment of the left- and right-ear signals and, consequently, to an imperfect cancellation of the masker signal. These processing errors were modified by Vom Hövel (1984) to better agree with data of Langford and Jeffress (1964) and Egan (1965). Beutelmann and Brand (2006) and Beutelmann et al. (2010) incorporated these processing errors into their binaural speech intelligibility model, which is referred to here as *BSIM2010*. The short-term version of BSIM2010 (Beutelmann et al., 2010) operates on time frames of 23.2 ms (1,024 samples at a sampling rate of 44,100 Hz with 50% overlap, leading to an effective time window of 11.6 ms) to process envelope-modulated interferers and to account for the phenomenon of ''listening in the dips,'' which refers to the observation that SRTs are lower in amplitude-modulated noise. This time frame is used in the EC mechanism and in the SII and is constant across all frequency channels, which is comparable with the best fitting frequency-independent time frame of 12 ms in the extended speech intelligibility index (ESII) (Rhebergen and Versfeld, 2005). Short-term EC processing is also performed in the models developed by Wan et al. (2014), Andersen et al. (2016), and Chabot-Leclerc et al. (2016).

In Wan et al. (2014) and Chabot-Leclerc et al. (2016), the EC mechanism is applied in overlapping 20-ms time frames (10-ms overlap), which is similar to the realization of short-term processing used in Beutelmann et al. (2010). In Andersen et al. (2016), time frames of 25.6 ms are used in the EC mechanism. As the STOI is used as back end, the short-time processed segments are averaged using a time constant of 386 ms. Therefore, the short-term BSIM2010 and the models from Wan et al. (2014) and Chabot-Leclerc et al. (2016) are not expected to be able to account for binaural sluggishness as they do not effectively differ in their binaural processing stages.

In BSIM2010, the time frame used by the SII is also used in the EC mechanism. However, previous studies suggest that 23.2 ms may be too short (Culling & Summerfield, 1998; Holube et al., 1998). Therefore, in this study, the temporal processing of the EC mechanism and of the short-term SII is decoupled to enable the EC mechanism to operate on a different time scale than the short-term SII. Binaural sluggishness is introduced in the model's front end in order to account for the ability of the auditory system to cope with time-varying IPDs as long as the changes are not too fast. A binaural time constant related to processing the IPD-modulated interferer in this

speech-in-noise task was estimated based on simulated SRTs obtained with the BSIM2010 in its short-term version with and without an extension for binaural sluggishness. Different time constants were tested in the BSIM framework in order to find the time constant yielding the best agreement with the perceptual data.

Method

Listeners

A total of 10 listeners with normal hearing (five men and five women) participated in this study and were paid for their effort. Their ages ranged from 23 to 26 years (mean age 24.8 years). The audiometric thresholds did not exceed 20 dB HL, except for one frequency in one listener, where 25 dB HL were measured at 1,500 Hz. All listeners had previous experience with sentence test procedures and psychoacoustic measurements.

Apparatus and Procedure

Speech intelligibility experiments were conducted using the Oldenburg Sentence Test (OLSA, Wagener, Brand, Kühnel, & Kollmeier, 1999a, 1999b, 1999c) in noise. For determining the SRT (the SNR where 50% of the words are understood), an adaptive procedure was used for controlling the level of the speech (Equation 9, Brand & Kollmeier, 2002).

Measurements were conducted using closed-set sentences (see subsequent text), where all test items of the OLSA were presented visually in a matrix, and the listener then marked those items they understood using a graphical user interface. The listeners were allowed to guess and no feedback was provided. The probability of guessing a single item correctly is 10% and the probability of guessing a whole sentence correctly is 0.001%. The stimuli were generated using MATLAB (MathWorks, Natick, MA, USA) using the AFC Toolbox (Version 1.4), developed by Stephan Ewert at Carl von Ossietzky University, Oldenburg, Germany, and presented binaurally via an RME Fireface UC soundcard (Audio AG, Haimhausen, Germany) and HD 650 headphones (Sennheiser, Wedemark, Germany).

The SRT was determined using test lists of 20 sentences. The test lists were randomly selected out of 45 lists. The order of stimuli was Latin-square balanced to avoid effects of the order of presentation. All listeners were trained using four lists to get used to the phase manipulation prior to testing. Two lists were presented at a fixed SNR of −2 dB and two lists used the adaptive procedure. Overall, two sessions (test and retest) were conducted, each lasting 1.5 hr including breaks.

Before conducting the experiments, the equipment was calibrated to dB SPL using a Brüel&Kjaer (B&K, Nærum, Denmark) 4153 artificial ear, a B&K 4134 half-inch microphone, a B&K 2669 preamplifier, and a B&K 2610 measuring amplifier. The noise level was set to 65 dB SPL and the speech level was varied to find the individual SRT. The experiments were conducted in a double-walled, sound-attenuated booth.

Stimuli

Each sentence of the OLSA is composed of five words, which are arranged in a fixed syntactical structure: noun, verb, numeral, adjective, and object. Each item is sampled from a list of 10 words to create a low-context sentence, such as, ''Peter kauft achtzehn nasse Dosen'' (Peter buys 18 wet cans). The speech material is always presented diotically to the listeners, creating the perception that the speech source is located in the center of the head. The interfering noise was generated by randomly superimposing the speech material, creating a stationary noise with the same long-term spectrum as the speech material (Wagener et al., 1999a).

To test the effect of binaural sluggishness on speech intelligibility in noise, the IPD of the interfering noise was modulated sinusoidally, which at slow rates creates the perception of noise oscillating between the left and right ears. The interaural phase modulation frequency (IPMF) was varied from 0 Hz to 64 Hz in the following steps: 0, 0.25, 0.5, 1, 2, 4, 8, 16, 32, and 64 Hz, resulting in 10 conditions. An IPMF of 0 Hz denotes the static condition with a constant IPD of 0°, that is, the diotic presentation of speech and noise. For the other modulation frequencies, the IPD were varied between $-\pi/2$ and $+\pi/2$ and the initial phase was set randomly. The phase manipulation was applied to the noise in the short-time Fourier transform (STFT) domain. The noise was segmented into frames of 5 ms with 50% overlap and windowed with a square root Hann window. After computing the fast Fourier transform (FFT), the complex spectral representation of the noise was separated into magnitude and phase. Later, the interaural phase shift was applied to the phase and the complex spectrum was resynthesized. Before reconstructing the time signal, the time frames were again windowed with a square root Hann window. The amplitude of the phase manipulation was divided by a factor of 2 and applied symmetrically to the left- and right-ear channels in order to minimize monaural phase distortions. The noise started 500 ms before the presentation of the speech and was terminated 500 ms after each sentence.

An IPD manipulation was preferred over ITD manipulation, because a high time resolution was required, leading to a poor frequency resolution. As an ITD leads to a frequency-dependent IPD, the required frequency resolution would have been too high. By applying a constant IPD offset, the frequency resolution

was decreased and a time resolution of 2.5 ms was obtained.

Extension of BSIM for sluggish EC processing

In order to investigate whether a second time constant is required in the EC mechanism for dynamic binaural cues, the EC stage of BSIM2010 was modified to account for binaural sluggishness. Figure 1 shows the scheme of the processing performed in BSIM2010. In Beutelmann et al. (2010), a long-term and a short-term version of BSIM2010 were introduced, where the short-term version performs the EC and SII calculations in time frames of $\tau_{SII} = 23.2$ ms. The long-term version of BSIM2010 performs the EC and SII calculations by considering the whole signal as a single time frame and shows good results in spatially and temporally stationary maskers (e.g., Beutelmann et al., 2010). The short-term version is modified in this study. In both the long-term and short-term versions, the following calculations are performed: First, left- and right-ear signals are processed by a peripheral filtering stage, using a gammatone (GT) filterbank (Hohmann, 2002) ranging from 146 Hz to 8,346 Hz in 30 ERB-spaced (Glasberg and Moore, 1990) frequency bands. In the EC mechanism, the ILD (α) and the ITD (τ) are estimated in each frequency channel independently. In the modified short-term version of BSIM2010, the estimation of the EC parameters is still performed in time frames of 23 ms, but the estimates are concatenated over a binaural temporal window, whose length is defined by the time constant $\tau_{sluggish}$. The median is calculated over the binaural temporal window and used in the EC process. This approach gave results slightly closer to the observed values than using the mean value. Furthermore, the median value was chosen in order to be more robust against outliers in the estimation of the interaural differences. Using the mean value would implicitly assume a normal distribution of ITD and ILD values. Consequently, the median was preferred to be independent of the underlying distribution of ITD and ILD values.

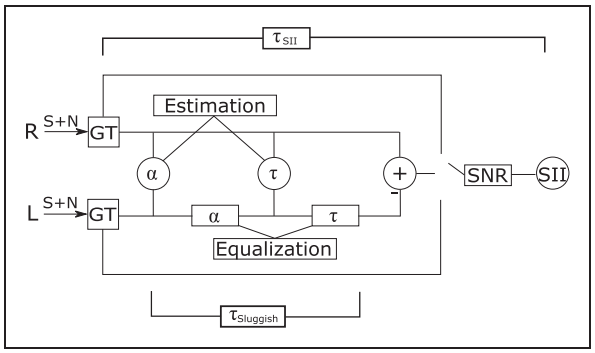

Figure 1. Scheme of BSIM2010: The left- and right-ear signals are first band pass filtered using a gammatone filterbank, denoted as GT. In each filter, the equalization parameters are estimated and concatenated over a temporal window defined by $\tau_{Sluggish}$. The median over this window is applied in the equalization step of the EC mechanism. Later, the signal providing the best SNR is selected in each frequency band and time frame (τ_{SII}), which can be either the left-/right-ear channel or the output of the EC model. The SNRs are transformed to SRTs using the SII in each time frame. The overall SRT is obtained by averaging the SRTs over all time frames.

Different binaural time constants $\tau_{sluggish}$ were tested: 23.2 ms (i.e., the SII time frame limits the resolution of the binaural processing stage, which is equivalent to the short-term version of BSIM2010) and, in addition, 50, 75, 100, 125, 150, and 200 ms. For further processing, the median values of the estimated EC parameters across the binaural temporal window were used for EC processing in the actual time frame. The realization of two separate time constants has the advantage of introducing a sluggish component in the binaural processing stage while maintaining the model's ability to cope with fast temporal envelope fluctuations of the masker. The long-term BSIM2010 was also used to predict the outcome of the experiment in order to investigate whether or not it can make use of long-term IPD information even though the IPD fluctuates over time. Conceptually, this corresponds to a long-term beamformer steered toward the direction yielding the largest long-term SNR improvement.

Results

Figure 2 depicts the results obtained in the speech intelligibility experiment, in which the IPDs of the interfering noise were varied temporally. Boxplots of the SRTs are shown as a function of the IPD modulation frequency. An IPMF of 0 Hz is equivalent to the diotic or N_0S_0 presentation of speech and noise. In the diotic condition, the median SRT was obtained at −9.2 dB SNR; this was decreased to −12.3 dB SNR for the lowest IPMF, at 0.25 Hz. With increasing IPMF, the SRTs gradually increased, showing a ceiling effect above an IPMF of 4 Hz, where the median SRT was −10.3 dB SNR.

Figure 3 shows the corresponding binaural intelligibility level difference (BILD), which is the difference between the conditions with time-varying IPDs and the diotic condition. The largest release from masking, in the range of 2 dB to 3 dB, was obtained for the lowest IPMF, at 0.25 Hz. However, the standard deviation was also largest in this condition. This was caused by the randomly selected initial phase of the time-varying IPDs. In this condition, the period of the sinusoidal interaural phase modulation is 4 s, while the length of an OLSA sentence is approximately 2 s. Therefore, the outcome in this condition is strongly affected by the initial

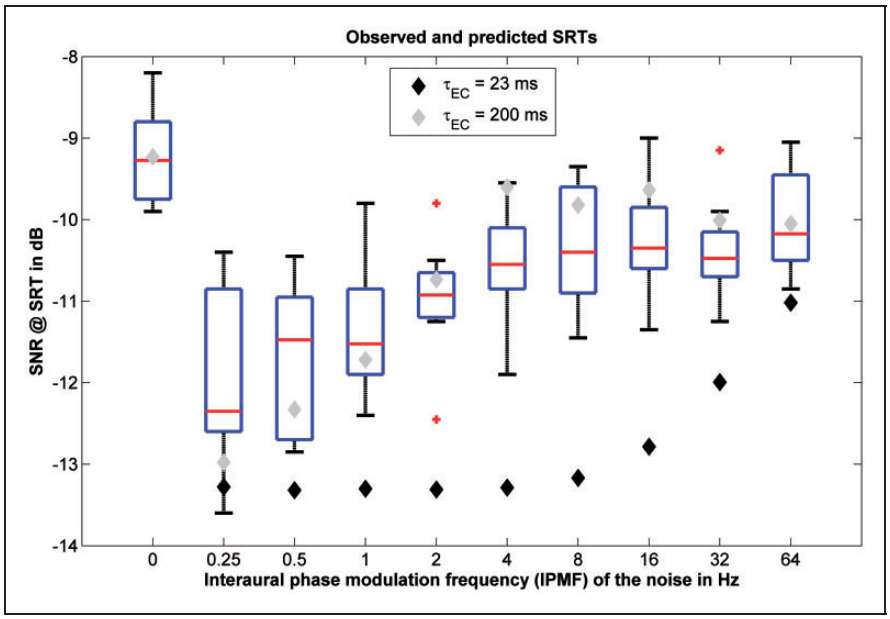

Figure 2. Predicted and observed SRTs (dB SNR) are shown as a function of IPMF (in hertz). The largest release from masking as well as the largest spread in the data can be observed for the lowest IPMF (0.25 Hz). With increasing IPMF, the SRTs also increase (get worse) up to an IPMF of 4 Hz. Above 4 Hz, the SRTs remain constant. An IPMF of 0 Hz denotes the diotic or N0S0 condition. Predictions are obtained using the short-term BSIM2010 and the sluggish short-term BSIM2010 with a binaural time constant of 200 ms in the EC mechanism. Note that the predictions obtained for an IPMF of 0 Hz overlap and thus only the marker for the EC time constant of 200 ms is visible.

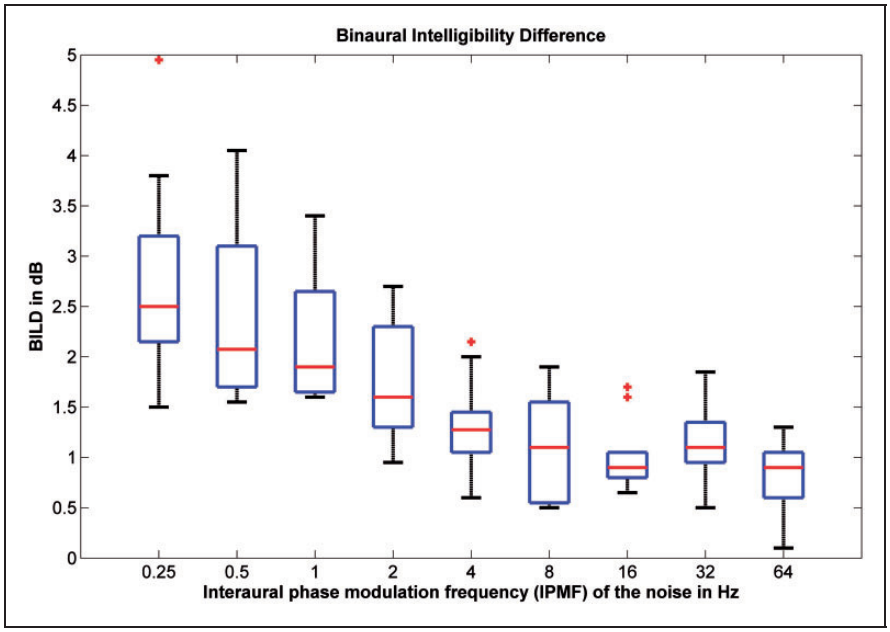

Figure 3. The BILD (in decibel) is plotted as a function of IPMF (in hertz). A decreasing BILD with increasing IPMF up to 4 Hz and a constant BILD of 1 dB to1.5 dB above 4 Hz can be observed.

phase. For IPMFs larger than 0.25 Hz, the BILD gradually decreased with increasing IPMF, up to an IPMF of 4 Hz. For modulation frequencies higher than 4 Hz, the release from masking remained constant between 1 dB and 1.5 dB.

For statistical evaluation of the results, a Shapiro–Wilk test ($\alpha = 0.05$) was conducted, which revealed normally distributed results for all tested conditions. A t test was used to investigate the effect of IPMF on SRT. Because a multiple pairwise comparison was used, a

Bonferroni correction was applied in order to avoid spurious positives. The t test ($\alpha = 0.01$) revealed a highly significant effect of IPMF in all conditions, compared with the diotic situation. Furthermore, a significant difference in SRT was observed for IPMFs that differed by two octaves up to an IPMF of 4 Hz. For IPMFs higher than 4 Hz, no statistical difference was observed between the tested IPMFs.

Figure 2 also shows the predicted SRTs using short-term BSIM2010 (black diamonds). The SII was calibrated to fit the median SRT in the 0 Hz condition, that is, to speech in stationary speech-shaped noise with no IPDs. The remaining IPMFs were then predicted using the same SII value (SII = 0.201). The calibration was not changed for the other tested EC time constants, as the effect of the additional time constant in the EC mechanism was expected to have no effect on the predicted SRT in the 0 Hz condition and the results also showed no difference in the diotic condition. In general, up to an IPMF of 16 Hz, the predicted SRTs using the short-term BSIM2010 are consistently reduced by 4 dB compared with the diotic condition. Above 16 Hz, the predicted SRTs are increased and approach the obtained data; this was caused by the relationship between the SII time frame and the period of the sinusoidal phase modulation: At 32 Hz, the period of the IPMF is 31.25 ms, so almost a whole cycle fits in the SII time frame. Therefore, no effective improvement of the SNR over the whole frame can be achieved by applying a constant delay in the EC process. The short-term version of BSIM2010, which, in principle, has been shown to be able to account for the phenomenon of listening in the dips (Beutelmann et al., 2010), is not optimal for predicting the effect of fast-changing IPD information.

In Table 1, the performance of the short-term BSIM2010 is evaluated using R^2, the root mean square error (RMSE) in decibels and the bias in decibels between predicted and the obtained mean SRTs as a measure, where R^2 is the coefficient of determination, which describes the amount of variation that can be explained by the binaural model, and the bias is the prediction error averaged across the tested IPMFs, which is the signed difference of predicted and measured SRTs. Predictions using the short-term BSIM2010 without sluggishness (one common time constant for binaural processing and SII) resulted in a relative low prediction accuracy ($R^2 = 0.37$, RMSE = 2.1 dB, bias = −1.98 dB). Introducing a second time constant for the EC mechanism led to an approximation of the predictions to the perceptual data. By increasing the time constant in the EC mechanism, predicted SRTs tended to increase for lower IPMFs, as seen in the listening data. Figure 2 also shows the predicted SRTs obtained using the refined version of the short-term BSIM2010 with a time constant of $\tau_{sluggish} = 200$ ms in the EC mechanism. This time constant showed best agreement with the perceptual data in terms of RMSE and bias ($R^2 = 0.92$, RMSE = 0.6 dB, bias = −0.1 dB).

Table 1. Accuracy of the Predicted SRTs Depending on the Length of the Binaural Temporal Window Applied in the Short-Term BSIM2010 in Terms of R^2, RMSE, and Bias.

Length of the binaural window $\tau_{Sluggish}$ in milliseconds	R^2	RMSE in decibels	Bias in decibels
23.2	.38	2.1	−1.98
50	.72	1.44	−1.28
75	.88	1.07	−0.83
100	.93	0.81	−0.45
125	.93	0.77	−0.23
150	.95	0.63	−0.18
200	.92	0.6	−0.08
Long term	.52	0.98	0.87

Note. The result obtained with the long-term BSIM2010 is denoted as "Long term." RMSE = root mean square error; SRT = speech reception threshold.

Figure 4 shows predicted SRTs against measured SRTs for all tested time constants. With increasing the time constant up to 100 ms, the predictions approach the perceptual data. By further increasing the time constant, the predictions are only slightly improved. RMSE and bias are within the standard deviation of the Oldenburg sentence test procedure, which is 1 dB, if the binaural time window is set to $\tau_{EC} = 100$ ms or higher. However, 200 ms is suggested as best-fitting binaural time constant because it produces the lowest RMSE and bias. The effect using an even longer time constant was investigated by modeling the outcome of the experiment with the long-term version of BSIM2010, where the whole incoming signal is considered as a single time frame. This model was not able to account for the data as binaural unmasking was only predicted for the slowest modulation of 0.25 Hz and to some extent for 0.5 Hz, which was moreover strongly dependent on the initial IPD of the masking noise ($R^2 = 0.52$, RMSE = 0.98 dB, bias = 0.87 dB). Using the long-term version, a binaural intelligibility difference ranging from 0 dB to 4 dB was predicted for an IPMF of 0.25 Hz. Summarizing the predictions with different time constants simulating binaural sluggishness, a time constant in the range from 150 ms to 200 ms should be considered to account for binaural sluggishness in this experiment.

The backward compatibility of the sluggish short-term BSIM2010 was evaluated by simulating an experiment from Beutelmann and Brand (2006), where speech intelligibility in stationary speech-shaped noise and different acoustic scenarios was measured. Simulations were conducted using the long-term BSIM2010, short-term BSIM2010, and the sluggish short-term

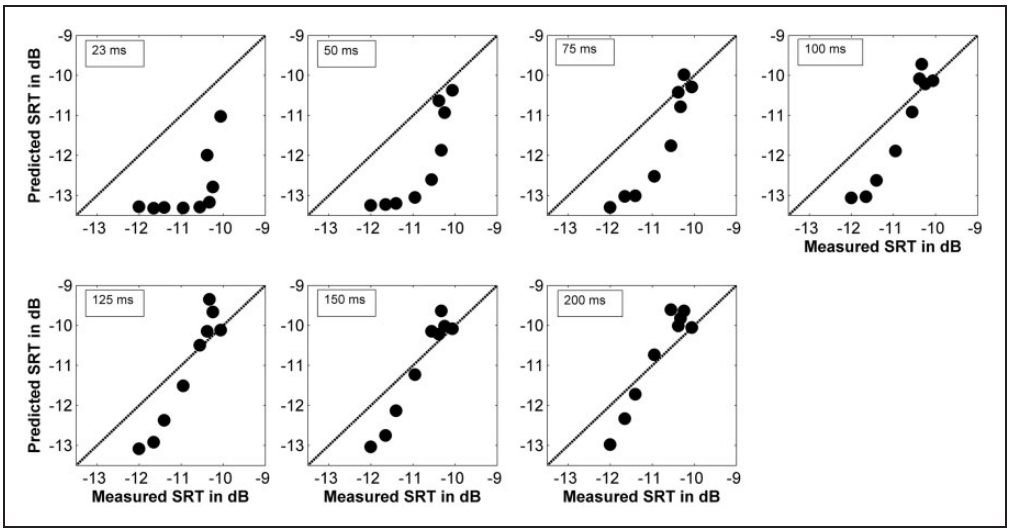

Figure 4. Predicted and observed SRTs for all tested IPMFs. Predicted SRTs are plotted against measured SRTs values. The tested time constants range from 23 ms (i.e., no sluggish processing in the EC mechanism) to 200 ms.

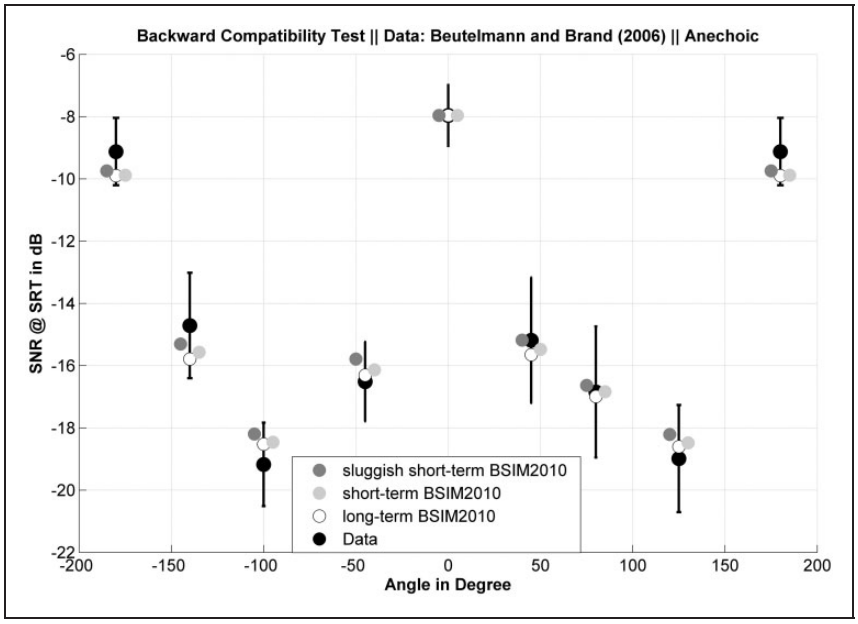

Figure 5. Predicted and observed SRTs (dB SNR) are shown as a function of the position of the noise source in the horizontal plane (*x*-axis). The speech source was always located at 0° azimuth. All model versions were calibrated to match the mean SRT obtained for a colocated speech and noise source (0° azimuth).

BSIM2010 in the anechoic condition. The anechoic condition was chosen, because the EC mechanism can be expected to provide a relatively large SNR improvement compared with more reverberant conditions. Therefore, differences across model realizations should also be largest in this condition. The results are shown in Figure 5. In general, similar results are obtained across all model realizations. Slightly higher SRTs are obtained if the short-term BSIM2010 or the sluggish version is used. The changes are below 0.3 dB in SRT, however, these higher SRTs lead to visually better agreement with perceptual data.

Discussion

This study is a time-domain analog of the experiment performed by Beutelmann et al. (2009), where the IPD of an interferer was modulated across frequency to investigate the hypothesis of independent binaural processing in auditory filters. In that study, effectively wider

binaural auditory filters were found. The current study investigated the effect of binaural sluggishness on binaural speech processing by varying the IPD of an interferer over time using different IPMFs. As hypothesized, manipulating the interaural phase of the noise resulted in a change in the release from masking. The release from masking was largest for slow interaural phase modulations and decreased gradually with increasing IPMF. As the SNR at both ears was the same and thus no better-ear listening could be performed, the decrease in speech intelligibility with increasing IPMF was most likely caused by sluggish binaural processing. A ceiling effect was observed at frequencies above 4 Hz: Above this frequency, speech intelligibility remained stable even if the IPMF was further increased. This is consistent with data from Grantham and Whightman (1979), who showed that pulsed tone detection thresholds in noise with time-varying interaural correlation information were nearly unaffected up to a modulation frequency of 4 Hz. They derived estimates of a binaural window ranging from 44 ms to 243 ms.

There are several possible explanations for this finding: For low IPMFs, it is easy to perceptually segregate target speech and interfering noise, as they are lateralized differently. The speech is perceived as being localized in the middle of the head, while the interfering noise is perceived as moving from left to right and back again. It becomes harder to attend to the target speech when both speech and noise are perceived as being located in the middle of the head. The interaural phase differences of speech and noise provide a localization cue enabling the listener to perceptually segregate speech from noise. For low IPMFs, the duration of a continuously perceived localization can be quite long. If the IPMF is increased, the duration is reduced. On the other hand, an increased IPMF between the left and right ears can be interpreted as a decorrelation of the noises on the left and right sides. A rapid change in IPD (or ITD) over time reduces the (long-term) correlation of the dichotic signal, as the maximum peak of the cross correlation appears at variable time lags. Averaging over time leads to a broadening and reduction of the cross correlation function. However, when listening to these signals, they do not appear as completely uncorrelated noises, as it is still audible that they originate from the same source.

Licklider (1948) observed a release from masking of 0.5 dB to 1 dB in situations where the noises between left and right sides were uncorrelated, while the speech signal was interaurally in phase. This is in line with the asymptotic release from masking for high IPMF values.

The outcome of the experiment was simulated using the short-term version of BSIM2010. The short-term BSIM2010 was able to predict the release from masking for the lowest IPMF (0.25 Hz). However, it was not able to predict the decrease in masking release with increasing frequency observed in the perception experiment. This was expected because binaural sluggishness was not incorporated in the original model.

As the short-term SII frame length of 23.2 ms was not sufficient to explain the data, the underlying time constant of the EC mechanism was separated from the SII calculation and a time constant defining a binaural window in the EC mechanism ($\tau_{sluggish}$) was introduced. In this approach, EC parameters are still estimated in short time frames of 23.2 ms, but the median EC estimate within the binaural window described by τ_{EC} is considered for further processing. Using this approach, the estimation of ITDs and ILDs is still fast, but using only the median value for further processing introduces a sluggish component in the binaural processing stage. It would have also been possible to consider the mean value, but the median was chosen because it is more robust against outliers in the estimation process and to be independent of the underlying distribution of ITD and ILD values. Time constants ranging from 50 ms to 200 ms were tested. The decrease in masking release with increasing IPMF observed in the experiment was only captured for $\tau_{sluggish} \geqslant 100$ ms. Increasing the time constant to values higher than 100 ms slightly improved predictions especially for the low IPMFs; this also led to an improvement in RMSE and bias. However, above 100 ms, the RMSE and bias between predictions and measured data were always within the standard deviation of the Oldenburg sentence test procedure. This is in line with previously reported binaural time constants (e.g., 110 ms; Culling & Summerfield, 1998). However, for high modulation frequencies, the release from masking was slightly underestimated. Consistent with the perceptual results, the predicted SRTs were constant for modulation frequencies above 4 Hz. In summary, it was necessary to separate the time constant involved in binaural processing (binaural sluggishness) from the time constant involved in speech processing (listening in the dips), to predict speech intelligibility in binaural dynamic scenes, while keeping the short-term back end to process speech in envelope-modulated interfering signals. Introducing a second time constant of approximately 200 ms in the EC mechanism to account for sluggish binaural processing led to successful SRT predictions in conditions with rapidly changing IPD information. In this experiment, the binaural time constants were derived from synthetic stimuli rather than realistic stimuli. The next step would be to test the refined BSIM2010 in scenarios, where the location of an interferer or the target is changed over time, leading to a congruent change in ITD and ILD information, instead of IPD information only. Nevertheless, only IPD cues were used in this study as they allow for analyzing pure binaural unmasking by excluding better ear cues. In this way, the experiments were focused on the EC stage of the model. In the

experiment of Culling and Mansell (2013), more natural stimuli were used than in our study. However, in their Experiment 1, the noise sources at $\pm 105^{\circ}$ were switched on and off using square wave modulation. Using this approach, also fast-switching better-ear listening can explain the results obtained in this experiment. Our study aimed at analyzing the benefit because of binaurally processing of interaural differences in temporal fine structure. Using their scenario, it is more difficult to derive a binaural time constant as the monaural component cannot be discarded. In their Experiment 2, they tried to investigate the effect of binaural sluggishness. The ITD and the ILD components of the masker-HRTFs were disentangled by manipulating the used HRTFs such that they either kept the interaural disparity in time or the interaural disparity in level. Our experiment can be seen as their ITD condition.

Backward compatibility of the sluggish short-term BSIM2010 was guaranteed by simulating an experiment from Beutelmann and Brand (2006), which showed no effect of the used time constant in the EC mechanism if a spatially stationary scenario is considered. Therefore, the extended model can be used for scenarios with stationary maskers, temporally fluctuating maskers, and spatially fluctuating maskers.

Conclusions

In this study, it was shown that binaural sluggishness has an effect on the binaural unmasking of speech if the IPD of the noise is changing over time. Up to an IPMF of 4 Hz, a release from masking was observed, which differed significantly from the release of masking for higher IPMFs. The binaural auditory system was only able to achieve substantial binaural unmasking if the IPD did not change too rapidly. The short-term version of BSIM2010 needed to be refined in order to account for the obtained data. It was shown that a temporal window of 200 ms applied to the EC mechanism provides best agreement with perceptual data.

Acknowledgment

The authors thank two anonymous reviewers and the associate editor Torsten Dau for their helpful comments on an earlier version of the article.

Declaration of Conflicting Interests

The authors declared no potential conflict 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 supported by a starting grant of the SFB/TRR31: "The active auditory system" at the University of Oldenburg and by the DFG Cluster of Excellence EXC 1077/1 "Hearing4all."

References

Akeroyd, M. A., & Bernstein, L. R. (2001). The variation across time of sensitivity to interaural disparities: Behavioral measurements and quantitative analyses. *Journal of the Acoustical Society of America*, *100*, 2516–2526. doi: 10.1121/1.1412442.

Akeroyd, M. A., & Summerfield, Q. (1999). A binaural analog of gap detection. *Journal of the Acoustical Society of America*, *105*, 2807–2820. doi: 10.1121/1.426897.

Andersen, A. H., de Haan, J. M., Tan, Z.-H., & Jensen, J. (2016). Predicting the intelligibility of noisy and non-linearly processed binaural speech. *IEEE/ACM Transactions on Speech, Audio and Language Processing*, *24*(11), 1908–1920. doi: 10.1109/TASLP.2016.2588002.

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

Bernstein, L. R., Trahiotis, C., Akeroyd, M. A., & Hartung, K. (2001). Sensitivity to brief changes of interaural time and intensity. *Journal of the Acoustical Society of America*, *109*, 1604–1615. doi: 10.1121/1.1354203.

Beutelmann, R., & Brand, T. (2006). Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. *Acoustical Society of America*, *120*(1), 331–342. doi: 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. doi: 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*(4), 2479–2497. doi: 10.1121/1.3295575.

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*(6), 2801–2810. doi: 10.1121/1.1479152.

Bronkhorst, A. W. (2000). The cocktail party phenomenon: A review on research on speech intelligibility in multiple-talker conditions. *Acta Acoustica United with Acustica*, *86*(1), 117–128.

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. doi: 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*(5), 75–79. doi: 10.1121/1.1907229.

Culling, J. F., & Mansel, E. R. (2013). Speech intelligibility among modulated and spatially distributed noise sources. *Journal of the Acoustical Society of America*, *133*, 2254–2261. doi: 10.1121/1.4794384.

Culling, J. F., & Summerfield, Q. (1998). Measurements of the binaural temporal window using a detection task. *Journal of the Acoustical Society of America*, *103*(6), 3540–3553. doi: 10.1121/1.423061.

Durlach, N. I. (1963). Equalization and cancellation theory of binaural masking-level differences. *Journal of the Acoustical Society of America*, *35*(8), 1206–1218. doi: 10.1121/1.1918675.

Egan, J. P. (1965). Masking-level differences as a function of interaural disparities in intensity of signal and noise. *Journal of the Acoustical Society of America*, *38*, 1043–1049. doi: 10.1121/1.1909836.

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

Grantham, D. W., & Whightman, F. L. (1979). Detectability of a pulsed tone in the presence of a masker with time-varying interaural correlation. *Journal of the Acoustical Society of America*, *65*(6), 1509–1517. doi: 10.1121/1.382915.

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

Holube, I., Kinkel, M., & Kollmeier, B. (1998). Binaural and monaural auditory filter bandwidths and time constants in probe tone detection experiments. *Journal of the Acoustical Society of America*, *104*(4), 2412–2425. doi: 10.1121/1.423773.

Jørgensen, S., Ewert, S. D., & Dau, T. (2013). A multi-resolution envelope power based model for speech intelligibility. *Journal of the Acoustical Society of America*, *134*(1), 436–446. doi: 10.1121/1.4807563.

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. doi: 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*(1), 387–399. doi: 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*(1), 218–231. doi: 10.1121/1.3662075.

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*(2), 150–159. doi: 10.1121/1.1906358.

Rhebergen, K. S., & 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. *Journal of the Acoustical Society of America*, *117*, 2181–2192. doi: 10.1121/1.1861713.

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*(7), 2125–2136. doi: 10.1109/TASL.2011.2114881.

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 Thesis). RTWH Aachen, Germany.

Wagener, K., Brand, T., Kühnel, V., & 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). *Zeitschrift. für Audiologie, Audiological Acoustics*, *38*, 4–15.

Wagener, K., Brand, T., Kühnel, V., & 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). *Zeitschrift für Audiologie, Audiological Acoustics*, *38*, 44–56.

Wagener, K., Brand, T., Kühnel, V., & 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). *Zeitschrift für Audiologie, Audiological Acoustics*, *38*, 86–95.

Wan, R., Durlach, N. I., & Colburn, H. S. (2014). Application of a short-time version of the equalization–cancellation model to speech intelligibility experiments with speech maskers. *Journal of the Acoustical Society of America*, *136*, 768–776. doi: 10.1121/1.4884767.