Further validation of a binaural model predicting speech intelligibility against envelope-modulated noises

Thibault Vicente, Mathieu Lavandier

Further validation of a binaural model predicting speech intelligibility against envelope-modulated noises

Thibault Vicente[a,*], Mathieu Lavandier[a]

[a] *Univ Lyon, ENTPE, Laboratoire Génie Civil et Bâtiment, Rue Maurice Audin, 69518 Vaulx-en-Velin Cedex, France*

Abstract

Collin and Lavandier [J. Acoust. Soc. Am. 134, 1146-1159 (2013)] proposed a binaural model predicting speech intelligibility against envelope-modulated noises, evaluated in 24 acoustic conditions, involving similar masker types. The aim of the present study was to test the model robustness modeling 80 additional conditions, and evaluate the influence of its parameters using an approach inspired by a variance-based sensitivity analysis. First, the data from four experiments from the literature and one specifically designed for the present study were used to evaluate the prediction performance of the model, investigate potential interactions between its parameters, and define their values leading to the best predictions. A revision of the model allowed to account for binaural sluggishness. Finally, the optimized model was tested on an additional dataset not used to define its parameters. Overall, one hundred conditions split into six experiments were modeled. Correlation between data and predictions ranged from 0.85 to 0.96 across experiments, and mean absolute prediction errors were between 0.5 and 1.4 dB.

Keywords: Auditory Modeling, Binaural Perception, Speech Intelligibility

1. Introduction

Our auditory system can use monaural and binaural mechanisms in order to improve the intelligibility of speech in noise. One monaural mechanism is our ability to catch target information when there is less energy in the masker signal due to masker envelope modulations, also known as "glimpsing" or "dip

*Corresponding author
Email address: thibault.vicente@entpe.fr (Thibault Vicente)

listening" (Festen and Plomp, 1990). When the target is spatially separated from the masking noise, intelligibility is improved. This spatial unmasking or spatial release from masking (SRM) is generally thought to be based on two mechanisms relying on binaural cues: better-ear listening and binaural unmasking. For instance, when a target speech is in front of the listener while the masker is placed on its side, the noise will arrive with less energy and later at the ear not on the masker side thus producing (1) interaural level differences (ILDs) and (2) interaural time differences (ITDs). Due to the ILD, the signal-to-noise ratios (SNRs) will be different at the two ears. The ear on the side opposite to the masker will provide a better SNR. Better-ear listening consists in using this ear to improve target intelligibility, which is also improved due to the difference in ITD of masker and target signals. According to the equalization-cancellation (EC-theory ; Durlach, 1972), the binaural system is able to cancel part of the noise to improve the internal SNR, by a mechanism known as binaural unmasking.

In the presence of multiple envelope-modulated noises, the SNR at each ear can change quickly over the time, so that the better ear is not always the same. The auditory system can take advantage of these variations, switching to the better ear. This ability is often called "better-ear glimpsing". The exact nature of this mechanism is not yet clear. Better-ear glimpsing could be a "true" binaural mechanism in which the auditory system compares the SNR at the two ears and then switch back and forth from one ear to the other to follow the ear with the best SNR (Culling and Mansell, 2013). It could also result from two simultaneous monaural mechanisms at each ear, providing the SNRs at both ears (Brungart and Iyer, 2012). These two interpretations might not involve the same time constants or limitations in terms of following changes across time at the ears. The binaural system appears to be sluggish compared to the monaural system (Grantham, 1982). This binaural "sluggishness" corresponding to a poorer temporal resolution can be modeled by using a longer time window when describing the mechanism. Values ranging from 40 to 250 ms have been proposed for a binaural temporal window (Culling and Summerfield, 1998; Culling and

Mansell, 2013; Grantham and Wightman, 1979; Hauth and Brand, 2018). In contrast, the time constant usually used to describe the monaural system is about 8-13 ms (Moore et al., 1988; Plack and Moore, 1990). Culling and Mansell (2013) provided evidence that better-ear glimpsing could be "truly" binaural and rely on switching across ears, since they found that performance was highly dependent on the required switching rate, and that this binaural switching could be rather sluggish.

In order to predict the effects of these mechanisms on speech intelligibility, different binaural models have been proposed (for a detailed review, see Lavandier and Best, in press, 2020). The current study will concentrate on the model proposed by Collin and Lavandier (2013) to predict speech intelligibility against multiple envelope-modulated noises in rooms. This model has four parameters: the size of the temporal windows used for computing the better-ear listening and binaural unmasking components, the degree of sampling of the spectral information, and a SNR ceiling used when estimating better-ear listening (see section 3 for the description of the model). The influence of these parameters has not been thoroughly evaluated yet, only a few values of SNR ceiling were tested while the other parameters were not varied. Moreover, the model has only been evaluated in a limited number of conditions, all involving the same speech material.

The main aim of the present study was to test the robustness of the model proposed by Collin and Lavandier (2013), considering critical conditions and also different speech materials (see section 2.1). The influence of the model parameters was evaluated using an approach inspired by a variance-based sensitivity analysis (see section 4.1). It involved the predictions — varying the model parameters — of four previously published experiments and one specifically designed for the present study. The results allowed highlighting the potential interactions between model parameters, as well as the parameter values leading to the best predictions across the five experiments. Another aim of the study was to analyze in details the model predictions, thus highlighting the effects and configurations accurately predicted and the remaining limitations of the model.

The size of the temporal window used to model binaural unmasking was revised, so that binaural sluggishness could be partly described. This study also tried to play its part in discussing the controversial concepts of better-ear glimpsing mentioned above. Finally, the optimized model was tested using an additional dataset not used to define its parameter values (see section 5).

With the proposed model, we want to provide a metric able to predict speech intelligibility in real-life listening. This is why we considered conditions involving running speech for the target, speech modulations for the maskers and real-room reverberation. However, it is sometimes useful to consider unrealistic synthetic stimuli. Isolating better-ear listening and binaural unmasking is not realistic, but tests whether the model can predict both effects correctly.

The proposed model is made available to the community and a code can be downloaded here: https://mathieulavandier.wordpress.com/home/models/.

2. Data

2.1. Data sets used to test the model parameter

Five experiments were used to test the model parameters. The experiments 1 and 2 of Culling and Mansell (2013) are abbreviated CM1 and CM2, the experiments 1 and 4 of Collin and Lavandier (2013) are CL1 and CL4, the experiment run in the present study is VL. A summary of the design of each experiment is presented in Table 1, for more details referred to the related publications (Culling and Mansell, 2013; Collin and Lavandier, 2013; and Appendix A, respectively). The "co-located" condition will refer to the configuration where target and noise(s) are at the same spatial position, otherwise the configuration will be referred to as "separated". Positive azimuths correspond to the right side of the listeners. All the noises used as masking sources had the same long-term spectrum as the target speech (speech-shaped noises, SSNs).

Two experiments from Culling and Mansell (2013) were chosen in order to test the model in anechoic conditions and in presence of artificially modulated maskers. In particular, CM2 investigated the influence of binaural sluggishness

Exp.	Number of Noises	Noise modulation	Noise Distance in room	Noise Azimuth	Cues available
CM1	1 or 2	Steady-state or modulated (10-Hz square wave, 50% duty cycle)	Anechoic	$0°^T$, 105° or ± 105°	ITD+ILD
CM2	1 or 2	Steady-state (0 Hz) or modulated (1-, 2-, 5-, 10-, or 20-Hz square wave, 50% duty cycle)	Anechoic	$0°^T$ or ± 105°	ITD+ILD, ILD-only or ITD-only
CL1	1	Steady-state or modulated (broadband envelope of 1, 2 or 4 voices)	0.65^T, 1.25 or 5 m	$0°^T$	ITD+ILD
CL4	1 or 2	Steady-state or modulated (broadband envelope of 1 or 2 voices)	0.65^T m	$0°^T$, 25° or ± 25°	ITD+ILD
VL	1	Steady-state or modulated (broadband envelope of 1 voice)	0.65^T or 5 m	$25°^T$ or -25°	ITD+ILD or no ITD/no tail

Table 1: Summary of the experimental designs used to test the model parameters. The superscript 'T' indicates the target's distance and azimuth and defines the co-located condition. The last column indicates the nature of the binaural cues available in the tested signals.

on better-ear listening and binaural unmasking independently, which is relevant to test the temporal resolutions used in the model. The SNRs at which the listener reports 50% of the target words correctly, the so-called speech reception thresholds (SRTs), are displayed on Fig. 2 as a function of noise azimuth, number of noises and type of noise modulation for CM1 and on Fig. 3 as a function of modulation rate for CM2.

CL1 was chosen to test the model performance at predicting the effect of reverberation on speech intelligibility in the presence of modulated noise and to consider envelope modulations more characteristic of real speech (rather than artificial modulations; see Table 1). The measured SRTs are plotted as a function of masker distance in Fig. 4, each panel corresponds to a modulation depth for the noise. CL4 was considered because it involved reverberation and speech modulations for the noises, but also asymmetrical configurations in which binaural hearing and SRM were involved. Figure 5 presents the SRTs measured for each type of masking noise.

VL was designed to evaluate the model at predicting the influence of reverberation filling in the masker modulation gaps in an asymmetrical condition (see Appendix A). The better-ear component of the model was tested on its own and in combination with the binaural unmasking component. The measured SRT are plotted in Fig. 6 as a function of the noise position, with one panel for each type of noise modulation.

2.2. Data set used to validate the revised model

In order to validate the revised model, the experiment of Ewert et al. (2017) was considered. The target was always simulated in front of the listener at 0.8 m. Two maskers were involved, either co-located with the target or symmetrically placed on both sides of the listener at $\pm$ 60°. Six types of masker were tested, but only the four energetic maskers are considered here. Our model is not designed to predict the effects of informational masking. A steady-state noise and three types of envelope-modulated noise were tested. The modulated noises were generated using: a 8-Hz sinusoidal amplitude modulation, the broadband en-

velope of a speech signal, and speech modulations incoherent across-frequency (named here sinusoidal noise, 1-voice noise and 1-voice Freq. Inc. noise, respectively). The last type of modulation was obtained by modulating different spectral regions of the noise with different speech envelopes.

Five head-related impulse response (HRIR) conditions were tested. (i) A natural HRIR (ITD+ILD) condition without processing (ii) An ILD-only condition (iii) An ITD-only condition for which the HRIRs spectra at 0° and 60° ("Magnitude 0" and "Magnitude 60", respectively) were averaged across ears (iv) An "Independent" condition was created using the natural HRIR at 0°. One noise source was convolved only with the right ear HRIR while the other was convolved only with the left ear HRIR, resulting in a listening without crosstalk and coherence between ears, thus creating an infinite ILD. (v) Two Ideal Monaural Better-ear Mask (IMBM, Brungart and Iyer, 2012) conditions were also created using the natural HRIR and the independent HRIR (resulting in an IMBM condition or an independent IMBM condition, respectively). The 4 noise modulation types, 5 HRIR conditions and 2 spatial configurations resulted in 40 conditions. The measured SRTs are plotted in Fig. 7 as a function of the masker type, each panel corresponding to a given type of HRIRs.

3. Model description

A block diagram of the model is provided on Fig. 1. The model takes as inputs the target and combined masker signal at the ears. It predicts the target intelligibility taking into account binaural unmasking and better-ear listening as proposed by Lavandier and Culling (2010). The model computes, per time frame (Rhebergen and Versfeld, 2005) and frequency band, the SNR at the better ear and the binaural unmasking advantage which is added to the better-ear SNR (Collin and Lavandier, 2013). After integration across frequency and averaging across time frames, the model output is a SNR in the corresponding condition, referred to as "binaural ratio" in the following. Differences in binaural ratios can be directly compared to differences in intelligibility thresholds measured in dB.

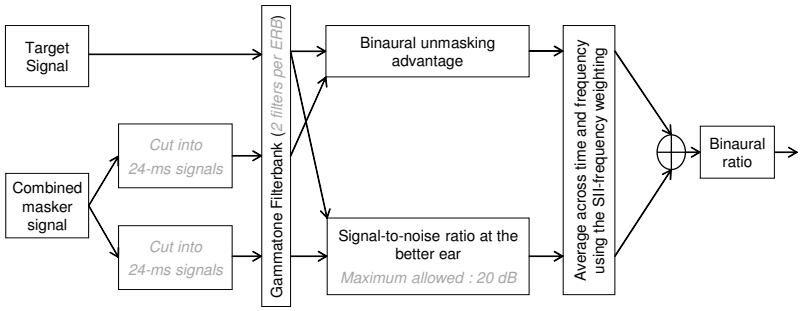

Figure 1: Block diagram of the original model (Collin and Lavandier, 2013), with the parameters tested in the present study highlighted in grey.

Binaural ratios are first inverted to be compared to SRTs, so that the inverted ratio decreases with the SRT when intelligibility increases. Because only relative differences across conditions can be predicted by the model, a reference needs to be chosen to compare the inverted ratios to the SRTs. A single constant is added to all inverted ratios, so that their mean equals this reference. For each experiment presented here, this reference is the average measured SRT across conditions (Jelfs et al., 2011; Lavandier et al., 2012).

Peaks in the masker signal induce an increase of target masking whereas pauses induce a decrease of this masking. Therefore, the model considers masking energy as a function of time. In order to consider the pauses/envelope modulations in the target speech as important information for its intelligibility (e.g. the gaps between words), the model considers the average level of the target across time rather than its instantaneous level within short-time frames (Rhebergen and Versfeld, 2005). Like Cubick et al. (2018), instead of replacing the target speech by a stationary signal with a similar long-term spectrum and interaural parameters and applying the short-term analysis on this signal (Collin and Lavandier, 2013), the present implementation of the model computes the long-term statistics of the target only once and combines these statistics with the short-term spectrum and interaural parameters of the noise to compute the

better-ear and binaural unmasking components within each time frame. Thus, as a model input, the target sentences at the ears are replaced by an averaged target signal generated by adding at least[1] 60 target sentences (truncated to the duration of the shortest sentence), and this averaged signal is not submitted to the temporal decomposition into short-time frames used for the masker.

The masker signals are cut into frames using half-overlapping Hann windows, before being passed through a Gammatone filterbank (Patterson et al., 1987) with two filters per equivalent rectangular bandwidth (ERB ; Moore and Glasberg, 1983). The bandwidth of the Gammatone filters is about 1 ERB, thus the filters are half-overlapping. Within each time frame and frequency band, the two components of spatial unmasking are modeled, (1) the binaural unmasking advantage is estimated using a formula proposed by Culling et al. (2005), which depends on the masker interaural coherence and on the target and masker interaural phase differences. The target and masker signals are both cross-correlated to derive these interaural parameters. The coherence is taken as the maximum of the cross-correlation function, and the phase difference is obtained by multiplying the corresponding delay by the center frequency of the band. The search of maximum delay in the cross-correlation functions is limited to the range plus/minus half the period of the channel center frequency, so that the model does not predict any binaural unmasking advantage at high frequency (Durlach, 1972). The binaural unmasking advantage is set to zero if the masking noise power is zero at one of the ears in the considered band and frame. (2) The SNR is also computed at each ear, and the best SNR across ear is selected (thus independently for each frequency band and each time frame). A ceiling parameter corresponding to the maximum better-ear ratio allowed by frequency band and time frame is introduced at this stage, to avoid the SNR

[1]For CM1 and CM2, 80 and 160 target sentences were used, respectively. Regarding CL1 and CL4, 60 target sentences were averaged. To model the experiment of the present study, 120 sentences of each target type were used. To model the experiment of Ewert et al. (2017), the 2100 target sentences of the Oldenburg Satztest corpus were used.

ratio tending to infinity in masker pauses. Conceptually, this parameter is implemented to explain the fact that a listener does not need an infinite SNR to fully understand the target. (3) The better-ear ratios and binaural unmasking advantages estimated per frequency bands and time frames are then integrated across frequency using the SII weighting (ANSI S3.5, 1997) and averaged across time. Finally, the two values are added to get the binaural ratio.

The first aim of the present study was to test the four parameters of the model (see Fig. 3): the duration of the Hann window used for computing the binaural unmasking advantage ("BU" in ms), the duration of the Hann window used for computing the better-ear SNR ("BE" in ms) — those are the two temporal resolutions of the model — the number of gammatone filters per ERB (the model spectral sampling "SpecSamp") and finally the ceiling parameter ("Ceiling" in dB). The parameter values were previously set to 24 ms, 24 ms, 2 filters per ERB and 20 dB, respectively (Collin and Lavandier, 2013; Cubick et al., 2018). In particular, the same temporal resolution was used to model better-ear listening and binaural unmasking; whereas the temporal resolution of the two mechanisms (and their susceptibility to binaural sluggishness) was investigated independently here.

4. Revision of the model

4.1. A method inspired by a sensitivity analysis

One of the aims of the present study was to quantify the relative influence of each parameter of the tested model and to identify potential interactions between these parameters. The method used was inspired by a variance-based sensitivity analysis, which has been described in details by Saltelli et al. (2010). Conceptually, the method consists in computing model predictions while varying the value of its parameters. Then, sensitivity indices can estimate the rate of model output variance due to a given parameter or to an interaction between parameters. For instance, the first order sensitivity index evaluates the direct impact of varying a given parameter on the model output, a second order sen-

sitivity index evaluates the amount of variance in the model output that can be attributed to an interaction between two parameters. The indices are computed so that they all take values between 0 and 1, the sum across all indices is equal to 1 and the higher the index the stronger the influence of the corresponding parameter or interaction. This analysis allows to determine whether strong interactions between model parameters prevent from defining these parameters values independently one from the other, and to identify the most influential parameters of the model.

Five values were tested for each of the 4 model parameters, resulting in 625 combinations. The equivalent rectangular window duration of a Hann window is only half of its full length (Beutelmann et al., 2010). The durations of the Hann windows were here converted into equivalent rectangular duration (ERD). The durations tested were (for both BU and BE): 8, 12, 40, 100, 200 ms (ERD). These values span the range of the monaural and binaural time constants proposed in the literature and mentioned in the Introduction. The values tested for SpecSamp were: 2, 1, 2/3, 1/2, 2/5 filter(s) per ERB. The values tested for Ceiling were 8, 12, 16, 20, 24 dB.

The mean absolute error (Mean Err.) and the correlation (r) between data and predictions were chosen as the outputs of the model for the sensitivity analysis. Mean Err. was computed as the averaged across conditions of the absolute difference between measured and predicted SRTs. The maximum of this absolute error (Max. Err.) was also considered in order to have an information on the worst predictions, but it was not used as a criterion in the sensitivity analysis. The root-mean-square error (RMS Err.) between data and prediction was also calculated but not used as a criterion in the sensitivity analysis either. It was computed as the square root of the quadratic mean of the difference between data and predictions.

The 625 combinations of parameter values were tested for the 5 experiments described above. The sensitivity indices were estimated using either r or Mean Err. as model output. The interactions between parameters and the relative influence of each parameter were studied using these indices. Afterwards, for each

experiment, the independent parameters were varied independently to define the value(s) leading to the best predictions, whereas for the interacting parameters, these values were defined while varying the parameters simultaneously. The best predictions values were then compared across experiments in order to find a single common value for each parameter leading to good predictions across all experiments. A qualitative analysis of the predictions was also considered, to eventually help define the final parameter values if there were more than a single value leading to best predictions across experiments. Values leading to predictions conceptually wrong (e.g., missing a basic effect observed in the data) were excluded prior to this analysis. For each type of parameter, independent or interacting, if its original value (Collin and Lavandier, 2013; Cubick et al., 2018) was among the values leading to the best predictions, then this value was selected for the parameter because there was no relevant argument for a change. The definition of the best parameter values was done while keeping in mind which parameters were the most influential.

4.2. Results

The conclusions of the sensitivity analysis were similar when considering r or Mean Err. as model output. The sensitivity index values were different but the observed trends were same. Only the results obtained with the Mean Err. are presented here. All the first order sensitivity indices and the second order sensitivity index between Ceiling and BE are displayed in Table 2.

The most directly influential parameter (displayed in bold for each experiment in Table 2) was Ceiling for CM1, CL1, CL4 and VL while it was BE for CM2. For example for CL4, the corresponding index was equal to 43%, meaning that 43% of the Mean Err. variance (over the 625 predictions) was due to the variations of Ceiling. To say it differently, if the sensitivity analysis had been ran with a constant Ceiling and only the three others parameters were varied, then the Mean Err. variance would have been at least 43% lower. The only non-negligible second order sensitivity index was for the interaction between Ceiling and BE (17% on average across experiments, while the second highest

Exp.	CM1	CM2	CL1	CL4	VL
	1^{st} order indices (%)				
Ceiling	**58**	21	**48**	**43**	**68**
BE	20	**54**	28	8	12
BU	1	3	0	0	18
SpecSamp	0	0	0	15	0
Sum of 1^{st} order indices	79	78	76	66	98
	2^{nd} order indices (%)				
Ceiling/BE	19	14	24	26	2

Table 2: First order indices, their sum and the second order index between Ceiling and BE for all experiments. The highest first order index for a given experiment is displayed in bold. Only the main interaction (between Ceiling and BE) is displayed here.

second order index was limited to 1% on average).

The sum of all first order sensitivity indices and the second order sensitivity index between Ceiling and BE (sum of the two last lines of Table 2), per experiment, led to rates higher or equal to 92% (including 100 % for CL1 and VL). In other words, across all experiments, the variation of Mean Err. was almost entirely due to direct impacts of the parameters and the interaction between BE and Ceiling. The few percent of variance left were split into the ten other sensitivity indices. From this observation, the choice of the final values to be used was done individually for SpecSamp and BU, but BE and Ceiling were considered together.

None of the experiments were discriminating to choose the SpecSamp value, in agreement with the fact that its first order sensitivity index was equal to 0% for four experiments. This parameter had some limited influence only for the predictions of CL1, but in practice all values led to accurate predictions. Therefore, the original value used by Collin and Lavandier (2013) was kept unchanged (2 Gammatone filters per ERB, see section 4.4 concerning this choice).

Concerning BU, the predictions for CL4 and CL1 were not affected by changing BU values. For CM1 and VL, the longer the window duration the lower Mean Err. ; and for CM2 the lowest Mean Err. was reached with the longest window duration. Hence, the analysis suggests to change the value of BU from 24 to 400 ms (from 12 to 200 ms ERD). It was decided to set BU to 300 ms (150 ms ERD), a value not tested above but which corresponds to the midst of the binaural temporal windows reported in the literature (Culling and Summerfield, 1998; Culling and Mansell, 2013; Grantham and Wightman, 1979; Hauth and Brand, 2018), which ranged from 80 ms to 500 ms (from 40 to 250 ms ERD, see Introduction). The difference in predicted SRT using a BU of 300 rather than 400 ms was below 0.1 dB in each of the five experiments considered above.

The values for Ceiling and BE giving the best predictions were deduced by removing the values leading to inconsistent predictions. The tested values for Ceiling were: 8, 12, 16, 20, 24 dB ; for BE, they were: 16, 24, 80, 200, 400 ms (or 8, 12, 40, 100, 200 ms ERD). In CM1 and CM2, when BE was set to 80, 200 or 400 ms the model predicted SRTs with obvious deviation from the data for all values of Ceiling (e.g. no difference in predicted SRT for stationary and modulated noises in CM1). These prediction errors are considerably reduced with the shortest window durations, so that only the values 16 and 24 ms (8 and 12 ms ERD) remained for BE.

Model predictions for CM1 and CM2 led to conflicting results concerning the choice of Ceiling. The best predictions for CM1 were obtained for values equal to 20 or 24 dB, whereas the best predictions for CM2 were obtained for a Ceiling of 8 dB. This value was not considered further, because CM2 is the only experiment well predicted with a Ceiling of 8 dB. A Ceiling of 12 dB led a 3.2-dB overestimation of the SRTs in the conditions with one modulated noise in CM1. In this case, the model also predicted identical SRTs for the steady-state and modulated noise in the separated condition. The best model performances for CL1 were reached with a 12-dB Ceiling. Because the value of 12 dB also led to conflicting results between CL1 and CM1, it was not considered further. The remaining possible values for Ceiling after this first analysis were 16, 20 and 24

Exp.	r Orig. ; Rev.	Mean Err. Orig. ; Rev.	RMS Err. Orig. ; Rev.	Max. Err. Orig. ; Rev.
CM1	0.93 ; 0.96	1.3 ; 1.0	1.6 ; 1.3	3.5 ; 2.3
CM2	0.92 ; 0.94	1.0 ; 1.0	1.2 ; 1.0	2.4 ; 1.8
CL1	0.85 ; 0.85	0.5 ; 0.5	0.6 ; 0.6	1.3 ; 1.3
CL4	0.92 ; 0.93	0.6 ; 0.6	0.8 ; 0.8	1.6 ; 1.4
VL	0.87 ; 0.90	1.0 ; 0.9	1.0 ; 1.2	2.0 ; 1.8
Ewert et al. (2017)	NA ; 0.91	NA ; 1.4	NA ; 2.0	NA ; 7.1

Table 3: Performance statistics of the original (Orig.) and revised (Rev.) model. Mean Err., RMS Err. and Max Err. are computed in dB. The experiment of Ewert et al. (2017) was only used to validate the revised model.

dB.

The original values of Ceiling and BE used by Cubick et al. (2018) have not been discarded, meaning that they did not lead to inconsistent model predictions. Ceiling and BE were thus set to these values, 20 dB and 24 ms, respectively.

4.3. Predictions of the revised model

The SRTs predicted with the revised model are presented as solid lines for each experiment in Fig. 2 to 6. The predictions of the original model are plotted for comparison with dashed lines. On each figure, the performance statistics of the revised model are indicated (r, Mean Err., RMS Err. and Max. Err.). A comparison of the performance statistics between the original and revised model is displayed in Table 3, which shows that they are similar across experiments and both models predict accurately the data. Mean Err. and RMS Err. provide also comparable values for each experiment and model.

In CM1, the steady-state noise conditions (Fig. 2) are well predicted with errors below 1 dB. The model overestimates the SRT in the presence of a single co-located modulated masker by 2.3 dB and it underestimates by 2.3 dB the

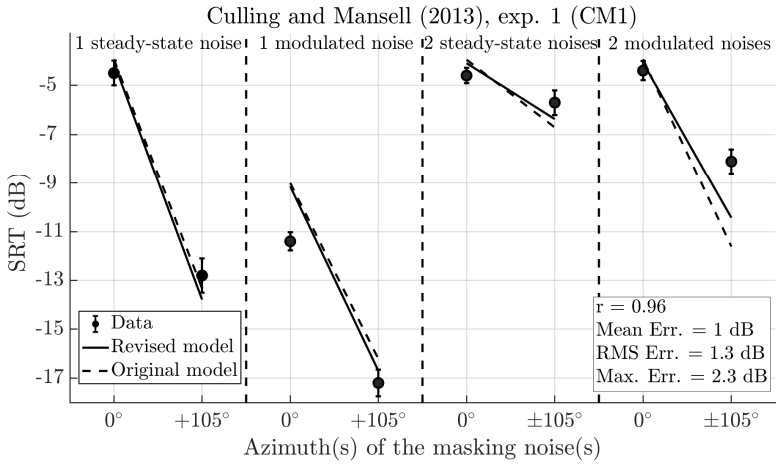

Figure 2: Mean SRTs with standard errors across listeners measured in CM1, involving 1 or 2 noises, steady-state or modulated by a 10-Hz square wave (50% duty cycle, modulated out-of-phase if they were two maskers), simulated as originating from different azimuths (0° and +105° or ±105° if there were two maskers) in an anechoic environment. The target was always at 0°. Model predictions are displayed as a solid line for the revised model and as a dashed line for the original model. Model performance statistics are displayed only for the revised model.

SRT for the symmetrical condition involving 2 separated modulated noises. The revised model improved this last prediction by 1.2 dB due to the longer duration of BU.

In CM2, changing BU from 24 ms to 300 ms enables to better predict the influence of the modulation rate (between 1 and 5 Hz only) for the ITD-only conditions, and as a result also for the ILD+ITD conditions. Concerning the ILD-only conditions, it is important to note that the original and revised models predict exactly the same binaural ratios, i.e. the predictions of these conditions are not affected by the revision. Because the average prediction (which is different for the two models because of the other conditions) is scaled to the average SRT in the experiment, the resulting predicted SRTs are different. The model predicts SRTs increasing by about 1.5 dB above the 5 Hz modulation rate for the ILD-only conditions, while the data show a 0.6-dB difference.

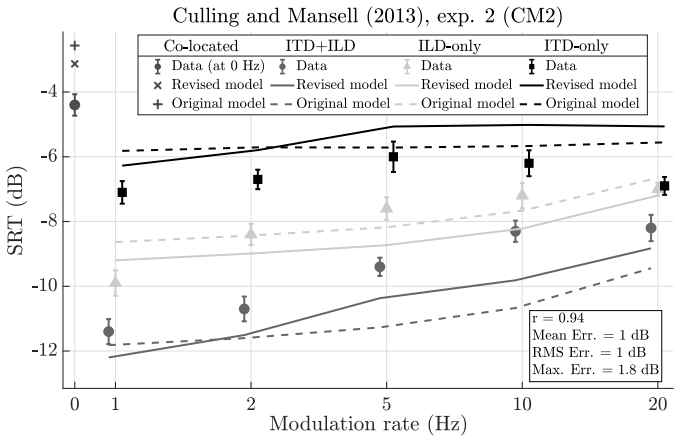

Figure 3: Mean SRTs with standard errors across listeners measured in CM2. The target was always presented at 0° in the presence of two noises placed on both sides of the listener (±105°). The noises were modulated out-of-phase by a square wave at 5 modulation rates (1, 2, 5, 10, 20 Hz). Three types of HRIR were involved (ILD+ITD, ILD-only, ITD-only). One reference condition involved a steady-state noise co-located with the target (modulation rate of 0 Hz). Model predictions for the separated configuration are displayed as a solid line for the revised model and as a dashed line for the original model. The predictions related to the co-located configuration are plotted using a cross and a plus sign for the revised and original model, respectively. Model performance statistics are displayed only for the revised model.

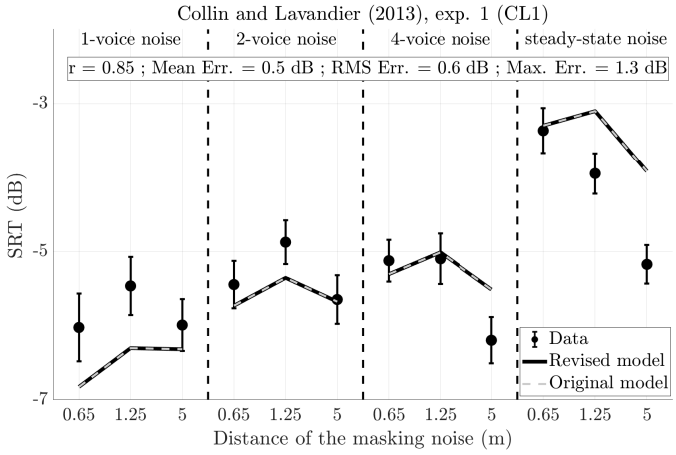

Figure 4: Mean SRTs with standard errors across listeners measured in CL1. The target was at 0.65 m in front of the listener in a lecture hall. The noise was placed at three distances (0.65, 1.25, 5 m), also in front of the listener. Four types of modulation were used for the noise (steady-state, 1-, 2- or 4-voice modulated). Model predictions are displayed as a solid line for the revised model and as a dashed line for the original model. Model performance statistics are displayed only for the revised model.

For CL1 (Fig. 4), there is no difference between the original and revised model, not surprisingly since target and masker were simulated in front of the listener, so that the influence of binaural unmasking was limited. In CL4 (Fig. 5), the model predicts accurately all the conditions involving a single masker (i.e. black symbols), only the one with a co-located steady-state noise leads to an error of about 1 dB. For the conditions with two maskers (grey symbols), the model seems to predict about 1 dB more SRM than measured in the data.

The predictions for VL are quantitatively correct (Fig. 6), suggesting that the model is able to predict the general trends measured in the data. The SRTs for the steady-state noise are better predicted than those for the 1-voice modulated noise. The model predicts a binaural unmasking advantage for the steady-state noise (difference between the black lines and the grey lines in the top panel) that was not observed in the data. In the bottom panel, the relative difference predicted between the no ITD/no tail conditions and the ILD+ITD

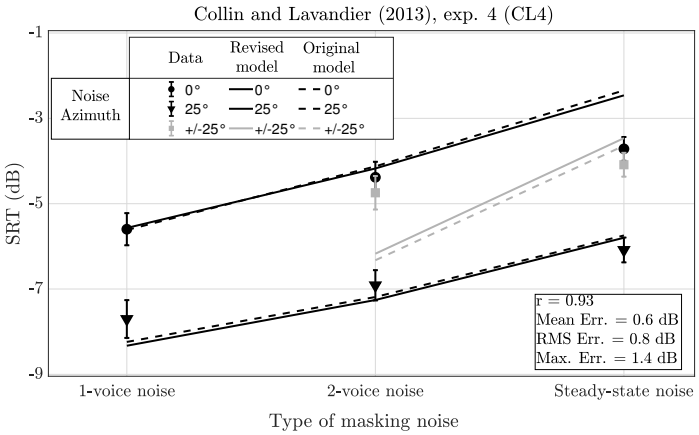

Figure 5: Mean SRTs with standard errors across listeners measured in CL4. The target was at 0.65 m in front of the listener in a meeting room. The single masker was always at 0.65 m but tested at two azimuths (0° and 25°). Three types of noise were involved (1-voice modulated, 2-voice modulated or steady-state). Two noises (steady-state or two 1-voice modulated) were tested in two configurations (0° or ±25°, 0.65 m). The revised and original model predictions are plotted as a solid and a dashed line, respectively. Model performance statistics are displayed only for the revised model.

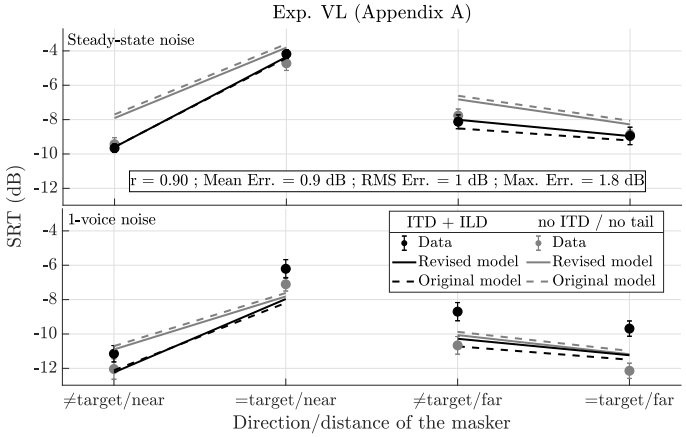

Figure 6: Mean SRTs with standard errors across listeners measured in the present study (VL). The target was placed at 0.65 m, +25° from the listener (=target/near). The noise was steady-state (top panel) or 1-voice modulated (bottom panel). It was tested at two distances (near at 0.65 m and far at 5 m) and two azimuths (+25°/=target, -25°/ =target) in a room. Two types of BRIR were involved (natural BRIRs with ITD+ILD, SEIRs with no ITD/no tail). Solid lines present the revised model predictions, while dashed lines present the original model predictions. Model performance statistics are displayed only for the revised model.

conditions (grey solid lines and black solid lines, respectively) for a given spatial configuration does not correspond to the relative differences measured in the data. This means that the model is not able to completely predict the conflicting effects of having no ITD, i.e. no binaural unmasking, and no reverberation tail, i.e. no filling in of the masker gaps.

4.4. Discussion

Considering the sensitivity analysis, it should be emphasized that the range of values over which the model parameters were varied will have influenced the magnitude of the sensitivity indices. For instance, a smaller range of Ceiling values could have led to a decrease of this parameter predominance. Inversely, a larger range of BU values could have led to higher sensitivity indices. The tested values were chosen based on previous results from the literature; but they should be kept in mind when considering the conclusions of the sensitivity

analysis.

The only strong interaction between model parameters was observed for Ceiling and BE, the two parameters involved in the computation of the SNR at the better ear. The window duration BE sets the time constant for the model to analyze an amplitude modulation in a noise envelope and Ceiling sets the maximum value of the by-band SNR/masker modulation depth from which the band contribution to intelligibility is assumed to plateau. Conceptually, if the window duration is too long, the fast modulations will not be detectable. As a result, Ceiling will not be used in the calculation of the SNR at the better ear for those modulations. Conversely, if the window duration is sufficiently short for detecting the modulation, then Ceiling will be used in the calculation and will influence the model output. So it is not surprising that these two parameters interact.

The window duration BU used to compute the binaural unmasking advantage presented lower first order sensitivity indices than the window duration BE, probably for two reasons. First, it should be noted that the better-ear listening component of the model is influenced both by the ILD/better-ear effects, but also by the effects associated with masker modulations (dip listening). Across experiments, less conditions were tested in which binaural unmasking played a role compared to those in which better-ear/dip listening played a role (e.g. in the ITD-only conditions, the better-ear component of the model was still influenced by the differences in masker modulations). As a result, the model predicts more differences across conditions that are associated with the better-ear/dip listening component. Hence, it seems normal that the model is more sensitive to the parameter associated with this latter component. The second reason is the following, as described in the previous section, the predictions were extremely far from the data when BE was set to the longer durations, only for the shorter BE the predictions described well the data. As a result, a considerable range of Mean Err. variations depended on BE. On the other hand, whatever the window duration BU was, the predictions were sufficiently close to the data, so that the range of Mean Err. variations induced by BU variations was narrower

than the range induced by the BE variations.

Considering the final choice of the values used for the model parameters, first, binaural sluggishness is better taken into account by the model with a window duration BU equal to 150 ms (ERD). As mentioned in the Introduction, several previous studies measured the binaural temporal window using different methods (Culling and Summerfield, 1998; Culling and Mansell, 2013; Grantham and Wightman, 1979; Hauth and Brand, 2018). The obtained values were between 80 to 500 ms (40 to 250 ms ERD).

Hauth and Brand (2018) investigated the effect of binaural sluggishness using a different short-time binaural speech intelligibility model (stBSIM ; Beutelmann et al., 2010). They designed an experiment in order to test the effect of binaural sluggishness on speech intelligibility. Stimuli were a steady-state noise for which interaural phase differences (IPDs) were modulated sinusoidally between $-\pi/2$ and $+\pi/2$ at different rates between 0 and 64 Hz. Increasing the modulation rate led to higher SRTs for rates up to 4 Hz, above which the rate had no significant influence on the SRT. These results are consistent with the results of CM2. When modeling their own experiment, Hauth and Brand modified the EC processing of their model to introduce binaural sluggishness that influences the definition of the EC parameter. However, the EC stage *per se* is still applied on short-time signals (for detail of implementation, see Hauth and Brand, 2018). In the current model, the binaural unmasking advantage is estimated using signals whose duration is influenced by binaural sluggishness, resulting in longer signals than in the "revised"' stBSIM. Despite this discrepancy of binaural sluggishness implementation in the models, the duration of the binaural/EC window proposed by Hauth and Brand allowing to predict accurately their data was 200 ms (ERD), which is similar to the duration (BU) highlighted in the current study.

SpecSamp did not influence the model predictions in any of the 5 experiments tested at this stage. So if one is interested in saving computing time, it seems appropriate to reduce the spectral sampling of the model. Reducing it to as low as 2 filters per 5 ERB did not impair the predictions in the 5 experiments

tested here. We choose to keep 2 filters per ERB for now, because computing time is not an issue for the current study and a better spectral sampling might be needed in future developments of the model (e.g. while considering harmonic maskers or hearing-impaired listeners).

Some studies have shown that speech intelligibility models could still lead to relevant predictions despite a loss of frequency resolution, which has to be understood as the accuracy to analyze the signals in the frequency domain. To change the frequency resolution in a model, the number and the bandwidth of the filters that analyze the signals have to be varied but keeping the same overall frequency bandwidth analysis. Kryter (1962), when developing the Articulation Index (AI, monaural speech intelligibility model), showed that its predictions using a 20-band, one-third-octave-band or octave-band method were in reasonable agreement. Steeneken and Houtgast (1980) developed and validated the Speech Transmission Index (monaural speech intelligibility model), with its computation done using an octave band method.

The conclusion of the present study is different because the results showed that the model predicted similar SRTs even if some frequency channels were not used for the computation while keeping the same filters (creating "holes" in the bandwidth in which the signals are analyzed). However, those conclusions lead to a common observation, which shows that a loss of spectral information in the signals, either by smoothing it (reduction of the frequency resolution) or not analyzing some frequency channels (reduction of the spectral sampling), still results in similar model predictions.

Regarding the Ceiling value, it has been set to 20 dB that is higher than the values implemented in the SII (ANSI S3.5, 1997) or in the AI (Kryter, 1962), +15 and +18 dB SNR, respectively. It means that the current model considers that the full target intelligibility is reached at a higher SNR. Studebaker and Sherbecoe (2002) showed that increasing the SNR up to 29 dB could still improve target intelligibility. Such a high value of Ceiling does not seem appropriate in the proposed model. Collin and Lavandier (2013) introduced Ceiling to the model. They found that a value of 10 or 15 dB reduced prediction errors. Cubick

et al. (2018) needed a Ceiling of 20 dB to optimize predictions. Implementation differences between both models (see section 3) may account for the different Ceiling values.

The values of the window duration BE used to compute the better-ear component of the model that predicted well most conditions tested here were 16 and 24 ms (8 and 12 ms ERD). These values are within the range 8 to 13 ms (ERD) of the measured monaural temporal resolution (Moore et al., 1988; Plack and Moore, 1990). The tested binaural window durations (i.e. 80, 200, 400 ms or 40, 100, 200 ms ERD) provided inconsistent predictions in some conditions. The model was not able to predict the advantage of listening in the masker dips when the temporal resolution of the better-ear component was not sufficient (i.e. when the window duration was too long). For instance, in CM1 a BE of 200 ms provided the same predicted SRT for the steady-state and modulated noises. Taking a too-long temporal window triggers an amplitude modulation smoothing in the model, so that the modulated masker appeared as a steady-state masker. Consequently, the window duration BE has to match a monaural time constant. It should be noted that Collin and Lavandier (2013) as well as Beutelmann et al. (2010) also used a temporal resolution of 24 ms. It corresponds to the best frequency-independent duration used in the monaural model of Rhebergen and Versfeld (2005).

The values retained for BE differed from Culling and Mansell's conclusion, which stated that better-ear listening is a mechanism affected by binaural sluggishness, because in CM2 there was an influence of the required ear-switching rate up to 5 Hz. Although a monaural time window is required for the proposed model, it does not mean that better-ear listening is a "double" monaural mechanism, which is not influenced by binaural sluggishness and across-ear switching. It just means that the current implementation of the model does not allow for predicting this effect of sluggishness on better-ear listening. Culling and Mansell (2013) concluded that better-ear listening is binaural because the listener has to choose which ear is more beneficial for listening to the target; but also that the monaural behavior of each ear allows for listening in the dips. So there may

be two time constants to consider for modelling better-ear listening. Modelling the effect of across-ear switching on better-ear listening is not straightforward and not implemented here. The present study however shows that better-ear listening cannot be simply modeled using a binaural temporal window. The monaural temporal resolution is required to predict the benefit associated with fast masker modulations.

5. Validation of the revised model

5.1. Predictions

The model predictions in the 5 HRIR conditions of Ewert et al. (2017) are plotted in the panels of Fig. 7. The SRTs were scaled using the mean SRT across all 40 conditions (i.e. the scaling was done only once for all panels rather than independently for each panel), in order to observe whether the model could predict the differences across HRIR conditions. The model performance across all conditions led to a r of 0.91, a MeanErr of 1.4 dB, a RMS Err. of 2.0 dB and a Max Err. of 7.1 dB. The correlation and Mean Err. are similar to those obtained for the other experiments presented above. Max Err. is considerably larger due to a single data point (last panel of Fig. 7). The performance statistics were also computed separately for each HRIR condition and are displayed in the corresponding panels of Fig. 7.

The general pattern of the predictions and the model performances are similar for the natural ITD+ILD, ILD-only and IMBM conditions (first, second and fourth panels, respectively). The solid black and grey lines represent the predictions for the co-located and separated conditions, respectively. Correlations are above 0.92 and Mean Err. around 1 dB. Max Err. is obtained for the separated conditions with the 1-voice Freq. Inc. noise. The differences in SRTs produced by the different types of masker modulation are well predicted, for both spatial configurations, except for the 1-voice Freq. Inc. modulation in all HRIR conditions and the sinusoidal modulation in the separated IMBM

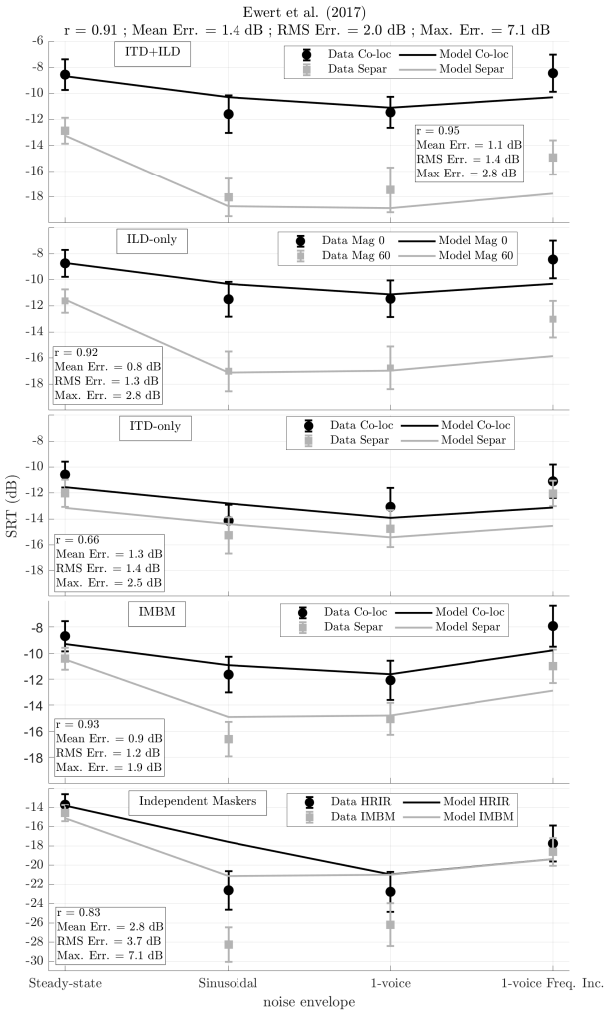

Figure 7: Mean SRTs with standard deviations across listeners measured by Ewert et al. (2017). SRTs are plotted as a function of the noise modulation type (steady-state, sinusoidal, 1-voice, 1-voice frequency incoherent). Each panel corresponds to a given type of HRIRs (ITD+ILD, ILD-only, ITD-only, IMBM, Independent Maskers). For the first four panels, two spatial configurations were tested : while the target was at 0°, the two noises were placed in front (at 0°) or on each side (±60°) of the listener (plotted in black circles and grey squares, respectively). For the last panel, two different HRIRs (natural and IMBM) at 0° were used to create the independent maskers represented with black circles and grey squares, respectively. The model predictions are plotted in solid black and grey lines in all panels. The model performance statistics across all conditions are indicated in the title, and the performances for each HRIR type are displayed in the corresponding panel.

condition, where the differences between the observed and predicted SRTs are around 2-3 dB.

The ITD-only predictions related to the Magnitude 0/60 are plotted on the third panel, with solid black and grey lines, respectively. The trends across masker modulations are not well predicted ($r = 0.66$) although Mean Err. is close to 1 dB. The 2.5-dB Max Err. occurs again for the separated configuration with the 1-voice Freq. Inc. noise. The difference between the black and grey lines is almost equal to the difference between the solid black and grey symbols, indicating that the model predicts the influence of spectral coloration.

The predictions for the independent masker conditions are shown in the last panel. The steady-state noise conditions are well predicted, while the SRTs for the modulated noises are less well predicted. The SRTs are overestimated for the sinusoidal and 1-voice modulation, while they are underestimated for the 1-voice Freq. Inc. modulation. The model does not predict any difference between the HRIR and IMBM for the 1-voice and the 1-voice Freq. Inc. modulation (while there is a difference in the data). Max. Err. occurs for the independent IMBM condition with sinusoidal modulation. Mean Err. is equal to 2.8 dB for this HRIR condition, which is the worstly predicted.

5.2. Discussion

First, the overall model performance on this experiment is relatively good, and comparable to the stBSIM performance (Ewert et al., 2017). The performance statistics of the two models cannot be compared because the prediction errors of the stBSIM were largely increased due to mispredictions of informational masking. Those conditions were not even attempted for here, because they cannot be described with our only-energetic-masking model. The proposed model is able to predict the difference across HRIR conditions, particularly the differences between natural ITD+ILD, ILD-only and ITD-only conditions. In other words, each model component is able to predict the effect of its associated binaural cue (ITD or ILD) and the combination of the two components leads to accurate predictions of the natural (ILD+ITD) HRIR conditions.

Predictions for the 1-voice Freq. Inc. noises always underestimate the measured SRTs. The model predicts too much advantage when the listeners were listening in gaps that were incoherent across frequency bands. Compared to the conditions with 1-voice modulated noises, the predicted SRTs are always higher, indicating that the model predicts a detrimental effect of the incoherence across frequency, but not enough to match the data. This pattern was also observed in the stBSIM predictions (Ewert et al., 2017).

Model predictions are less accurate for differences amongst the ITD-only conditions, r equals 0.66 indicates that the model is not able to predict correctly the trends in the data across masker types. The model predictions show a pattern similar to the stBSIM predictions. The SRTs for the steady-state, 1-voice and 1-voice Freq. Inc. noises are underestimated (by 1 to 2 dB), while the SRTs for the sinusoidal masker modulation are overestimated (by 1 dB).

The model predicts correctly only half of the data measured with the independent HRIR/IMBM conditions. The predictions for the steady-state and 1-voice Freq. Inc. modulated noises match the data but the predicted SRTs for the sinusoidally and 1-voice modulated noises are largely overestimated, leading to an error of 7.1 dB. The stBSIM was more accurate to predict the magnitude of the variations across masker types. Max. Err. was probably below 5 dB, occurring for the SRT measured with 1-voice Freq. Inc. modulated noises and the independent HRIR. Conversely, the stBSIM predicted higher SRTs with modulated noises for the independent IMBM conditions than for the independent HRIR conditions. Therefore, the stBSIM and the present model show limits (different for each model) to predict the influence of these types of artificial HRIRs.

6. General discussion

In the end, the present study led to a single change of the parameter values used in the original model. The original values were inspired from the literature when the model was developed (Collin and Lavandier, 2013). Only Ceiling was

roughly tested and then fixed by Cubick et al. (2018). The present study confirms that these values are indeed required for optimal predictions. The value of BU has been modified to take into account the effect of binaural sluggishness allowing to better predict the influence of the modulation rate on binaural unmasking in CM2. The influence of this revision is of course limited here because it models an effect that is not dominating in the experiments considered.

Despite the model's revision, some conditions are still not well predicted, as is the case for the effect of reverberation when it fills the masker's gaps. For instance in Fig. 6, the difference between model predictions (black lines) on each panel, which represents the dip-listening advantage, is around 1.5 dB higher than the difference in the data (black circles). Therefore, the model overestimates the dip-listening advantage even if the trends are well predicted. For the conditions without reverberation tail (grey squares and grey solid lines), the dip-listening advantage is better predicted. Hence, the current model does not fully take into account the negative effect of reverberation filling in the gaps in the masking noise. Earlier, Beutelmann et al. (2010) observed a similar behavior for their model.

The model overestimates the SRT for the co-located modulated noise in CM1 (Fig. 2) while the corresponding separated condition is well predicted. This might be explained by the predictability of the dip occurrences within the masker. Fogerty et al. (2018) turned on and off a noise at different rates, with a 50% duty cycle, roughly resulting in a masker modulated by a square wave at different rates. They showed that listeners were able to benefit from the predictability of the dip occurrences for gating rates below 16 Hz. Collin and Lavandier (2013) found similar results concerning the predictability of masker dips using noise modulated by a broadband speech envelope. Culling and Mansell's masker was modulated by a 10-Hz square wave in CM1, so listeners were probably able to at least partly benefit from the predictability of the masker dips. The difference observed between data and prediction could be due to this effect, which is not taken into account by the model, the parameters of which were set to predict unpredictable speech modulations (in CL1, CL4 and VL).

For the symmetrical configurations with modulated maskers on both sides of the target (e.g. in CM1, CM2 (ILD+ITD), CL4), the model predicts more better-ear glimpsing and/or binaural unmasking than measured in the experiments. Increasing the duration of the temporal window BE used for computing the SNR at the better-ear in order to simulate binaural slughisness and the across-ear switching cost did not produce better predictions in the present study. A future update of the better-ear listening model component — which could take into account binaural sluggishness along with the monaural ability for listening in the dips — could improve the predictions for these conditions.

Some other effects could have been tested in the present study and potentially added in the model, which might further improve its prediction accuracy, even if being detrimental to its simplicity. While adapting the monaural SII for fluctuating noise using temporal windows to decompose the signals, and inspiring the binaural models presented here and by Beutelmann et al. (2010), Rhebergen and Versfeld (2005) showed that their best predictions were obtained with frequency-dependent durations for the temporal window. The 12 ms-ERD used in the present model comes from their best value for a frequency-independent window, but it is only an approximation of a more complex frequency-dependent decomposition of the signals. Also, Rhebergen et al. (2006) later implemented forward masking in their model. This additional component led to better predictions in the case of a periodically modulated noise. The shape of the temporal windows used in the present model (Hann windows) could have been varied. Culling and Summerfield (1998), Moore et al. (1988), as well as Plack and Moore (1990) indicated that the shape of this window is probably asymmetric and depends on the frequency and level of the stimulus. Finally, Beutelmann et al. (2009) and Kolarik and Culling (2010) demonstrated that binaural auditory filters are probably wider than monaural auditory filters. This feature could be incorporated and tested in the proposed model, particularly for the prediction of the binaural unmasking advantage.

The revised model proposed here provides predictions similar to the original models it is based on (Lavandier and Culling, 2010; Jelfs et al., 2011; Lavandier

et al., 2012; Collin and Lavandier, 2013) and other models proposed in the literature (Beutelmann et al., 2010; Wan et al., 2014), with a r ranging from 0.85 to 0.96 (across experiments) and a Mean Err. between 0.5 and 1.4 dB. The value of only one parameter was changed compared to the model of Collin and Lavandier (2013). This change allows to take at least partly into account the effect of binaural sluggishness on binaural unmasking. More importantly, all model parameters have been thoroughly tested, and it was demonstrated that the parameter values proposed are those giving the best results. The model has been validated on three speech corpora (German, English and French), in anechoic and reverberant rooms, in the presence of different number of maskers and different types of masker modulations (steady-state, speech modulated or periodically modulated), with maskers placed at various azimuths and distances from the listener. In total, 60 conditions (CM1, CM2, CL1, CL4, VL) were used to set the value of the four model parameters, 20 of which (CL1, CL4) were previously used to validate the original version of the model. The revised model, using the new set of parameter values, was validated with an additional 40 conditions (Ewert et al., 2017). Thus, the robustness of the model has been improved through this study and it is more in line with the literature by the implementing a window accounting for the binaural sluggishness.

The proposed model is available to the community. A code can be downloaded here: https://mathieulavandier.wordpress.com/home/models/.

Acknowledgments

The authors would like to thank the reviewers for their helpful comments. The authors are grateful to John Culling and Stephan Ewert for sharing their data, and to Emmanuel Gourdon for his valued advice on the sensitivity analysis. This work was performed within the LabEx CeLyA (ANR-10-LABX-0060/ANR-16-IDEX-0005). This paper was written while TV was funded by the “Fondation Pour l’Audition” (Speech2Ears grant).

Appendix A. Description of the experiment (VL) conducted in the present study

Aim

VL employed modulated maskers, different levels of reverberation, in an asymmetrical configuration. It also aimed at emphasizing the contribution of better ear listening to speech intelligibility. Hence, stimuli with and without ITDs were considered using binaural room impulse responses (BRIRs) and spectral-envelope impulse responses (SEIRs, Lavandier et al., 2012). SEIRs were obtained by removing the ITDs and reverberation tails of the BRIRs, while preserving their long-term spectrum (at each ear). In the following, BRIRs and SEIRs are associated with the "ITD+ILD" and "no ITD/no tail" conditions, respectively. Because the stimuli in the no ITD/no tail condition did not contain reverberation tails, the influence of reverberation filling in the masker modulation gaps was varied here in an asymmetrical condition.

It was hypothesized that, for the conditions with SRM, higher SRTs should be obtained in the no ITD/no tail condition compared to the ITD+ILD condition due to the absence of ITDs/binaural unmasking. It was also hypothesized that the difference between SRTs measured with steady-state and modulated noises should be larger in the no ITD/no tail conditions, at least at large distances, when reverberation tails fill in the dips of the modulated noise so that it becomes steady-state.

Stimuli and apparatus

The stimuli were produced as done by Collin and Lavandier (2013). A male speaker uttered semantically unpredictable sentences in French that contained four key words. The anechoic recordings were used as the basis of all stimuli. The maskers were noises (SSNs) either steady-state or modulated by an envelope extracted from a speech signal (1-voice modulated noises). A long steady-state noise was obtained by concatenating several lists of sentences, taking the Fourier

transform of the resulting signal, randomizing its phase, and finally taking its inverse Fourier transform.

To create the speech-modulated noises, the envelopes of the sentences were extracted as proposed by Festen and Plomp (1990), then concatenated by pairs keeping a 100-ms silence between them. The modulated noises were obtained by multiplying these envelopes with the steady-state noises. During the test, a masker envelope was never the same as the target envelope.

Real-room listening was simulated over headphones by convolving the anechoic stimuli with the BRIRs. These BRIRs were measured by Lavandier et al. (2012) in a meeting room (meeting room 1). SEIRs were also used in order to evaluate the contributions of reverberation tails and binaural unmasking. SEIRs were designed to remove the ITDs and reverberation tails of the BRIRs, but preserve room coloration and long-term ILDs when present, since SEIRs retain the same long-term spectrum as their corresponding BRIR. Full information about the measurements and processing can be found in Lavandier et al. (2012).

The convolution by a BRIR can introduce level differences in the resulting signals across different positions. To avoid these level effects, the left-right average of the RMS power of the convolved stimuli was equalized before the experiment, i.e., the levels of the spatialized stimuli were equalized at the ears of the listeners while preserving the ILDs.

Signals were digitally mixed, D/A converted, and amplified using a Lynx TWO sound card. They were presented to listeners over Sennheiser HD 650 headphones in a double-walled soundproof booth. A computer screen was visible outside the booth window. A keyboard was inside the booth to gather the transcripts.

Design

The target was simulated at 0.65 m and +25° from the listener. Two types of noise (steady-state or 1-voice modulated) were tested at two distances, 0.65 m and 5 m referred to as “Near” and “Far”. Two noise azimuths were also tested, one identical (+25°) and one different (-25°) from the one of the tar-

get (+25°). The target (+25°/Near) was presented against a single noise in each condition. The combination of all these experimental factors resulted in 16 conditions (NOISE MODULATION{steady-state, modulated} x NOISE DISTANCE{near, far} x NOISE AZIMUTH{= target, ≠ target} x IMPULSE RESPONSE TYPE{itd+ild, no itd/no tail}).

Procedure

The adaptive procedure used to measure the SRTs was similar to the one used by Collin and Lavandier (2013), except that Collin and Lavandier varied the target level and kept the noise level constant to control the SNR, but the overall sound level varied during the measurements. In the current experiment the overall level was fixed at 70 dB SPL (calibrated using a MK2/NCF1 dummy head, Neutrik Cortex Instrument), and instead of applying formula 1 in Collin and Lavandier (2013) to the target level as they did, it was applied here on the SNR.

The results of a listener were discarded from the data if there was no inversion in the adaptive procedure to measure a SRT. It occurred only once during the experiment. Another listener was enrolled to substitute the participant whose results were discarded.

Listeners

Seventeen French native speakers participated in the experiment. The data of one participant was discarded because one SRT measurement failed (see previous section). All participants had an hearing threshold equal to or better than 20 dB HL from 125 Hz to 8 000 Hz. None of them was familiar with the speech material. All provided written informed consent and were paid for their participation.

Results

Figure 6 presents the SRTs measured in VL, averaged across listeners and plotted as a function of the noise position, with one panel for each type of noise

modulation. There was no difference between the no ITD/no tail and ILD+ITD conditions for the steady-state noise. For the modulated noise, the SRTs were lower for the no ITD/no tail condition than for the ITD+ILD condition. For both types of noise modulation and impulse response, when the masker was spatially separated from the target, listeners had SRMs of at least 4 dB. In the "far" conditions, SRMs were lower but at least 2 dB.

A repeated-measure analysis of variance (ANOVA) confirmed significant effects of the impulse response type [$F(1,15) = 15.5$, $p = 0.001$], the noise distance [$F(1,15) = 35.0$, $p < 0.0001$], the noise modulation [$F(1,15) = 238.4$, $p < 0.0001$] and the noise azimuth [$F(1,15) = 195.4$, $p < 0.0001$]. Two interactions were significant, between the impulse response type and the noise modulation [$F(1,15) = 38.8$, $p < 0.0001$] and between the noise distance and azimuth [$F(1,15) = 188.6$, $p < 0.0001$].

Discussion

SRTs in the modulated noise are consistently lower than those obtained in steady-state noise, i.e., listeners benefited from the masker gaps. A 5-dB SRM can be observed at near distance when the masker was moved from the =target-near position (co-located configuration) to the ≠target-near position. The SRM was reduced to about 2 dB in the far conditions (resulting in the interaction of the effects of masker distance and azimuth). This can be explained by the increased effect of reverberation on the masker (Lavandier et al., 2012), which impairs both better-ear listening (by reducing head-shadow) and binaural unmasking (by decorrelating the masker at the two ears). SRTs were on average lower in the far conditions compared to the near conditions, probably highlighting a beneficial effect of room coloration in this particular configuration, as already observed previously (Lavandier et al., 2012; Collin and Lavandier, 2013).

Two effects associated with the no ITD/no tail condition may account for the interaction of the effects of noise modulation and impulse response. Stimuli without ITDs prevent listeners to use binaural unmasking, hence impair intel-

ligibility. Under no ITD/no tail conditions reverberation does not fill in the masker dips. This allows listeners to use dip listening, thus enhancing speech intelligibility. These two counteracting effects could explain the difference between the ILD+ITD and the no ITD/no tail data for the modulated maskers in the bottom panel of Fig. 6 (black symbols vs. grey symbols). SRTs are lower in the no ITD/no tail condition compared to the ILD+ITD condition, suggesting that the positive effect of having no reverberation tails is stronger that the negative effect of removing binaural unmasking. The difference is about 1 dB for the near distance and 2 dB for the far distance. The 1-dB intelligibility enhancement, observed when the masker distance increased, is consistent with the following explanation: when the masker is further away from the listener, it has more energy in its reverberation tails. It fills the masker gaps more, hence triggers a larger difference than under the (no ITD/no tail) SEIR conditions.

With the steady-state noise, no effect of the reverberation tails was expected. The difference between ITD+ILD and no ITD/no tail conditions should be limited to the involvement of binaural unmasking in the ITD+ILD condition. This should have led to lower SRTs compared to the no ITD/no tail. No significant effect of binaural unmasking was observed here at near distance even if it was previously observed using the same impulse responses (Lavandier et al., 2012). That effect was limited to about 1 dB, which could explain its lack of significance in the present experiment.

References

ANSI S3.5, 1997. Methods for calculation of the speech intelligibility index. American National Standards Institute, New York .

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

Beutelmann, R., Brand, T., Kollmeier, B., 2010. Revision, extension, and evalu-

ation of a binaural speech intelligibility model. The Journal of the Acoustical Society of America 127, 2479–2497.

Brungart, D.S., Iyer, N., 2012. Better-ear glimpsing efficiency with symmetrically-placed interfering talkers. The Journal of the Acoustical Society of America 132, 2545–2556.

Collin, B., Lavandier, M., 2013. Binaural speech intelligibility in rooms with variations in spatial location of sources and modulation depth of noise interferers. The Journal of the Acoustical Society of America 134, 1146–1159.

Cubick, J., Buchholz, J.M., Best, V., Lavandier, M., Dau, T., 2018. Listening through hearing aids affects spatial perception and speech intelligibility in normal-hearing listeners. The Journal of the Acoustical Society of America 144, 2896–2905.

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)]. The Journal of the Acoustical Society of America 118, 552.

Culling, J.F., Mansell, E.R., 2013. Speech intelligibility among modulated and spatially distributed noise sources. The Journal of the Acoustical Society of America 133, 2254–2261.

Culling, J.F., Summerfield, Q., 1998. Measurements of the binaural temporal window using a detection task. The Journal of the Acoustical Society of America 103, 3540–3553.

Durlach, N.I., 1972. Binaural signal detection: Equalization and cancellation theory, in: Tobias, J. (Ed.), Foundations of Modern Auditory Theory. Academic, New York. volume II, pp. 371–462.

Ewert, S.D., Schubotz, W., Brand, T., Kollmeier, B., 2017. Binaural masking release in symmetric listening conditions with spectro-temporally modulated maskers. The Journal of the Acoustical Society of America 142, 12–28.

Festen, J.M., Plomp, R., 1990. Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. The Journal of the Acoustical Society of America 88, 1725–1736.

Fogerty, D., Carter, B.L., Healy, E.W., 2018. Glimpsing speech in temporally and spectro-temporally modulated noise. The Journal of the Acoustical Society of America 143, 3047–3057.

Grantham, D.W., 1982. Detectability of time-varying interaural correlation in narrow-band noise stimuli. J. Acoust. Soc. Am. 72, 1178.

Grantham, D.W., Wightman, F.L., 1979. Detectability of a pulsed tone in the presence of a masker with time-varying interaural correlation. The Journal of the Acoustical Society of America 65, 1509–1517.

Hauth, C.F., Brand, T., 2018. Modeling sluggishness in binaural unmasking of speech for maskers with time-varying interaural phase differences. Trends in Hearing 22, 1–10.

Jelfs, S., Culling, J.F., Lavandier, M., 2011. Revision and validation of a binaural model for speech intelligibility in noise. Hearing Research 275, 96–104.

Kolarik, A.J., Culling, J.F., 2010. Measurement of the binaural auditory filter using a detection task. The Journal of the Acoustical Society of America 127, 3009–3017.

Kryter, K.D., 1962. Methods for the Calculation and Use of the Articulation Index. The Journal of the Acoustical Society of America 34, 1689.

Lavandier, M., Best, V., in press, 2020. Modeling binaural speech understanding in complex situations, in: Blauert, J., Braasch, J. (Eds.), The technology of binaural understanding, chapter 19. Springer, Berlin–Heidelberg–New York NY.

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

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. The Journal of the Acoustical Society of America 131, 218–231.

Moore, B.C., Glasberg, B.R., Plack, C.J., Biswas, a.K., 1988. The shape of the ear's temporal window. The Journal of the Acoustical Society of America 83, 1102–1116.

Moore, B.C.J., Glasberg, B.R., 1983. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. The Journal of the Acoustical Society of America 74, 750–753.

Patterson, R.D., Nimmo-Smith, I., Holdsworth, J., Rice, P., 1987. An efficient auditory filterbank based on the gammatone function. presented to the Institute of Acoustics speech group on auditory modelling at the Royal Signal Research Establishment.

Plack, C.J., Moore, B.C.J., 1990. Temporal window shape as a function of frequency and level. The Journal of the Acoustical Society of America 87, 2178–2187.

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. The Journal of the Acoustical Society of America 117, 2181–2192.

Rhebergen, K.S., Versfeld, N.J., Dreschler, W.A., 2006. Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. The Journal of the Acoustical Society of America 120, 3988–3997.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., Tarantola, S., 2010. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. Computer Physics Communications 181, 259–270.

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

Studebaker, G.A., Sherbecoe, R.L., 2002. Intensity-importance functions for bandlimited monosyllabic words. The Journal of the Acoustical Society of America 111, 1422–1436.

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. The Journal of the Acoustical Society of America 136, 768–776.