



HAL
open science

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

Thibault Vicente, Mathieu Lavandier

► To cite this version:

Thibault Vicente, Mathieu Lavandier. Further validation of a binaural model predicting speech intelligibility against envelope-modulated noises. *Hearing Research*, 2020, 390, pp.107937. 10.1016/j.heares.2020.107937 . hal-02631613

HAL Id: hal-02631613

<https://hal.science/hal-02631613>

Submitted on 22 Aug 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

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

3 Thibault Vicente^{a,*}, Mathieu Lavandier^a

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

6 **Abstract**

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

22 *Keywords:* Auditory Modeling, Binaural Perception, Speech Intelligibility

23 **1. Introduction**

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

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

February 25, 2020

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

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

59 Mansell, 2013; Grantham and Wightman, 1979; Hauth and Brand, 2018). In
60 contrast, the time constant usually used to describe the monaural system is
61 about 8-13 ms (Moore et al., 1988; Plack and Moore, 1990). Culling and Mansell
62 (2013) provided evidence that better-ear glimpsing could be “truly” binaural
63 and rely on switching across ears, since they found that performance was highly
64 dependent on the required switching rate, and that this binaural switching could
65 be rather sluggish.

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

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

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

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

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

103 **2. Data**

104 *2.1. Data sets used to test the model parameter*

105 Five experiments were used to test the model parameters. The experiments
106 1 and 2 of Culling and Mansell (2013) are abbreviated CM1 and CM2, the exper-
107 iments 1 and 4 of Collin and Lavandier (2013) are CL1 and CL4, the experiment
108 run in the present study is VL. A summary of the design of each experiment
109 is presented in Table 1, for more details referred to the related publications
110 (Culling and Mansell, 2013; Collin and Lavandier, 2013; and Appendix A, re-
111 spectively). The “co-located” condition will refer to the configuration where
112 target and noise(s) are at the same spatial position, otherwise the configuration
113 will be referred to as “separated”. Positive azimuths correspond to the right side
114 of the listeners. All the noises used as masking sources had the same long-term
115 spectrum as the target speech (speech-shaped noises, SSNs).

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

Exp.	Number of Noises	Noise modulation	Noise Distance in room	Noise Azimuth	Cues available
<i>CM1</i>	1 or 2	Steady-state or modulated (10-Hz square wave, 50% duty cycle)	Anechoic	$0^{\circ T}$, 105° or $\pm 105^{\circ}$	ITD+ILD
<i>CM2</i>	1 or 2	Steady-state (0 Hz) or modulated (1-, 2-, 5-, 10-, or 20-Hz square wave, 50% duty cycle)	Anechoic	$0^{\circ T}$ or $\pm 105^{\circ}$	ITD+ILD, ILD-only or ITD-only
<i>CL1</i>	1	Steady-state or modulated (broadband envelope of 1, 2 or 4 voices)	0.65^T , 1.25 or 5 m	$0^{\circ T}$	ITD+ILD
<i>CL4</i>	1 or 2	Steady-state or modulated (broadband envelope of 1 or 2 voices)	0.65^T m	$0^{\circ T}$, 25° or $\pm 25^{\circ}$	ITD+ILD
<i>VL</i>	1	Steady-state or modulated (broadband envelope of 1 voice)	0.65^T or 5 m	$25^{\circ T}$ 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.

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

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

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

140 2.2. Data set used to validate the revised model

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

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

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

167 **3. Model description**

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

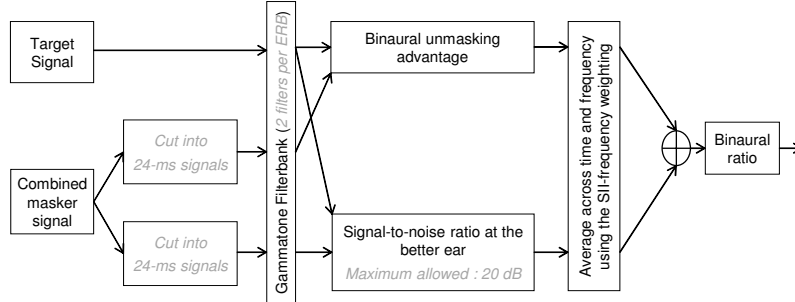


Figure 1: Block diagram of the original model (Collin and Lavandier, 2013), with the parameters tested in the present study highlighted in grey.

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

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

197 better-ear and binaural unmasking components within each time frame. Thus,
198 as a model input, the target sentences at the ears are replaced by an averaged
199 target signal generated by adding at least¹ 60 target sentences (truncated to the
200 duration of the shortest sentence), and this averaged signal is not submitted to
201 the temporal decomposition into short-time frames used for the masker.

202 The masker signals are cut into frames using half-overlapping Hann win-
203 dows, before being passed through a Gammatone filterbank (Patterson et al.,
204 1987) with two filters per equivalent rectangular bandwidth (ERB ; Moore and
205 Glasberg, 1983). The bandwidth of the Gammatone filters is about 1 ERB,
206 thus the filters are half-overlapping. Within each time frame and frequency
207 band, the two components of spatial unmasking are modeled, (1) the binaural
208 unmasking advantage is estimated using a formula proposed by Culling et al.
209 (2005), which depends on the masker interaural coherence and on the target
210 and masker interaural phase differences. The target and masker signals are
211 both cross-correlated to derive these interaural parameters. The coherence is
212 taken as the maximum of the cross-correlation function, and the phase differ-
213 ence is obtained by multiplying the corresponding delay by the center frequency
214 of the band. The search of maximum delay in the cross-correlation functions is
215 limited to the range plus/minus half the period of the channel center frequency,
216 so that the model does not predict any binaural unmasking advantage at high
217 frequency (Durlach, 1972). The binaural unmasking advantage is set to zero if
218 the masking noise power is zero at one of the ears in the considered band and
219 frame. (2) The SNR is also computed at each ear, and the best SNR across ear
220 is selected (thus independently for each frequency band and each time frame).
221 A ceiling parameter corresponding to the maximum better-ear ratio allowed by
222 frequency band and time frame is introduced at this stage, to avoid the SNR

¹For 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.

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

229 The first aim of the present study was to test the four parameters of the
230 model (see Fig. 3): the duration of the Hann window used for computing the
231 binaural unmasking advantage (“BU” in ms), the duration of the Hann window
232 used for computing the better-ear SNR (“BE” in ms) — those are the two
233 temporal resolutions of the model — the number of gammatone filters per ERB
234 (the model spectral sampling “SpecSamp”) and finally the ceiling parameter
235 (“Ceiling” in dB). The parameter values were previously set to 24 ms, 24 ms,
236 2 filters per ERB and 20 dB, respectively (Collin and Lavandier, 2013; Cubick
237 et al., 2018). In particular, the same temporal resolution was used to model
238 better-ear listening and binaural unmasking; whereas the temporal resolution
239 of the two mechanisms (and their susceptibility to binaural sluggishness) was
240 investigated independently here.

241 4. Revision of the model

242 4.1. A method inspired by a sensitivity analysis

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

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

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

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

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

283 experiment, the independent parameters were varied independently to define the
284 value(s) leading to the best predictions, whereas for the interacting parameters,
285 these values were defined while varying the parameters simultaneously. The
286 best predictions values were then compared across experiments in order to find
287 a single common value for each parameter leading to good predictions across
288 all experiments. A qualitative analysis of the predictions was also considered,
289 to eventually help define the final parameter values if there were more than a
290 single value leading to best predictions across experiments. Values leading to
291 predictions conceptually wrong (e.g., missing a basic effect observed in the data)
292 were excluded prior to this analysis. For each type of parameter, independent
293 or interacting, if its original value (Collin and Lavandier, 2013; Cubick et al.,
294 2018) was among the values leading to the best predictions, then this value
295 was selected for the parameter because there was no relevant argument for a
296 change. The definition of the best parameter values was done while keeping in
297 mind which parameters were the most influential.

298 *4.2. Results*

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

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

Exp.	CM1	CM2	CL1	CL4	VL
1st order indices (%)					
<i>Ceiling</i>	58	21	48	43	68
<i>BE</i>	20	54	28	8	12
<i>BU</i>	1	3	0	0	18
<i>SpecSamp</i>	0	0	0	15	0
<i>Sum of 1st order indices</i>	79	78	76	66	98
2nd order indices (%)					
<i>Ceiling/BE</i>	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.

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

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

323 None of the experiments were discriminating to choose the SpecSamp value,
324 in agreement with the fact that its first order sensitivity index was equal to 0%
325 for four experiments. This parameter had some limited influence only for the
326 predictions of CL1, but in practice all values led to accurate predictions. There-
327 fore, the original value used by Collin and Lavandier (2013) was kept unchanged
328 (2 Gammatone filters per ERB, see section 4.4 concerning this choice).

329 Concerning BU, the predictions for CL4 and CL1 were not affected by chang-
330 ing BU values. For CM1 and VL, the longer the window duration the lower Mean
331 Err. ; and for CM2 the lowest Mean Err. was reached with the longest window
332 duration. Hence, the analysis suggests to change the value of BU from 24 to
333 400 ms (from 12 to 200 ms ERD). It was decided to set BU to 300 ms (150
334 ms ERD), a value not tested above but which corresponds to the midst of the
335 binaural temporal windows reported in the literature (Culling and Summerfield,
336 1998; Culling and Mansell, 2013; Grantham and Wightman, 1979; Hauth and
337 Brand, 2018), which ranged from 80 ms to 500 ms (from 40 to 250 ms ERD, see
338 Introduction). The difference in predicted SRT using a BU of 300 rather than
339 400 ms was below 0.1 dB in each of the five experiments considered above.

340 The values for Ceiling and BE giving the best predictions were deduced by
341 removing the values leading to inconsistent predictions. The tested values for
342 Ceiling were: 8, 12, 16, 20, 24 dB ; for BE, they were: 16, 24, 80, 200, 400 ms
343 (or 8, 12, 40, 100, 200 ms ERD). In CM1 and CM2, when BE was set to 80,
344 200 or 400 ms the model predicted SRTs with obvious deviation from the data
345 for all values of Ceiling (e.g. no difference in predicted SRT for stationary and
346 modulated noises in CM1). These prediction errors are considerably reduced
347 with the shortest window durations, so that only the values 16 and 24 ms (8
348 and 12 ms ERD) remained for BE.

349 Model predictions for CM1 and CM2 led to conflicting results concerning
350 the choice of Ceiling. The best predictions for CM1 were obtained for values
351 equal to 20 or 24 dB, whereas the best predictions for CM2 were obtained for a
352 Ceiling of 8 dB. This value was not considered further, because CM2 is the only
353 experiment well predicted with a Ceiling of 8 dB. A Ceiling of 12 dB led a 3.2-dB
354 overestimation of the SRTs in the conditions with one modulated noise in CM1.
355 In this case, the model also predicted identical SRTs for the steady-state and
356 modulated noise in the separated condition. The best model performances for
357 CL1 were reached with a 12-dB Ceiling. Because the value of 12 dB also led to
358 conflicting results between CL1 and CM1, it was not considered further. The
359 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.
<i>CM1</i>	0.93 ; 0.96	1.3 ; 1.0	1.6 ; 1.3	3.5 ; 2.3
<i>CM2</i>	0.92 ; 0.94	1.0 ; 1.0	1.2 ; 1.0	2.4 ; 1.8
<i>CL1</i>	0.85 ; 0.85	0.5 ; 0.5	0.6 ; 0.6	1.3 ; 1.3
<i>CL4</i>	0.92 ; 0.93	0.6 ; 0.6	0.8 ; 0.8	1.6 ; 1.4
<i>VL</i>	0.87 ; 0.90	1.0 ; 0.9	1.0 ; 1.2	2.0 ; 1.8
<i>Ewert et al.</i> <i>(2017)</i>	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.

360 dB.

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

365 4.3. Predictions of the revised model

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

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

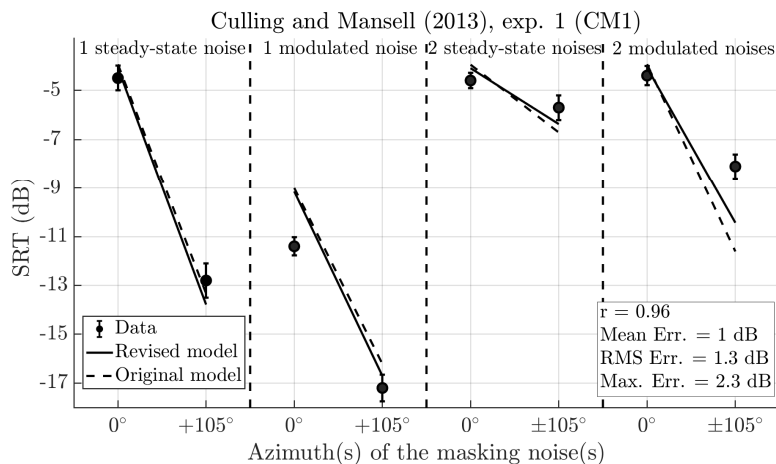


Figure 2: Mean SRTs with standard errors across listeners measured in CM1, involving 1 or 2 noises, steady-state or modulated by a 10-Hz square wave (50% duty cycle, modulated out-of-phase if they were two maskers), simulated as originating from different azimuths (0° and $+105^\circ$ or $\pm 105^\circ$ 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.

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

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

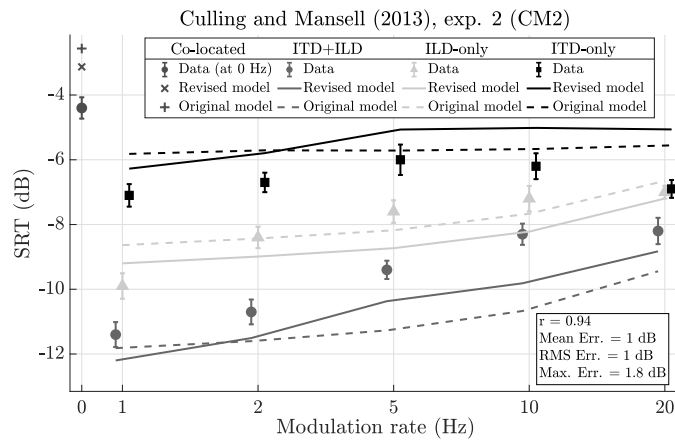


Figure 3: Mean SRTs with standard errors across listeners measured in CM2. The target was always presented at 0° in the presence of two noises placed on both sides of the listener ($\pm 105^\circ$). The noises were modulated out-of-phase by a square wave at 5 modulation rates (1, 2, 5, 10, 20 Hz). Three 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.

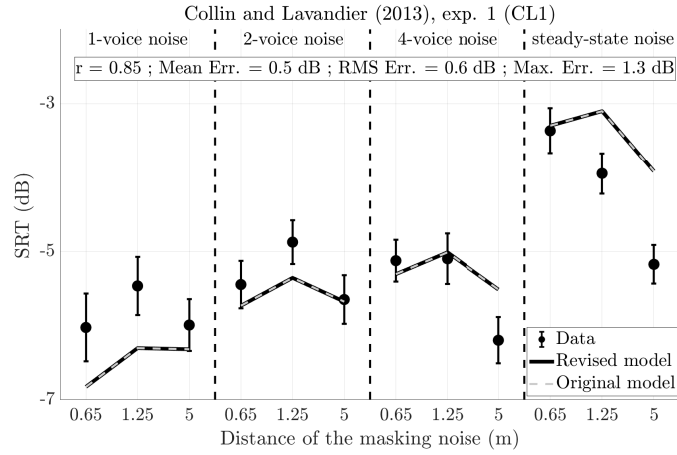


Figure 4: Mean SRTs with standard errors across listeners measured in CL1. The target was at 0.65 m in front of the listener in a lecture hall. The noise was placed at three distances (0.65, 1.25, 5 m), also in front of the listener. Four types of modulation were used for the noise (steady-state, 1-, 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.

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

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

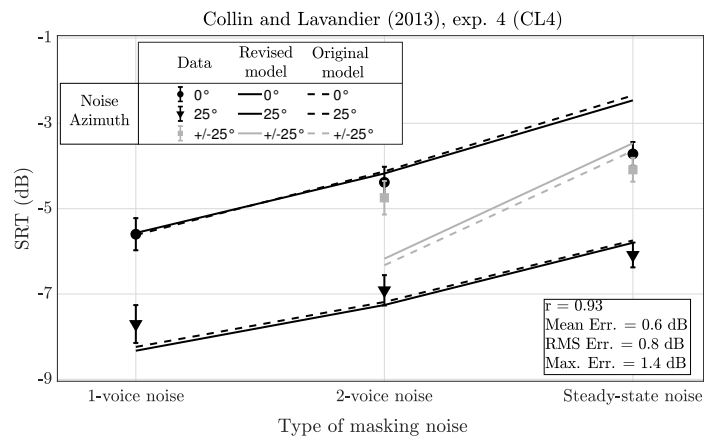


Figure 5: Mean SRTs with standard errors across listeners measured in CL4. The target was at 0.65 m in front of the listener in a meeting room. The single masker was always at 0.65 m but tested at two azimuths (0° and 25°). Three types of noise were involved (1-voice modulated, 2-voice modulated or steady-state). Two noises (steady-state or two 1-voice modulated) were tested in two configurations (0° or $\pm 25^\circ$, 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.

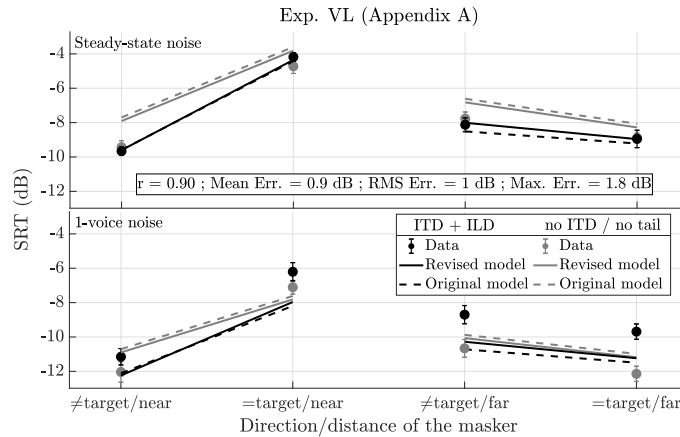


Figure 6: Mean SRTs with standard errors across listeners measured in the present study (VL). The target was placed at 0.65 m, +25° from the listener (=target/near). The noise was steady-state (top panel) or 1-voice modulated (bottom panel). It was tested at two distances (near at 0.65 m and far at 5 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.

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

409 4.4. Discussion

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

417 analysis.

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

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

448 than the range induced by the BE variations.

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

456 Hauth and Brand (2018) investigated the effect of binaural sluggishness us-
457 ing a different short-time binaural speech intelligibility model (stBSIM ; Beutel-
458 mann et al., 2010). They designed an experiment in order to test the effect
459 of binaural sluggishness on speech intelligibility. Stimuli were a steady-state
460 noise for which interaural phase differences (IPDs) were modulated sinusoidally
461 between $-\pi/2$ and $+\pi/2$ at different rates between 0 and 64 Hz. Increasing
462 the modulation rate led to higher SRTs for rates up to 4 Hz, above which the
463 rate had no significant influence on the SRT. These results are consistent with
464 the results of CM2. When modeling their own experiment, Hauth and Brand
465 modified the EC processing of their model to introduce binaural sluggishness
466 that influences the definition of the EC parameter. However, the EC stage *per*
467 *se* is still applied on short-time signals (for detail of implementation, see Hauth
468 and Brand, 2018). In the current model, the binaural unmasking advantage is
469 estimated using signals whose duration is influenced by binaural sluggishness,
470 resulting in longer signals than in the “revised” stBSIM. Despite this discrep-
471 ancy of binaural sluggishness implementation in the models, the duration of
472 the binaural/EC window proposed by Hauth and Brand allowing to predict ac-
473 curately their data was 200 ms (ERD), which is similar to the duration (BU)
474 highlighted in the current study.

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

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

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

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

502 Regarding the Ceiling value, it has been set to 20 dB that is higher than the
503 values implemented in the SII (ANSI S3.5, 1997) or in the AI (Kryter, 1962),
504 +15 and +18 dB SNR, respectively. It means that the current model considers
505 that the full target intelligibility is reached at a higher SNR. Studebaker and
506 Sherbecoe (2002) showed that increasing the SNR up to 29 dB could still improve
507 target intelligibility. Such a high value of Ceiling does not seem appropriate
508 in the proposed model. Collin and Lavandier (2013) introduced Ceiling to the
509 model. They found that a value of 10 or 15 dB reduced prediction errors. Cubick

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

513 The values of the window duration BE used to compute the better-ear com-
514 ponent of the model that predicted well most conditions tested here were 16
515 and 24 ms (8 and 12 ms ERD). These values are within the range 8 to 13 ms
516 (ERD) of the measured monaural temporal resolution (Moore et al., 1988; Plack
517 and Moore, 1990). The tested binaural window durations (i.e. 80, 200, 400 ms
518 or 40, 100, 200 ms ERD) provided inconsistent predictions in some conditions.
519 The model was not able to predict the advantage of listening in the masker dips
520 when the temporal resolution of the better-ear component was not sufficient
521 (i.e. when the window duration was too long). For instance, in CM1 a BE of
522 200 ms provided the same predicted SRT for the steady-state and modulated
523 noises. Taking a too-long temporal window triggers an amplitude modulation
524 smoothing in the model, so that the modulated masker appeared as a steady-
525 state masker. Consequently, the window duration BE has to match a monaural
526 time constant. It should be noted that Collin and Lavandier (2013) as well as
527 Beutelmann et al. (2010) also used a temporal resolution of 24 ms. It corre-
528 sponds to the best frequency-independent duration used in the monaural model
529 of Rhebergen and Versfeld (2005).

530 The values retained for BE differed from Culling and Mansell’s conclusion,
531 which stated that better-ear listening is a mechanism affected by binaural slug-
532 gishness, because in CM2 there was an influence of the required ear-switching
533 rate up to 5 Hz. Although a monaural time window is required for the proposed
534 model, it does not mean that better-ear listening is a “double” monaural mecha-
535 nism, which is not influenced by binaural sluggishness and across-ear switching.
536 It just means that the current implementation of the model does not allow for
537 predicting this effect of sluggishness on better-ear listening. Culling and Mansell
538 (2013) concluded that better-ear listening is binaural because the listener has
539 to choose which ear is more beneficial for listening to the target; but also that
540 the monaural behavior of each ear allows for listening in the dips. So there may

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

547 **5. Validation of the revised model**

548 *5.1. Predictions*

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

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

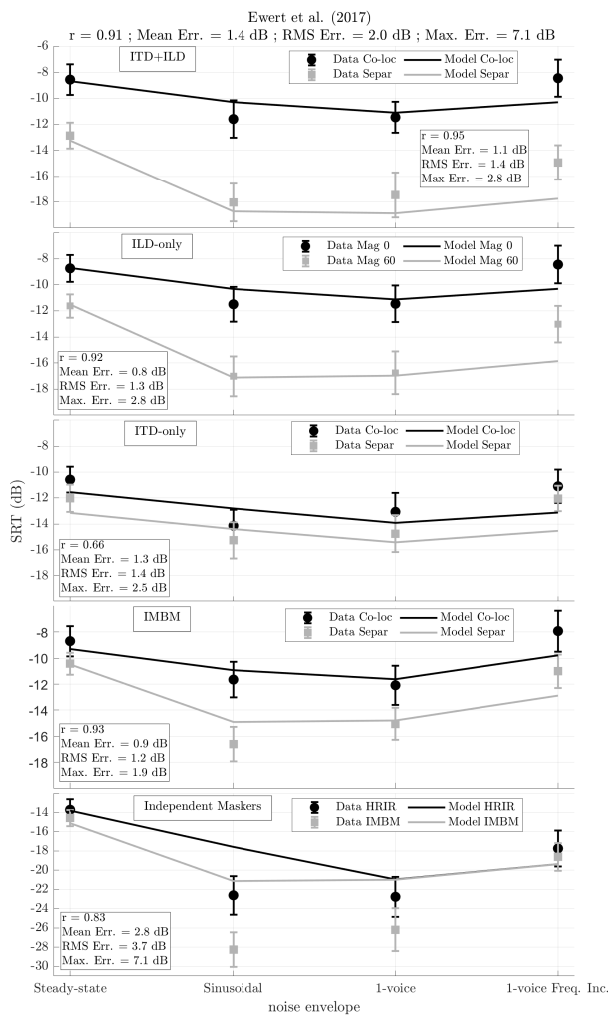


Figure 7: Mean SRTs with standard deviations across listeners measured by Ewert et al. (2017). SRTs are plotted as a function of the noise modulation type (steady-state, sinusoidal, 1-voice, 1-voice frequency incoherent). Each panel corresponds to a given type of HRIRs (ITD+ILD, ILD-only, ITD-only, IMBM, 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 ($\pm 60^\circ$) 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.

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

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

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

587 *5.2. Discussion*

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

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

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

612 The model predicts correctly only half of the data measured with the in-
613 dependent HRIR/IMBM conditions. The predictions for the steady-state and
614 1-voice Freq. Inc. modulated noises match the data but the predicted SRTs for
615 the sinusoidally and 1-voice modulated noises are largely overestimated, leading
616 to an error of 7.1 dB. The stBSIM was more accurate to predict the magnitude
617 of the variations across masker types. Max. Err. was probably below 5 dB, oc-
618 ccurring for the SRT measured with 1-voice Freq. Inc. modulated noises and the
619 independent HRIR. Conversely, the stBSIM predicted higher SRTs with mod-
620 ulated noises for the independent IMBM conditions than for the independent
621 HRIR conditions. Therefore, the stBSIM and the present model show limits
622 (different for each model) to predict the influence of these types of artificial
623 HRIRs.

624 **6. General discussion**

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

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

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

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

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

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

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

690 et al., 2012; Collin and Lavandier, 2013) and other models proposed in the
691 literature (Beutelmann et al., 2010; Wan et al., 2014), with a r ranging from
692 0.85 to 0.96 (across experiments) and a Mean Err. between 0.5 and 1.4 dB. The
693 value of only one parameter was changed compared to the model of Collin and
694 Lavandier (2013). This change allows to take at least partly into account the
695 effect of binaural sluggishness on binaural unmasking. More importantly, all
696 model parameters have been thoroughly tested, and it was demonstrated that
697 the parameter values proposed are those giving the best results. The model
698 has been validated on three speech corpora (German, English and French), in
699 anechoic and reverberant rooms, in the presence of different number of maskers
700 and different types of masker modulations (steady-state, speech modulated or
701 periodically modulated), with maskers placed at various azimuths and distances
702 from the listener. In total, 60 conditions (CM1, CM2, CL1, CL4, VL) were
703 used to set the value of the four model parameters, 20 of which (CL1, CL4)
704 were previously used to validate the original version of the model. The revised
705 model, using the new set of parameter values, was validated with an additional
706 40 conditions (Ewert et al., 2017). Thus, the robustness of the model has been
707 improved through this study and it is more in line with the literature by the
708 implementing a window accounting for the binaural sluggishness.

709 The proposed model is available to the community. A code can be down-
710 loaded here: <https://mathieulavandier.wordpress.com/home/models/>.

711 **Acknowledgments**

712 The authors would like to thank the reviewers for their helpful comments.
713 The authors are grateful to John Culling and Stephan Ewert for sharing their
714 data, and to Emmanuel Gourdon for his valued advice on the sensitivity anal-
715 ysis. This work was performed within the LabEx CeLyA (ANR-10-LABX-
716 0060/ANR-16-IDEX-0005). This paper was written while TV was funded by
717 the “Fondation Pour l’Audition” (Speech2Ears grant).

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

720 *Aim*

721 VL employed modulated maskers, different levels of reverberation, in an
722 asymmetrical configuration. It also aimed at emphasizing the contribution
723 of better ear listening to speech intelligibility. Hence, stimuli with and with-
724 out ITDs were considered using binaural room impulse responses (BRIRs) and
725 spectral-envelope impulse responses (SEIRs, Lavandier et al., 2012). SEIRs were
726 obtained by removing the ITDs and reverberation tails of the BRIRs, while pre-
727 serving their long-term spectrum (at each ear). In the following, BRIRs and
728 SEIRs are associated with the “ITD+ILD” and “no ITD/no tail” conditions,
729 respectively. Because the stimuli in the no ITD/no tail condition did not contain
730 reverberation tails, the influence of reverberation filling in the masker modula-
731 tion gaps was varied here in an asymmetrical condition.

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

739 *Stimuli and apparatus*

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

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

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

753 Real-room listening was simulated over headphones by convolving the ane-
754 choic stimuli with the BRIRs. These BRIRs were measured by Lavandier et al.
755 (2012) in a meeting room (meeting room 1). SEIRs were also used in order to
756 evaluate the contributions of reverberation tails and binaural unmasking. SEIRs
757 were designed to remove the ITDs and reverberation tails of the BRIRs, but pre-
758 serve room coloration and long-term ILDs when present, since SEIRs retain the
759 same long-term spectrum as their corresponding BRIR. Full information about
760 the measurements and processing can be found in Lavandier et al. (2012).

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

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

771 *Design*

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

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

781 *Procedure*

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

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

794 *Listeners*

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

801 *Results*

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

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

810 A repeated-measure analysis of variance (ANOVA) confirmed significant ef-
811 fects of the impulse response type [$F(1, 15) = 15.5, p = 0.001$], the noise dis-
812 tance [$F(1, 15) = 35.0, p < 0.0001$], the noise modulation [$F(1, 15) = 238.4,$
813 $p < 0.0001$] and the noise azimuth [$F(1, 15) = 195.4, p < 0.0001$]. Two in-
814 teractions were significant, between the impulse response type and the noise
815 modulation [$F(1, 15) = 38.8, p < 0.0001$] and between the noise distance and
816 azimuth [$F(1, 15) = 188.6, p < 0.0001$].

817 *Discussion*

818 SRTs in the modulated noise are consistently lower than those obtained in
819 steady-state noise, i.e., listeners benefited from the masker gaps. A 5-dB SRM
820 can be observed at near distance when the masker was moved from the =target-
821 near position (co-located configuration) to the \neq target-near position. The SRM
822 was reduced to about 2 dB in the far conditions (resulting in the interaction
823 of the effects of masker distance and azimuth). This can be explained by the
824 increased effect of reverberation on the masker (Lavandier et al., 2012), which
825 impairs both better-ear listening (by reducing head-shadow) and binaural un-
826 masking (by decorrelating the masker at the two ears). SRTs were on average
827 lower in the far conditions compared to the near conditions, probably high-
828 lighting a beneficial effect of room coloration in this particular configuration,
829 as already observed previously (Lavandier et al., 2012; Collin and Lavandier,
830 2013).

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

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

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

855 **References**

856 ANSI S3.5, 1997. Methods for calculation of the speech intelligibility index.

857 American National Standards Institute, New York .

858 Beutelmann, R., Brand, T., Kollmeier, B., 2009. Prediction of binaural speech
859 intelligibility with frequency-dependent interaural phase differences. The

860 Journal of the Acoustical Society of America 126, 1359–1368.

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

- 862 ation of a binaural speech intelligibility model. *The Journal of the Acoustical*
863 *Society of America* 127, 2479–2497.
- 864 Brungart, D.S., Iyer, N., 2012. Better-ear glimpsing efficiency with
865 symmetrically-placed interfering talkers. *The Journal of the Acoustical Soci-*
866 *ety of America* 132, 2545–2556.
- 867 Collin, B., Lavandier, M., 2013. Binaural speech intelligibility in rooms with
868 variations in spatial location of sources and modulation depth of noise inter-
869 ferers. *The Journal of the Acoustical Society of America* 134, 1146–1159.
- 870 Cubick, J., Buchholz, J.M., Best, V., Lavandier, M., Dau, T., 2018. Listening
871 through hearing aids affects spatial perception and speech intelligibility in
872 normal-hearing listeners. *The Journal of the Acoustical Society of America*
873 144, 2896–2905.
- 874 Culling, J.F., Hawley, M.L., Litovsky, R.Y., 2005. Erratum: The role head-
875 induced interaural time and level differences in the speech reception threshold
876 for multiple interfering sound sources [*J. Acoust. Soc. Am.* 116, 1057 (2004)].
877 *The Journal of the Acoustical Society of America* 118, 552.
- 878 Culling, J.F., Mansell, E.R., 2013. Speech intelligibility among modulated and
879 spatially distributed noise sources. *The Journal of the Acoustical Society of*
880 *America* 133, 2254–2261.
- 881 Culling, J.F., Summerfield, Q., 1998. Measurements of the binaural temporal
882 window using a detection task. *The Journal of the Acoustical Society of*
883 *America* 103, 3540–3553.
- 884 Durlach, N.I., 1972. Binaural signal detection: Equalization and cancellation
885 theory, in: Tobias, J. (Ed.), *Foundations of Modern Auditory Theory*. Aca-
886 *ademic, New York*. volume II, pp. 371–462.
- 887 Ewert, S.D., Schubotz, W., Brand, T., Kollmeier, B., 2017. Binaural masking
888 release in symmetric listening conditions with spectro-temporally modulated
889 maskers. *The Journal of the Acoustical Society of America* 142, 12–28.

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

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

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

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

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

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

906 Kolarik, A.J., Culling, J.F., 2010. Measurement of the binaural auditory filter
907 using a detection task. *The Journal of the Acoustical Society of America* 127,
908 3009–3017.

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

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

915 Lavandier, M., Culling, J.F., 2010. Prediction of binaural speech intelligibility
916 against noise in rooms. *The Journal of the Acoustical Society of America* 127,
917 387–399.

918 Lavandier, M., Jelfs, S., Culling, J.F., Watkins, A.J., Raimond, A.P., Makin,
919 S.J., 2012. Binaural prediction of speech intelligibility in reverberant rooms
920 with multiple noise sources. *The Journal of the Acoustical Society of America*
921 131, 218–231.

922 Moore, B.C., Glasberg, B.R., Plack, C.J., Biswas, a.K., 1988. The shape of the
923 ear’s temporal window. *The Journal of the Acoustical Society of America* 83,
924 1102–1116.

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

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

932 Plack, C.J., Moore, B.C.J., 1990. Temporal window shape as a function of
933 frequency and level. *The Journal of the Acoustical Society of America* 87,
934 2178–2187.

935 Rhebergen, K.S., Versfeld, N.J., 2005. A Speech Intelligibility Index-based ap-
936 proach to predict the speech reception threshold for sentences in fluctuating
937 noise for normal-hearing listeners. *The Journal of the Acoustical Society of*
938 *America* 117, 2181–2192.

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

942 Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., Tarantola, S.,
943 2010. Variance based sensitivity analysis of model output. Design and estima-
944 tor for the total sensitivity index. *Computer Physics Communications* 181,
945 259–270.

- 946 Steeneken, H.J.M., Houtgast, T., 1980. A physical method for measuring speech-
947 transmission quality. *The Journal of the Acoustical Society of America* 67,
948 318–326.
- 949 Studebaker, G.A., Sherbecoe, R.L., 2002. Intensity-importance functions for
950 bandlimited monosyllabic words. *The Journal of the Acoustical Society of*
951 *America* 111, 1422–1436.
- 952 Wan, R., Durlach, N.I., Colburn, H.S., 2014. Application of a short-time version
953 of the Equalization-Cancellation model to speech intelligibility experiments
954 with speech maskers. *The Journal of the Acoustical Society of America* 136,
955 768–776.