Acta Acustica 2022, **6**, 20
© The Author(s), Published by EDP Sciences, 2022
https://doi.org/10.1051/aacus/2022017

Available online at:
https://acta-acustica.edpsciences.org

Topical Issue - Auditory models: from binaural processing to multimodal cognition

TECHNICAL & APPLIED ARTICLE

OPEN ACCESS

A series of SNR-based speech intelligibility models in the Auditory Modeling Toolbox

Mathieu Lavandier*, Thibault Vicente, and Luna Prud'homme

Univ. Lyon, ENTPE, Laboratoire de Tribologie et Dynamique des Systèmes UMR 5513, Rue Maurice Audin, 69518 Vaulx-en-Velin Cedex, France

Received 4 August 2021, Accepted 21 April 2022

Abstract – This technical paper presents a series of speech intelligibility models that have been developed since the original version proposed by Lavandier and Culling [(2010). Journal of the Acoustical Society of America 127, 387–399]. This binaural model accounts for better-ear listening and binaural unmasking to predict the intelligibility of a near-field target speech among multiple stationary noise sources in rooms for normal-hearing listeners. Subsequent model versions allowed to consider a reverberated speech target in the far-field, envelope-modulated noise sources, and hearing-impaired listeners. As an intermediate step before considering speech maskers, a monaural version incorporating a harmonic-cancellation mechanism was recently developed to account for the effect of a stationary harmonic masker. This technical review is oriented towards model users and explains when and how each model should be used, points at its advantages and limitations, and provides an example of predictions using a data set from the literature. All these models along with the data, signals and code used to prepare the presented figures are made available within the Auditory Modeling Toolbox (AMT 1.1).

Keywords: Speech intelligibility, Binaural hearing, Auditory models, Hearing impairment, Harmonic cancellation

1 Introduction

Binaural hearing can improve speech intelligibility in noise: an interfering sound source causes less masking when it is spatially separated from the target speech [1, 2]. This spatial release from (energetic[1]) masking (SRM) is thought to be based on two mechanisms [3]: better-ear listening and binaural unmasking. Better-ear listening is associated with the difference in sound level produced by the competing sources at the two ears (interaural level differences, ILDs). For spatially separated sources, one ear usually offers a better signal-to-noise power ratio (SNR) than the other, and listeners can use the information coming from whichever ear offers the better SNR. Binaural unmasking corresponds to the release from masking associated with the difference in the timing of the sound at the two ears (interaural time differences, ITDs) between the target and masker signals, which as per the equalization-cancellation (E-C) theory [4] allows the auditory system to cancel part of the masker, thus improving the internal SNR. These two mechanisms are described in detail in a recent review chapter [5]. In rooms, SRM is reduced by reverberation [1, 6, 7]. Sound reflections traveling around the listener reduce ILDs, thus critically impairing better-ear listening [1]. Because these reflections are generally not identical at the two ears, they also impair binaural unmasking by decorrelating the interfering sound at the ears [8].

Several binaural models have been proposed to predict the effect of SRM on speech intelligibility. They are also presented in a recent review [9]. The aim of the present paper is to compare the model versions within a series of SNR-based intelligibility models that have been developed since the original version proposed by Lavandier and Culling (2010) [10]. All these models are made available within the Auditory Modeling Toolbox (AMT 1.1 [11]). The first versions are binaural models predicting the intelligibility of a near-field (close to the listener) speech target among multiple stationary[2] noise sources in rooms for normal-hearing

*Corresponding author: mathieu.lavandier@entpe.fr

[1] The factors limiting intelligibility are often separated into energetic masking [77] and informational masking [67]. Energetic masking occurs when the competing sounds overlap acoustically with the target and render it less audible. Informational masking can occur when the interfering sources are competing talkers similar to the target. It is associated with difficulties in determining which parts of the speech mixture belong to the target (achieving segregation) and difficulties in attending to the right source in the mixture (overcoming distraction).

[2] Stationary is used here to characterize the absence of envelope modulations in the masker, not to be confused with *static*, which indicates that the position of the source is fixed. Only static sources are considered here. None of the presented models have been tested for moving sources.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

(NH) listeners, using as inputs either the signals reaching the listener's ears (`lavandier2022`) or the binaural room impulse responses (BRIRs) measured in the considered configurations (`jelfs2011`). Subsequent model versions were designed to consider a reverberated speech target further away from the listener in the far-field (`leclere2015`), or envelope-modulated noise sources (`vicente2020nh`), and hearing-impaired (HI) listeners (`vicente2020`). A SNR-based monaural model incorporating a harmonic-cancellation mechanism that accounts for the effect of a stationary harmonic masker is also presented (`prudhomme2020`), and constitutes an intermediate step before considering speech maskers. All these models have been described in detail and validated using various data sets in their corresponding original publication. Our aim here is to explain when and how each model should be used, focusing on the differences across models and the practical details particularly relevant to model users rather than on the underlying theories (fully described in the original publications). For each model, predictions using a data set from the literature is provided. These predictions are not here to test or validate the model, but to give an example on how to use it. The data and signals are available within the AMT 1.1 [11] along with the code used to run the models for all the examples provided, so that the user can check how to handle a particular model before using it to predict intelligibility in other conditions. The code used to compute the predictions displayed in each of the six figures of this paper is integrated as `exp_lavandier2022` in the AMT 1.1 (see section Experiments in the Documentation).

2 Common structure

The original model proposed by Lavandier and Culling (2010) [10] accounts for better-ear listening and binaural unmasking. Even if the subsequent models differ from the original in their exact computations, their underlying concept and structure are identical. Like other binaural models [7, 12], this model is based on the E-C theory [4]; but the direct implementation of equalization and cancellation is replaced by a predictive equation and the resulting prediction of binaural unmasking advantage is added to a better-ear SNR, making it more conceptually similar to earlier anechoic models [13, 14].

The better-ear listening and binaural unmasking components of the model are predicted independently, from the signals produced by the sources at the ears of the listeners. The target and interferer signals must be available separately. When multiple masking sources are present, the overall interfering signal resulting from these sources is used as input. The better-ear SNR is estimated from the SNR computed as a function of frequency at each ear, selecting band-by-band the ear for which the SNR is higher. SNRs are weighted according to their relevance for speech using the speech intelligibility index (SII) weightings [15], and integrated across frequency to provide a broadband better-ear SNR in dB. Binaural unmasking is modeled by increasing the SNR by the size of the binaural masking level difference (BMLD) for pure tone detection in noise in each frequency band. BMLDs are estimated from the interaural phase differences of target and interferer (Φ_T and Φ_I) and the interaural coherence of the interferer (ρ_I). The BMLD is obtained in each frequency band using equation (1) following a development of the E-C theory [16, 17]. Where equation (1) returns a negative value, the BMLD is set to zero, following the assumption that binaural thresholds are never above either of the corresponding monaural thresholds [18]. The BMLD values are then SII-weighted and integrated across frequency to provide a broadband binaural unmasking advantage. The target and interferer signals are both cross-correlated to derive the interaural parameters used in equation (1). The coherence is taken as the maximum of the cross-correlation function, and the interaural phase difference is obtained by multiplying the corresponding delay by the center frequency of the band. The search of a maximum in the cross-correlation function is limited to delays within the range plus/minus half the period of the band center frequency. As a result, the model does not predict any BMLD at high frequencies.

$$\mathrm{BMLD} = 10\log_{10}([k - \cos(\Phi_T - \Phi_I)]/[k - \rho_I]) \quad (1)$$

with

$$k = (1 + \sigma_\varepsilon^2)\exp(\omega^2\sigma_\delta^2) \quad (2)$$

and ω = center frequency of the band in rad/s, σ_δ = 105 µs and σ_ε = 0.25 the standard deviations of the E-C time and amplitude jitters, respectively. This equation was adapted from those of Durlach (1972) for tone detection in noise and uses the same jitter values [4], characterizing an internal noise in the EC model.

To predict the overall effect of binaural hearing, the *effective* SNR is obtained by adding the binaural unmasking advantage to the better-ear SNR, assuming additive contributions[3] of the two mechanisms. This assumption, previously discussed [5, 19, 20] and not necessarily used in all binaural intelligibility models (see [9] for a review), allowed for accurate predictions of several data sets where the two mechanisms were involved both in isolation and combination [19, 20]. The model does not provide an absolute evaluation of intelligibility, the prediction method is relative. Effective SNRs can be used to predict measured differences in speech reception threshold (SRT, the SNR for 50% intelligibility). Effective SNRs are simply inverted, so that high SNRs correspond to low predicted SRTs. Predicted differences in inverted effective SNR can be directly compared to SRT differences across experimental conditions. To predict absolute SRTs rather than relative differences, a reference needs to be chosen. When modeling an experiment, we generally choose the average SRT across listeners and conditions as a reference. Inverted ratios are

[3] In the AMT code provided with this paper, the better-ear SNR and binaural unmasking advantage can be obtained as additional outputs to the binaural models, so that the relative contributions of the corresponding mechanisms can be evaluated.

centered to this average SRT (by subtracting their mean and adding the average SRT) to obtain the predicted SRTs, or in other words, the average predicted SRT is aligned to the average measured SRT. The predicted differences in SRT across conditions remain unaffected by this alignment.

The target and overall masker signals used as model inputs need to be equalized in level. During SRT measurements, the level of the target or maskers[4] is varied to control the SNR and reach 50% intelligibility. In the same way, for inputs equalized in level (input SNR = 0 dB), the models will predict differences in (effective/output) SNR across compared conditions, which can be compared to the differences in SNR at threshold during the measurements, the SRTs. The absolute level at which target and masker signals are calibrated is important only when audibility is an issue, i.e. for the model `vicente2020` that takes into account hearing impairment, in which case the signals are calibrated to the (maskers or target) level that was fixed during the measurements. For the five other models presented here, which do not account for audibility, changing the absolute level of the calibration does not change the effective SNR at the model output.

Except for `jelfs2011` and `leclere2015`, the models presented below take as inputs the signals at the ears. To produce reliable predictions, they require averaging across signals (i.e. across time). Thus, predictions are generally averaged across several realizations of the stimuli. With the stationary model `lavandier2022`, predictions can be computed using single realizations of the stationary noise maskers (if they are at least 4- to 5-s long). The predictions change very little when concatenating more stationary noises (or averaging prediction across more noise realizations). This is not the case for the non-stationary models (`vicente2020nh` and `vicente2020`) and modulated noise maskers for which more realizations are required (except in the particular case of maskers with identical modulations across realizations; e.g. for sinusoidally-modulated noises with the same modulation frequency). Even if the model `prudhomme2020` is stationary, it also requires averaging across several realizations of the harmonic masker because of the highly stochastic nature of its predictions (see below). Moreover, for all these models, the target signal used as model input is the average of several realizations of the target sentences. While for the stationary models (`lavandier2022` and `prudhomme2020`) this only reduces the prediction variability across realizations, applying the non-stationary models (`vicente2020nh` and `vicente2020`) directly on the original speech waveforms could mistakenly lead to a reduced effective SNR in the target pauses, resulting in reduced predicted intelligibility, implicitly considering these pauses as an absence of information instead of relevant information [21]. To avoid this, the non-stationary models consider target energy averaged across time.

[4] In this paper, only experiments in which SRTs were measured varying the target level are presented.

Model performances are evaluated here using three statistics: the Pearson's correlation coefficient r between measured and predicted SRTs, the mean absolute error (MeanErr) computed as the average across conditions of the absolute difference between data and predictions, and the maximum absolute error (MaxErr). Note that, when evaluating the prediction performance of a model, it is important to consider both prediction errors and correlation between data and predictions: predictions can be well correlated with the data but with a general offset so that prediction errors are large; if the effects in the data are small and the model does not predict any difference across compared conditions (not capturing any of these small effects), then predictions errors can be small but predictions poorly correlated with the data.

3 Binaural models for stationary noise maskers

3.1 Models

The first two models, `lavandier2022` and `jelfs2011`, are relevant when considering a near-field (or anechoic) target speech in the presence of single or multiple stationary noise sources in rooms (that can be anechoic) for NH listeners. The model `jelfs2011` [19, 20] is a revision of the original model [10] presented in the previous section. In addition to some computational changes, the revised model is applied on the BRIRs measured between the listener's ears and the different source positions rather than on the full stimuli produced at the ears. Applying the model directly on BRIRs allows to produce fast and accurate non-stochastic predictions (no averaging across realizations). The model `lavandier2022` has the same computational steps as `jelfs2011`, except that it is applied on the ear signals (like the original model [10]) rather than on the BRIRs. In both models, inputs are decomposed into simulated peripheral frequency channels using a gammatone filterbank. The switch from signals (`lavandier2022`) to BRIRs (`jelfs2011`) requires signal power to be replaced by BRIR energy when calculating SNRs, so that this calculation is independent of the presence of a silence at the end of the BRIR, in the same way root-mean-square (RMS) power calculation is independent of (stationary) signal duration.

When using `lavandier2022`, the signals produced by multiple interferers are simply summed to obtain the overall interferer at the ears. When using `jelfs2011`, the interferer BRIRs are concatenated rather than added. Concatenation has the effect of summing the frequency-dependent energy of each contributing impulse response, and generating an averaged cross-correlation function. Summing directly the BRIRs would result in spectral distortion due to interference, which does not occur when summing statistically independent interfering signals convolved with those BRIRs. Concatenation is the appropriate approach when the interfering sources are independent. Only in the particular case of different interfering sources driven by the same signal (e.g., different loudspeakers driven by the same input [22]) should the BRIRs be summed.

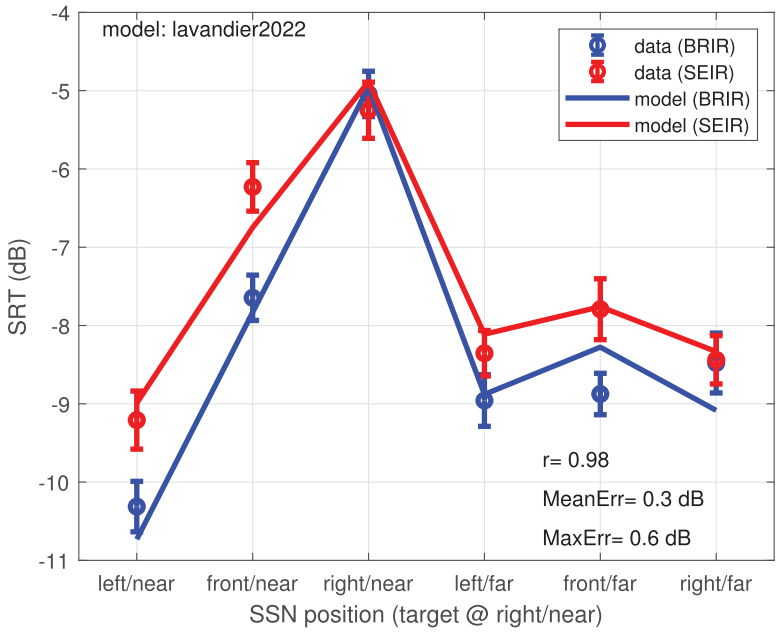

Figure 1. Mean SRTs with standard errors across listeners measured by Lavandier et al. (2012) [20] with a stationary speech-spectrum noise (SSN) simulated at two distances (near, far) in three directions (left, front, right) in a meeting room, using BRIRs (blue; binaural unmasking and better-ear listening involved) and SEIRs (red; no binaural unmasking). The target was always at near-right. Predicted SRTs and performance statistics are displayed for the model `lavandier2022`.

The model `jelfs2011` applied on BRIRs is more computationally efficient and also particularly suited for extension towards reverberated targets affected by temporal smearing (see `leclere2015` below). The version `lavandier2022` using ear signals as inputs is well adapted for extension towards non-stationary maskers (see `vicente2020nh` below). That being written, the two models are conceptually identical and very close in terms of computation, thus producing very similar predictions, validated across a wide range of anechoic [19, 23] and reverberated [20] conditions.

3.2 Data

Lavandier *et al.* (2012) [20] measured SRTs in the presence of a single stationary noise interferer (Fig. 1). In order to evaluate the relative contributions of better-ear listening and binaural unmasking, the stimuli were manipulated so that they contained normal ITDs (binaural unmasking and better-ear listening involved) or no ITD (no binaural unmasking). Real-room listening over headphones was simulated by convolving anechoic stimuli with BRIRs measured in a meeting room. Spectral-envelope impulse responses (SEIRs) were also used. They were obtained by removing the temporal characteristics of the BRIRs whilst preserving their spectral envelopes, thus removing the ITDs necessary for binaural unmasking while preserving the frequency-dependent ILDs necessary for better-ear listening. The difference between BRIRs (blue) and SEIRs (red) indicates the contribution of binaural unmasking. The interferer was tested at two distances (0.65 m "near", 5 m "far") and three directions (−25° "left", 0° "front", 25° "right") relative to the listener; whereas the target speech was always at near-right. Moving the noise away from the target, from right to left, improved intelligibility (reduced SRTs) due to SRM. Increasing the noise distance in the room increased the influence of reverberation on this source, leading to a strong reduction of SRM. For nearby interferers, the contribution to SRM of better-ear listening (SEIR data, red) was larger than that of binaural unmasking. Increasing reverberation reduced the influence of azimuth separation between sources, indicating that head shadow was very limited in the far conditions; but intelligibility then benefited from room coloration. Binaural unmasking was still apparent in the far conditions.

3.3 Implementation of the predictions

The `lavandier2022` predictions presented in Figure 1 were computed from the ear signals, using a single realization of the masker signal in each of the twelve tested conditions. The target was represented by averaging 120 target sentences in each condition (BRIR or SEIR), whereby all sentences were truncated to the duration of the shortest sentence. Because the BRIR-processed maskers end with a non-stationary portion corresponding to the reverberation decay (absent in the SEIRs [20]), only the first 4.2 s of these signals were used for the predictions (there was no effect of removing this non-stationary part on the predictions presented here, but this could become important when considering long BRIRs). To mimic the level equalization of the stimuli during the experiment, the mean of the left and right RMS powers of the averaged target signals was equalized to that of the maskers, which were equalized across

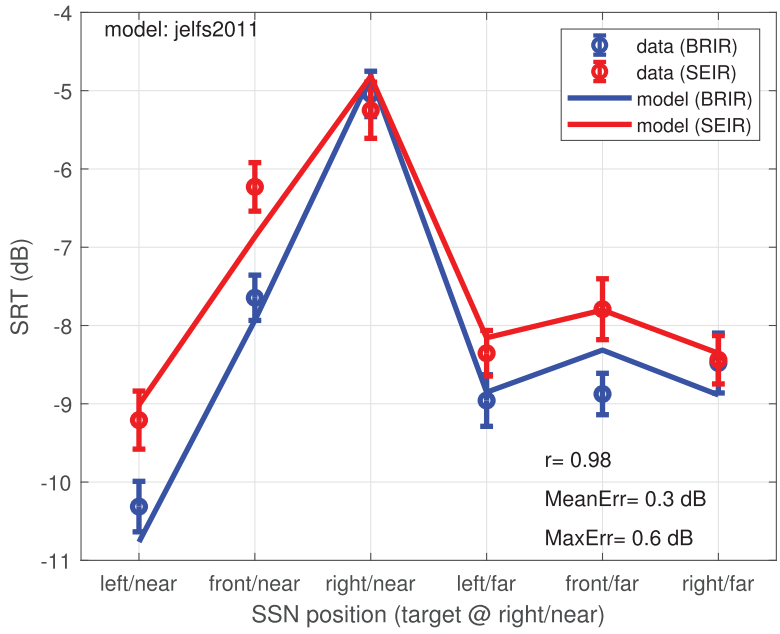

Figure 2. Mean SRTs with standard errors across listeners measured by Lavandier et al. (2012) [20] with a stationary speech-spectrum noise (SSN) simulated at two distances (near, far) in three directions (left, front, right) in a meeting room, using BRIRs (blue; binaural unmasking and better-ear listening involved) and SEIRs (red; no binaural unmasking). The target was always at near-right. Predicted SRTs and performance statistics are displayed for the model `jelfs2011`.

tested conditions. The model was then applied on the equalized signals. Because a single realization of the maskers was considered here, no averaging of predictions across realizations was required. The reference SRT used to transform model outputs into predicted SRTs was the average across the 12 conditions of the mean SRT across listeners. The average of the inverted effective SNRs was aligned to this average measured SRT to obtain the predicted SRTs.

Averaging across realizations is never required when making predictions with `jelfs2011`. The impulse responses used as inputs still need to be equalized in level according to the stimuli used during the experiment.[5] To do so, the impulse responses were first filtered by a speech-spectrum filter (used to create the stimuli in the experiment [20]) and the filtered target and masker impulse responses were equalized in mean energy (not RMS power) across left and right channels. The model was then applied on these equalized responses and produced non-stochastic predictions (Fig. 2). As for the `lavandier2022` predictions, the `jelfs2011` outputs were transformed into predicted SRTs using the average SRT across the 12 conditions as a reference. Note that for anechoic predictions involving much shorter head-related impulse responses, zero padding of the impulses responses is often required. Having impulses responses of at least 1024 samples allows for the gammatone filter to complete its response and produces accurate predictions [19].

[5] When multiple interferers are involved and the level of the overall interferer is the reference, the target BRIR is equalized (in energy) to the level of the overall interferer BRIR (obtained by concatenating the individual interferer BRIRs).

The predictions in Figures 1 and 2 confirm that `lavandier2022` and `jelfs2011` produce very similar predictions (the effective SNRs were on averaged 0.25 dB higher for `jelfs2011`, the difference between the two model outputs being always between 0.05 and 0.37 dB across the 12 tested conditions). The two models can predict SRM, binaural unmasking and better-ear listening, as well as the deleterious effect of reverberation on SRM (for both models $r = 0.98$, MeanErr $= 0.3$ dB, and MaxErr $= 0.6$ dB).

3.4 Limitations

The models `lavandier2022` and `jelfs2011` should not be used when considering envelope-modulated interferers, nor HI listeners, nor (highly) reverberated target speech having its intrinsic intelligibility impaired by reverberation (which is not the case for near-field or anechoic targets). Even if the models can predict the influence of having target and maskers with different spectra, providing that the impulse responses are filtered accordingly for `jelfs2011`, only broadband stimuli should be considered. In our experience and as in other modelling approaches, predictions fail for sharply filtered stimuli [15, 24].

4 Binaural model for non-stationary noise maskers

Modulations in the temporal envelope of the interferer can reduce its masking. In the temporal dips of the masker, the SNR is momentarily increased, allowing one to hear the target better [25, 26]. This ability is often called (temporal)

dip listening, listening in the gaps, or glimpsing [27]. Sound reflections in rooms reduce the possibility for dip listening by reducing the envelope modulations of the interferer, making it more masking by filling in the gaps through which the target could be heard [28–31].

4.1 Model

To account for the effect of envelope modulations in noise maskers, Collin and Lavandier (2013) [31] revised the stationary model `lavandier2022`, applying it on the ear signals within short-time frames and averaging the resulting predictions across frames/time, as also done in other models [21, 30]. The latest version of this non-stationary model is `vicente2020nh` [32]. The input masker signal is segmented using half-overlapping Hann windows before the gammatone filtering. The better-ear SNR calculation uses 24-ms windows to predict *monaural* dip listening [21, 30, 33], while the binaural unmasking advantage calculation uses 300-ms windows to take into account binaural sluggishness [33–36]. The long-term[6] characteristics (spectrum, interaural phase) of the target are computed once and combined with the short-term[7] characteristics (spectrum, interaural phase and coherence) of the masker to compute the effective SNR within each time frame before averaging. To prevent the better-ear SNR tending to infinity in the interferer pauses, a 20-dB ceiling is applied when computing this SNR in each time frame and frequency band. Since a binaural unmasking advantage should not be computed in the absence of noise, this advantage is set to zero if there is no interferer energy at one of the ears in the considered band and frame. As for `lavandier2022`, the signals produced by multiple interferers are simply summed to obtain the overall interferer signal used as model input.

This third model `vicente2020nh` is relevant when considering a near-field/anechoic target speech in the presence of multiple non-stationary noise sources in rooms for NH listeners, that is to say in the same conditions than `lavandier2022` and `jelfs2011` while additionally accounting for the effect of envelope modulations in the noise maskers. This model has been validated considering about one hundred conditions [32] and further used to evaluate (energetic) masking in similar conditions [37].

4.2 Data

Collin and Lavandier (2013) [31] measured SRTs for stationary noises and noises modulated by the envelope of one or two voices (Fig. 3). All sources were simulated at the same short distance (65 cm) from the listener in a meeting room. The speech target was always presented at 0° (front). The three types of noise were tested at 0° (co-located condition) and +25°. The stationary and 2-voice modulated noises were also tested in a symmetrical configuration involving two independent interferers placed on each side of the target, at −25° and +25°. The 2-voice modulated noises were obtained by adding two independent 1-voice modulated noises simulated at the same (0° or +25°) or different positions (−25° and +25°). Intelligibility improved when increasing masker modulations from stationary to 1-voice modulated noise, and when spatially separating target and interferer(s).

4.3 Implementation of the predictions

The `vicente2020nh` predictions presented in Figure 3 were computed using 36 realizations of the masker signals in each of the eight tested conditions. The target was identical in all conditions and represented by averaging 60 target sentences, whereby all sentences were truncated to the duration of the shortest sentence and trimmed of the 150-ms silence at their beginning. Only on-going portions of the masker signals, between 150 ms and 3.5 s, were used for the predictions. The mean of the left and right RMS powers of the averaged target signal was then equalized to that of the maskers (equalized across conditions as in the experiment). The model was applied on the equalized signals and predictions were averaged across the 36 masker realizations. To transform the model outputs (averaged across realizations) to predicted SRTs, the average of the inverted effective SNRs across the 8 conditions was aligned to the average measured SRT (across listeners and 8 conditions).

The predictions in Figure 3 indicate that `vicente2020nh` captures the effects of masker modulations and SRM in the data ($r = 0.93$, MeanErr = 0.6 dB, and MaxErr = 1.4 dB). Note that the model overestimates SRM by about 1.5 dB in the ±25° condition of this particular data set, but such overestimation was not observed for other data sets [32]. Again, these predictions are presented here to illustrate how to use the model, not to test or validate it, for which one should refer to the original publication [32].

4.4 Limitations

The model `vicente2020nh` should not be used when considering HI listeners, nor reverberated target speech. Vicente and Lavandier (2020) [32] also indicated that the model tends to slightly underestimate the effect of reverberation filling in the interferer gaps, as also observed with another binaural model [30].

5 Binaural model for reverberated speech target and stationary noise maskers

In addition to its effects of SRM (Figs. 1 or 2), reverberation exerts a well-known temporal smearing on the target speech, which occurs even in quiet. When the speech signal at the ears is mixed with the multiple delayed versions of itself reflected off room boundaries, it is temporally smeared

[6] As mentioned previously, target level is always averaged across time to avoid mistakenly predicting reduced intelligibility in the target pauses.

[7] The model computes the interferer level as a function of time because peaks in the interferer signal induce an increase of masking whereas pauses induce a decrease of masking.

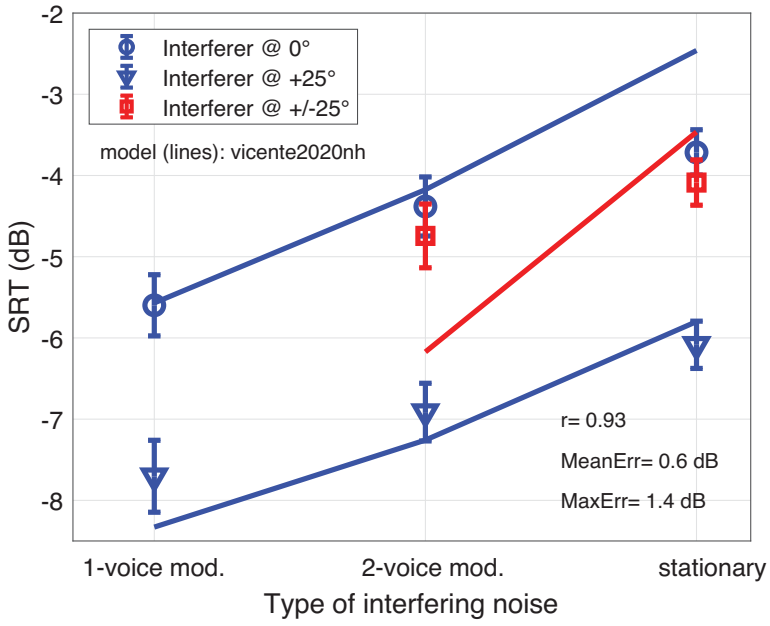

Figure 3. Mean SRTs with standard errors across listeners measured by Collin and Lavandier (2013) [31] with stationary, 1-voice modulated or 2-voice modulated noises tested at 0° (co-located condition), +25°, or ±25°. The target was always at 0°. Predicted SRTs and performance statistics are displayed for the model `vicente2020nh`.

and can be self-masked. This smearing reduces the amplitude modulations in the target speech and its intelligibility can be impaired [38]. The level of reverberation needs to be sufficiently high for temporal smearing of the target to be the overriding factor for intelligibility [8, 39]. Having two ears may ameliorate this smearing effect on intelligibility slightly, thanks to binaural de-reverberation [8, 40, 41].

5.1 Model

Leclère *et al.* (2015) [42] revised the stationary model `jelfs2011` taking BRIRs as inputs, to compute binaural useful-to-detrimental (U/D) ratios and simultaneously account for temporal smearing, SRM, and binaural de-reverberation. It combines the binaural model `jelfs2011` predicting SRM of a near-field target from multiple stationary noise interferers and a U/D decomposition taking into account the temporal smearing effect of reverberation on speech transmission. The U/D decomposition regards the early reflections of the target as useful and part of the *signal* because they reinforce the direct sound [43], whereas the late reflections are regarded as detrimental and effectively a part of the noise [44–46]. The revised model `leclere2015` is identical to `jelfs2011`, except that it incorporates a front end realizing the U/D decomposition. The target BRIR is first separated into an early and a late part. The early part constitutes the useful component. The late part is combined with the BRIRs of the interferers to form the detrimental component.[8] These BRIRs are concatenated (not added) and 1024-sample zero padding is added to the early and late parts of the target BRIR during the U/D decomposition (following the recommendations on the use of `jelfs2011` described above). The binaural model `jelfs2011` is then applied on the useful and detrimental components in the same way as it was previously applied on the target and interferer BRIRs.

The separation of the target BRIR into early and late parts uses two complementary temporal weighting windows: the early and the late windows that isolate each part by multiplying the original BRIR with the corresponding time-domain window. A rectangular window is a common way to split an impulse response: the early part is defined as the original impulse response until the early/late limit [43, 45, 47–49]. Despite its simplicity, the frontier between useful/early and detrimental/late is then infinitely sharp, so that two reflections can be considered very differently even if they are separated by only few samples. Warzybok *et al.* (2013) [50] highlighted this limitation in the presence of a single reflection; while Lochner and Burger (1964) [44] showed that only a part of the early reflections energy can be considered useful for intelligibility. The linear window used here progressively weights the reflections across time. The early window samples are equal to one from the beginning of the window to the early-late limit of 30 ms, then their amplitude decreases linearly from one to zero during 25 ms, the later samples of the early window are equal to zero. The late window is the complement of the early window, such that their sum is always equal to one. These parameters correspond to the room-independent model proposed by Leclère *et al.* (2015) [42].

This fourth model `leclere2015` is relevant when considering a target speech in the presence of multiple

[8] Note that the energy equalization of the target and interferer BRIRs sets the relative level of the late target compare to that of the interferers at the model input, which might differ from its relative level at the SRT during the experiment.

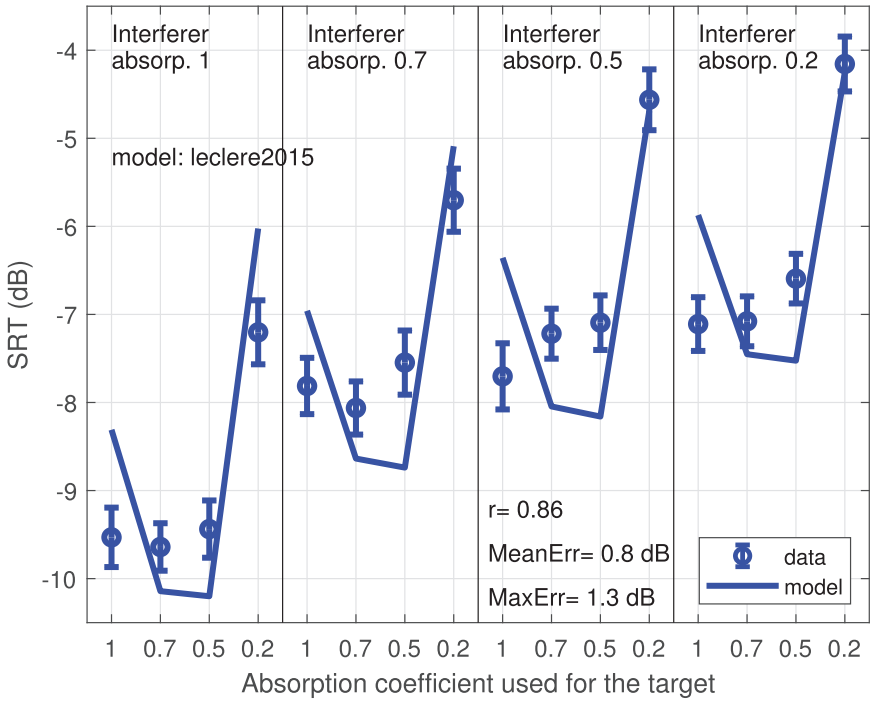

Figure 4. Mean SRTs with standard errors across listeners measured by Lavandier and Culling (2008) [8] with a stationary noise spatially separated from the speech target in a virtual room, plotted as a function of the room absorption coefficients used for the noise (four panels) and for the speech (x-axis). Predicted SRTs and performance statistics are displayed for the model `leclere2015`.

stationary noise sources in rooms for NH listeners [42], extending the conditions of use of `jelfs2011` to far/reverberated targets. However, it is not appropriate when considering non-stationary maskers (the effect of masker envelope modulations is not accounted for as it is in the model `vicente2020nh`).

5.2 Data

Lavandier and Culling (2008) [8] measured SRTs for a stationary noise simulated spatially separated from the speech target (65° to the left and right of the listener, respectively, 2 m away) in a virtual room. Different room absorption coefficients were used for the target and interferer, so that their reverberation level could be controlled independently. This experimental design is not realistic as it implies listening simultaneously to two sources in a room having different walls depending on the source, but it allowed to decompose the effect of reverberation into its temporal smearing of the target and its decorrelation of the interferer at the ears (reducing SRM). Four absorption coefficient were tested for each source (0.2, 0.5, 0.7, and 1 corresponding to an anechoic room). Intelligibility suffered from the effect of reverberation on the noise at lower levels of reverberation than those affecting intelligibility when applied to the target (Fig. 4).

5.3 Implementation of the predictions

The BRIRs used as model inputs were equalized in level according to the stimuli used in the experiment [8]. First, the target and interferer BRIRs were convolved with a filter mimicking the long-term spectrum of male and female speech, respectively; then the filtered BRIRs were equalized in energy independently for the left and right channels (as the stimuli were in the experiment). The model was applied on the equalized BRIRs. Note that the BRIRs of the virtual room contained a DC component, as it is often the case in room acoustics simulations that use an impulse with a non-zero mean to create the direct sound (and resulting reflections). Thus, the energy calculation used in the equalization disregarded any energy below 20 Hz (this precaution is not necessary when manipulating real BRIRs that do not have a DC component).

For the U/D decomposition, the direct sound was defined using a recursive algorithm applied to each channel of the BRIR to locate the direct left and right sounds, and then the earlier of the two was selected as the arrival time of the direct sound of the BRIR. The algorithm finds the first sample that is at least 20% greater than all previous samples in the impulse response. It was used to avoid defining the direct sound as the maximum value or the first non-zero sample, which could induce errors if a combination of reflections is more energetic than the direct sound or if some ambient noise is recorded before the direct sound in the impulse response [42].

The reference SRT used to transform the model outputs into predicted SRTs was the average across the 16 conditions of the mean SRT across listeners. The predictions in Figure 4 indicates that the model can predict the effects of reverberation (r = 0.86, MeanErr = 0.8 dB, and Max Err = 1.3 dB) temporally smearing the target (effect of target absorption coefficient) and reducing SRM via

the reduction of interferer coherence (effect of interferer absorption coefficient). The SRT is overestimated by about 1 dB in the anechoic target condition, but such systematic overestimation was not observed for other data sets [42].

5.4 Limitations

In the study of Leclère *et al.* (2015) [42], the best model performance was achieved by adjusting the early/late separation for each tested room. The room-independent parameters used here did not lead to similar performances, suggesting that a fixed early/late separation might not be sufficient to predict speech intelligibility in any room. This room dependence might indicate an inherent limitation of the approach and could partially explain the wide range of early/late limits encountered in the literature.[9] To overcome this limitation, one could try to make the early/late separation room-dependent [51, 52], or one might change the definition of useful/early and detrimental/late. Rennies *et al.* (2019) [53] recently showed that a reflection could be characterized as useful or detrimental depending on whether it is binaurally useful/detrimental, i.e. contributing or not to SRM, rather than early or late. The distinction between useful and detrimental might also be influenced by room adaptation [54, 55] and speech rate (delayed reflections of previous words can overlap with the direct sound word depending on how fast the words are spoken and how delayed the reflections are).

As stated previously, *leclere2015* cannot account for effects of envelope modulations in the masker nor hearing impairment, so that it is appropriate only for conditions involving stationary noises and NH listeners.

6 Binaural model for hearing-impaired listeners

Hearing impairment is clinically characterized as hearing threshold elevation (measured by a pure-tone audiogram) that leads to lower audibility of the target and masker information [56, 57]. It can also involve suprathreshold deficits, i.e., the information that is audible for HI listeners is not as well processed as for NH listeners [58]. For these reasons, HI listeners generally show higher SRT and lower SRM (see [59] for a review). The models described above are not able to predict these reduced intelligibility and SRM; in particular, they do not consider any information about the listener's hearing status.

6.1 Model

Vicente *et al.* (2020) [60] revised the model `vicente2020nh`, so that it could be used to predict intelligibility for both NH and HI listeners. The revised model `vicente2020` accounts for hearing impairment by implementing internal noise levels at the ears. It takes as input the individual listener's audiograms, along with the target and masker signals at the ears. These signals need to be equalized in level (like for the other models presented here) but also calibrated to the sound presentation level in dB SPL used in the considered conditions. Here, the sound level that was fixed during the SRT measurements was used for the calibration, i.e. the overall level of the maskers.

The internal noise spectrum is spectrally shaped according to the ear-specific listener's audiogram. In each frequency band, the hearing loss is decomposed into two contributions interpreted as rough estimates of the outer and inner hair cell losses [61, 62]. The outer hair cell loss defines a constant internal noise floor, to which is added an internal noise component (based on the inner hair cell loss) that increases linearly with the level of the external stimuli [63]. This level was approximated here as the external (overall) masker long-term broadband level.[10] The external level dependence of the internal noise proved particularly relevant to avoid overestimating the difference in SRT between NH and HI listeners (i.e. overestimating the HI impairment) at low sensation levels [60].

The time-frequency analysis of the signals is similar to the one in `vicente2020nh`. To compute the SNR at the better ear, the SNR at each ear is determined using the higher between the external and internal noise levels, and limited to 20 dB. The higher SNR across ears is selected as the better-ear SNR. The binaural unmasking advantage is computed as in `vicente2020nh` but only if the masker and target levels are above the internal noise levels at both ears in the frame and band considered (otherwise it is set to zero). Then, the values are SII-weighted, integrated across frequency, averaged across time, and added to obtain a listener-specific effective SNR.

It is worth noting that the internal noise implementation simulates increased audiometric thresholds but does not simulate any reduced spectral or temporal resolution often associated with hearing loss. This fifth model `vicente2020` extends the predictions of `vicente2020nh` to HI listeners, and is otherwise relevant in the same conditions: a near-field/anechoic target speech in the presence of multiple non-stationary noise sources in rooms (the detrimental effect of temporal smearing of the target by reverberation is not accounted for as it is by `leclere2015`). The model has been validated considering such conditions [60] and further used to evaluate (energetic) masking in speech-in-speech conditions [64]. It is worth noting that the backward compatibility of `vicente2020` was verified [60], where `vicente2020` provided similar results as `lavandier2022` and `vicente2020nh` when tested with the same data.

[9] An early/late limit of 50 ms has been used very commonly [43, 47–49], but other studies also used a limit of 35 ms [45], 80 ms [45], and 100 ms [44, 81, 82].

[10] This approximation is valid when the external level is dominated by the masker level, that is to say when the SNR/SRT is below 0 dB, which was the case in the conditions considered here. If the sound level fixed/known during the SRT measurements were the target level rather than the masker level, then the target level might have to be used as a proxy for the level of the external stimuli, this approximation being valid only in the less frequent conditions with a SNR/SRT above 0 dB.

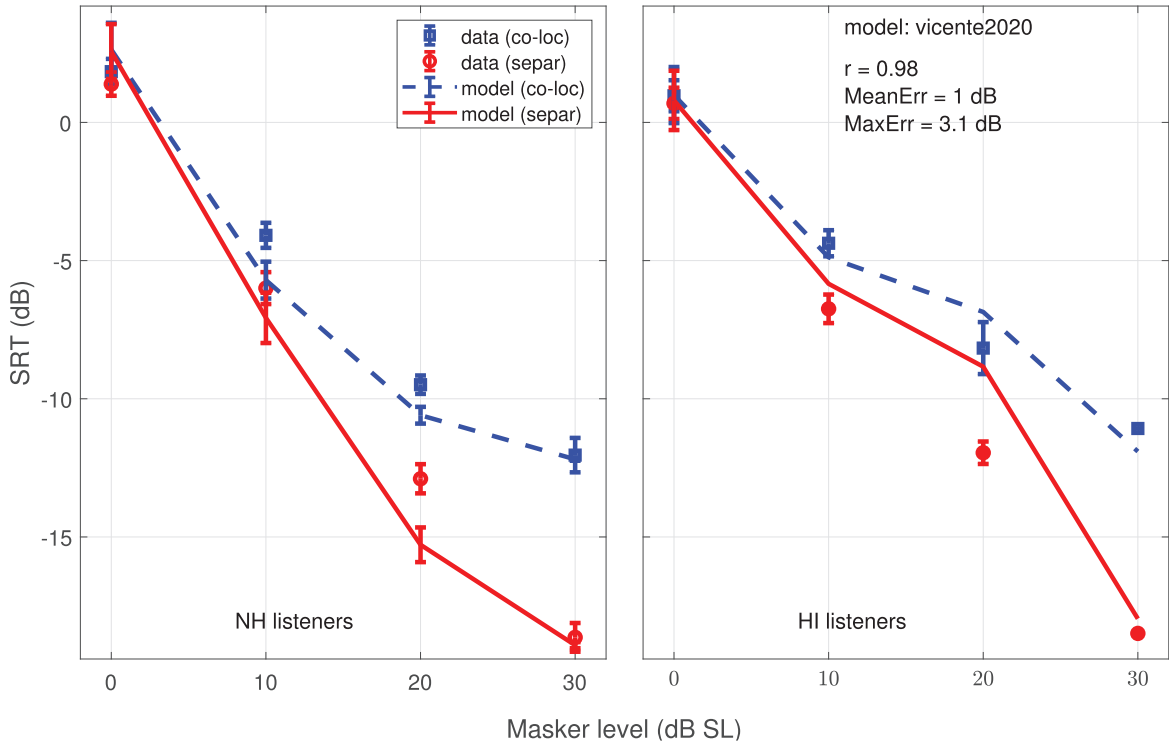

Figure 5. Mean SRTs with standard errors across NH (left panel) and HI (right panel) listeners measured by Rana and Buchholz (2018) [57] as a function of masker sensation level (dB SL, using the individual SRT in quiet as reference). The two vocoded-speech maskers were either co-located with the speech target in front of the listener ("co-loc", blue) or simulated on each side of the listener ("separ", red). Predicted SRTs and performance statistics are displayed for the model `vicente2020`.

6.2 Data

Rana and Buchholz (2018) [57] measured SRTs in the presence of two noise-vocoded speech maskers[11] (Fig. 5), for 10 young NH listeners aged between 20 and 30 years old (mean age: 23.2 years old) along with 10 older HI listeners aged between 57 and 78 years old (mean age: 70.3 years old). The group of HI listeners presented a four-frequency (0.5, 1, 2, 4 kHz) average hearing loss (4FAHL) equal to 29.1 ± 8.0 dB HL. The sound sources were simulated through headphones, with the target always in front of the listener and the maskers either at the target position ("co-loc") or placed on both sides of the listener at ± 90° ("separ"). This was done by playing the left masker only through the left channel of the headphones and the right masker only through the right channel of the headphones, removing crosstalk between ears, thus involving infinite ILD. Target sentences and maskers were filtered to individually equalize audibility across frequency, and then played at four different sensation levels (0, 10, 20 and 30 dB SL) relative to the individual SRT in quiet. This could induce different ILDs and presentation levels across listeners. Amongst the 10 HI listeners, 1, 6 and 9 of them could not be tested at 10, 20 and 30 dB SL, respectively, due to loudness discomfort. The measured SRTs on Figure 5 show that intelligibility and SRM (difference between blue and red) increased with sensation level.

6.3 Implementation of the predictions

To compute the `vicente2020` predictions presented in Figure 5, the input masker and target signals are listener-dependent due to the individual amplification applied to the stimuli during the experiment. For each listener, a single long masker input signal was used per condition. It was obtained by trimming the first second of the masker signal used during the experiment and truncating it at 120 s. The target signal was identical in all conditions and represented by averaging 128 target sentences, whereby all sentences were truncated to the duration of the shortest sentence and trimmed of the 1.2-s silence and alerting beep at their beginning. The signals provided in the AMT already contain the appropriate ILDs associated with the ear-specific audibility equalization. Hence, only the absolute overall level must be calibrated to the overall masker level used during the experiment. The reference chosen for the 0 dB SL condition was the target level (averaged across ears, in dB SPL) at the individual SRT in quiet. Thus, the target and masker signals were calibrated to this level

[11] Only the broadband condition measured by Rana and Buchholz (2018) is considered here, while band-limited conditions were also tested in their original study [57].

plus 0, 10, 20, or 30 dB according to the condition tested. The number of participants per condition varied across the 16 tested conditions. Hence, the effective SNRs at the model output had to be converted into predicted SRTs with caution. The measured SRTs and effective SNRs were averaged across listeners before being averaged across the 16 conditions to give equal weight to the conditions regardless of the number of participants per condition. The average of the inverted effective SNRs was then aligned to the average measured SRT to obtain the predicted SRTs.

The predictions in Figure 5 indicates that the model captures the influence of sensation level on SRM and SRTs for both NH and HI listeners ($r = 0.98$, MeanErr = 1 dB, and MaxErr = 3.1 dB).

6.4 Limitations

The model `vicente2020` has been validated on three data sets involving SRTs measured with stationary noise and non-stationary (noise-vocoded speech) maskers for NH and HI listeners [60]. Even if only one condition involved realistic ITDs, the binaural unmasking component of the model was further tested in conditions involving real ITDs and reverberation [65]. The model produced accurate predictions for both NH and HI listeners at moderate noise levels (50 and 60 dB SPL), while the interaural jitters in equation 1 had to be revised to describe the data measured at a lower level (40 dB SPL).

The hearing loss profiles considered in the first studies testing the model [60, 65] were quite similar across listeners; especially, listeners suffering from severe hearing loss were not considered nor did asymmetric HI listeners. Also, the model was able to predict the influence of hearing impairment on average across listeners, but it had not been thoroughly tested at predicting individual SRTs. A more recent study [66] showed that the model can be used to predict individual differences among (young) HI listeners for SRTs in noise, modulated noise, and competing speech; as long as cues are available to prevent informational masking from the competing speech (spatial separation in this particular case), and as long as HI listeners do not have a severe hearing loss. To handle this latter case, a floor SNR needs to be introduced in the model to limit extreme negative better-ear SNRs, otherwise the model overestimates impairment and underestimates performance for the severely impaired listeners. Because of the introduction of this new parameter, the model requires as input the group average target level at SRT in the condition considered, so that it can be used to account *a posteriori* for differences across listeners but not across conditions [66]. Note that the model was not able to account for the variability among the NH SRTs, suggesting that other factors limit performance when audibility (as measured with the audiogram) is not compromised.

As stated previously, `vicente2020` does not account for the intelligibility loss associated with the temporal smearing of strongly reverberated targets, nor for the effects of informational masking that can be observed with speech maskers that are similar to the target [67].

7 Monaural model for a stationary harmonic masker

A masker impairs less intelligibility when it has a harmonic structure [68–70], at least in part[12] because it produces less energetic masking when its fundamental frequency F0 is different from that of the target ($\Delta F0 \neq 0$). Two mechanisms have been proposed to explain this effect in SNR terms: spectral glimpsing and harmonic cancellation. Spectral glimpsing assumes that listeners can glimpse target information in the spectral dips of the masker, thus improving speech intelligibility [70]. The harmonic cancellation theory proposes that listeners can detect the harmonic structure of a masker and suppress energy at the corresponding (harmonic) frequencies in order to improve intelligibility [68, 71]. These two mechanisms seem to be impaired for an intonated masker with a F0 varying across time, so that masking increases when F0 variations are introduced [72]. These F0 effects might also be explained (instead or in addition) in terms of modulation masking [73, 74] (the fact that modulation in the masker might prevent the listener from detecting and processing the useful temporal fluctuations of the target speech), in which the harmonic maskers exhibit reduced envelope modulations (and associated masking) compared to noise maskers [75], while varying the F0 in intonated harmonic maskers might introduce additional envelope modulations causing more modulation masking.

Prud'homme *et al.* (2020) [76] proposed a monaural model incorporating a harmonic-cancellation mechanism to account for the effect of stationary harmonic complex tones as a masker. This model `prudhomme2020` is presented within this speech intelligibility model series for three reasons. First, harmonic complex tones with a F0 constitute an intermediate step before considering speech maskers, which voiced parts are harmonic with a F0, and which many model users will be interested in. Second, while being monaural, this model has the same underlying philosophy and structure as the other (binaural) models in the series. It is an SNR-based model [9] that uses the same signal as inputs and produces a similar effective SNR as output. The SNR analysis is very similar to the better-ear SNR analysis of the binaural models, while the E-C mechanism [4] underlying binaural unmasking is replaced here by harmonic cancellation (that is then based on the F0 of the masker rather than on its ITDs). Third, because of this common structure, this model is highly compatible with the binaural models. We have developed a binaural non-stationary version of the harmonic-cancellation model (a hybrid between `prudhomme2020` and `vicente2020nh`). Once this new version is validated/published, it will be made available within the AMT. We anticipate that the presentation of the specificity of `prudhomme2020` (and

[12] A difference in F0 is also a strong cue for the segregation of competing sounds [83, 84] or voices [85] into different auditory streams, thus providing for a release from informational masking.

vicente2020nh) here will also help the future users to navigate the subsequent model versions.

7.1 Model

The model prudhomme2020 is based on the monaural version of the stationary model lavandier2022 applied on the ear signals. It can account for spectral glimpsing because the SNR is computed by frequency band. In addition, harmonic cancellation is implemented by creating a comb filter to remove energy at the masker F0 and its harmonics. The comb filter is applied to both the target and masker input signals. In each frequency band, the best SNR between the SNR with or without applying harmonic cancellation is chosen, the idea being that harmonic cancellation is used only when it provides an advantage[13] (improves the SNR in the band). A jitter is introduced in the estimation of the masker F0 to simulate the fact that the auditory system might not be able to perfectly estimate the F0 nor create a perfect comb filter to remove energy at this F0 and its harmonics. The width of the normal distribution used to select the random jitter value is proportional to the masker F0 (0.25F0). The width of the notches of the filter is also proportional to this F0 (0.6F0). Harmonic cancellation is applied only up to 5 kHz. A 40-dB ceiling is applied to the SNR selected in each frequency band. The resulting SNRs are SII-weighted and integrated across frequency to obtain the effective SNR in the corresponding condition. The model parameters have been defined using data from two experiments of Deroche *et al.* (2014) [70], with single maskers presenting different F0s and different degrees of harmonicity.

This sixth model prudhomme2020 is relevant when considering a monotonous, stationary, diotic/monaural harmonic complex as the single masker. The effects of multiple maskers, binaural hearing, masker amplitude modulations, reverberation on the target, and hearing impairment are not accounted for. Note that the jitter in the F0 estimation leads to stochastic predictions that need to be averaged across several realizations of the jitter (typically 800 [76]), in addition to considering several realizations of the input signals, so that prudhomme2020 is less computationally efficient than lavandier2022 on which it is based.

7.2 Data

Deroche *et al.* (2014) [70] measured SRTs for stationary monotonous harmonic and inharmonic complex tones with different F0s (50, 100, 200 and 400 Hz). The inharmonic complexes were created by randomly jittering each partial from their harmonic position (the size of each jitter being drawn from a uniform distribution between −F0/2 and F0/2). SRTs were measured for frozen (the same masker was used throughout one block of target sentences used to measure one SRT) or fresh (the masker was changed for each sentence) conditions. As there was no significant difference between the two conditions, the results presented in Figure 6 were averaged across frozen and fresh conditions. Intelligibility improved when increasing the masker F0 for both masker types. The effect was even more pronounced for the inharmonic maskers. The harmonic maskers led to lower SRTs than the inharmonic maskers, the difference in SRT decreasing with increasing masker F0.

7.3 Implementation of the predictions

The prudhomme2020 predictions presented in Figure 6 were computed using 160 realizations of the masker signal and 800 trials for each realization, using for each trial a different jitter value taken from a 0.25F0-width normal distribution (where F0 is the fundamental frequency of the considered masker). The target was represented by averaging 160 target sentences (identical in all conditions), whereby all sentences were truncated to the duration of the shortest sentence and trimmed of the 150-ms silence at their beginning. In the study of Prud'homme *et al.* (2020) [76], the target was represented by concatenating the target sentences. However, averaging instead of concatenation considerably reduces the computation time without affecting the prediction results. Only on-going portions of the masker signals, between 150 ms and 2.5 s, were used for the predictions. The RMS power of the averaged target signal was equalized to that of the maskers (equalized across conditions as in the experiment). For each trial, the model was applied on the equalized signals, and then predictions were averaged across the 800 trials and 160 masker realizations. The randomness introduced in the model by the jitter is responsible for the fact that the model predictions vary slightly each time they are computed. The reference SRT used to transform model outputs (averaged across trials and masker realizations) into predicted SRTs was the average across the 8 conditions of the mean SRT across listeners.

The predictions in Figure 6 indicate that the model can account for the difference between harmonic and inharmonic maskers as well as for the decrease in SRT due to increasing masker F0 (r = 0.99, MeanErr = 0.5 dB, and MaxErr = 1.5 dB).

7.4 Limitations

To this point, the model prudhomme2020 has only been validated for NH listeners with a monotonized, stationary, diotic, anechoic tone complex masker. The model should not be used outside this rather limited framework. Moreover, it does not predict the small interaction observed in Figure 6, the difference between harmonic and inharmonic maskers being reduced when increasing the masker F0. The effects of spectral glimpsing and harmonic cancellation are greatly reduced when the masker F0 varies across time [72], this is not accounted for in this stationary model.

[13] A parallel can be drawn with the E-C theory and binaural unmasking that assume that binaural thresholds are never above the corresponding monaural thresholds [18], so that equalization-cancellation is used only when it provides an advantage (decreases threshold). For all the binaural models presented here, if equation 1 returns a negative value, the BMLD is set to zero in the corresponding frequency band.

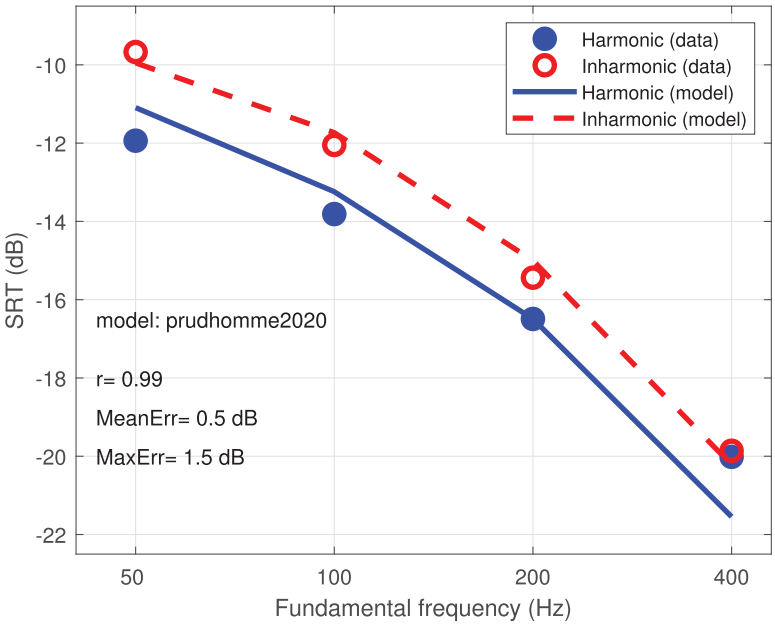

Figure 6. Mean SRTs across listeners measured by Deroche *et al.* (2014) [70] with stationary monotonous harmonic and inharmonic complex tones with a F0 of 50, 100, 200 and 400 Hz. Predicted SRTs and performance statistics are displayed for the model `prudhomme2020`.

Moreover, `prudhomme2020` is currently not able to handle multiple harmonic maskers with different F0s.

It is still unknown how relevant harmonic cancellation would be for more complex stimuli like speech maskers, which are intonated and contain unvoiced segments (with no F0).

8 General discussion

When willing to predict intelligibility, it is important to choose the right model for the considered conditions. Every model has its limitations and underlying assumptions, and will be appropriate or not depending on the type of listeners (NH or HI) and interferers (stationary noise, envelope-modulated noise, harmonic complex). Table 1 summarizes the characteristics of the six models presented here. One might be particularly interested in predicting the intelligibility of speech among competing talkers. No current model is able to do so [9]. Given the models currently available, those validated for non-stationary noise maskers seem to be the most appropriate to evaluate intelligibility differences associated with variations in energetic masking across speech-in-speech conditions. But one should keep in mind that, first, no such model exists that also accounts for the energetic masking effects associated with F0 differences [77]. The importance of such effects highlighted with harmonic complex maskers [70, 72] remains to be investigated for speech maskers that are generally intonated, contain unvoiced segments, and can involve multiple F0s (for multiple speech maskers). Second, these intelligibility models generally do not account for the effects of informational masking [67] that can occur when the speech maskers are similar to the speech target. On the other hand, models can be used to predict the contribution of energetic masking in speech-in-speech conditions, so that the contribution of informational masking can be quantified *a posteriori* as the remaining source of variations in the data once predicted energetic masking effects have been factored out. This has been done previously for NH listeners with `vicente2020nh` [37, 78] and for NH and HI listeners with `vicente2020` [64].

It is important to keep in mind that the absolute predictions of the models presented here do not carry any interesting meaning, it is the relative differences across predictions that are relevant. For example, `lavandier2022` and `jelfs2011` will produce an effective SNR of 0 dB for a stationary noise interferer with the same long-term spectrum as and co-located with the target speech, while `vicente2020nh` will output a 4.3-dB SNR in the same condition, because the Hann window used in the temporal segmentation systematically reduces the noise levels by 4.3 dB. However, the relative differences in prediction across different conditions with stationary noises will be very similar for these three models. Interestingly, to compare differences in predicted SRT across conditions with this model series, there is no requirement to calculate speech indices such as the articulation index (AI) [79] or the SII [15], or to conduct index-to-intelligibility mapping [7, 13]. The predicted differences in model output (effective SNR) can be directly compared to the measured differences in SRT (SNR at threshold). It is important to note that this assumes a linear relationship between SNR and percent correct, which is not verified at low and high SNRs, as

Table 1. Summary of the characteristics of the six models.

		`lavandier2022`	`jelfs2011`	`leclere2015`	`vicente2020nh`	`vicente2020`	`prudhomme2020`
Source type	Stationary noise	X	X	X	X	X	Never tested
	Modulated noise	0	0	0	X	X	0
	Harmonic masker	0	0	0	0	0	X
	Reverberated target	0	0	X	0	0	0
Listener type	NH listeners	X	X	X	X	X	X
	HI listeners	0	0	0	0	X	0
Unmasking/masking type	SRM	X	X	X	X	X	0
	$\Delta F0$	0	0	0	0	0	X
	Intonation	0	0	0	0	0	0
	Audibility	0	0	0	0	X	0
	Informational	0	0	0	0	0	0
Specific input			BRIRs	BRIRs		Audiogram	Masker F0
Specific feature				Early/late limit		Internal noise	Harmonic cancellation

indicated by the sigmoid form of the psychometric functions relating SNR and percent correct (which saturates at high and low SNRs). Thus, the direct comparison seems appropriate when considering the SRT defined at 50% intelligibility, which corresponds to the linear part of the psychometric functions, but it might not be appropriate when other threshold values are chosen (e.g., the SRT at 70% intelligibility). This direct comparison also assumes that the psychometric functions (underlying the SRTs) measured in the different conditions only differ by their SRT and in particular that they have the same slope, which might not always be verified depending on the conditions being compared. For example, the models would not be able to predict any effect on intelligibility associated with a difference in target speech material (which can differ in word frequency or in the presence of syntactic and/or semantic constraints). The nature of this speech material directly affects the slope of the psychometric function relating SNR and percent correct [15]. When one wants to predict absolute SRTs rather than relative differences, then a mapping between predictions and measurements is required: holding the assumptions just mentioned, the predictions are just linearly offset to match a reference SRT (here the average SRT across listeners and conditions in each experiment).

When a particular model has been chosen, it is also important to use the model properly, depending on its design. For example, the level equalization of the input signals is simple (most of the time), but also very important. In the model series presented here, the target and masker input signals need to be equalized in level. Any equalization error affecting differently the compared conditions will be directly transferred to the effective SNRs and thus to the prediction errors. For example, it has been explained above that this equalization needs to consider the energy of the BRIRs or the RMS power of the ear signals. The BRIRs also need to be filtered according to the long-term spectrum of the sources before the equalization (so that it mirrors the equalization of the stimuli). Any (inaudible) DC component in the (virtual) BRIRs needs to be discarded when computing the energy levels. These equalization techniques are illustrated in `exp_lavandier2022` (AMT 1.1 [11]), using the examples presented here.

The performance of a model can be evaluated in different ways. It has already been mentioned that it is important to verify that predictions are both well correlated with the data and lead to low prediction errors. It is of course crucial to use the absolute value of the predictions errors before computing the mean prediction error (to prevent over- and under-estimations to cancel each other). One can also choose to compute the RMS prediction error, even if this can imply taking the square of a dB value. Another way to estimate model performance, compare to equalizing the level of the input signals and then compare the variations in model output to the variations in the data, could have been inversely to calibrate the input signals using the SNR at the SRT (thus using the measured data to set the levels of the input signals) and then verify that the model output is constant (plus/minus the prediction

errors), corresponding to the constant intelligibility level at the SRT. This approach, even if elegant, can be very misleading: evaluating an intelligibility model by checking that it produces a constant output at the SRT is not appropriate because, for example, a "model" that produces a constant value of 0 for any input signal would pass this performance test perfectly even if it cannot predict anything.

The aim of this technical paper was to clarify when and how to use a series of six SNR-based speech intelligibility models. It is intended as a user guide, providing a code example to run each model (`exp_lavandier2022`, AMT 1.1 [11]), so that the user can choose a particular model and verify how to handle its details before predicting intelligibility in other conditions. The questions raised here might also prove helpful when considering other modelling approaches, not limited to intelligibility.

Conflict of interest

The authors declare no conflict of interest.

Data availability statement

Implementations of both the models (`lavandier2022`, `jelfs2011`, `leclere2015`, `vicente2020nh`, `vicente2020`, `prudhomme2020`) and the predictions presented here (`exp_lavandier2022`) are publicly available as part of the Auditory Modeling Toolbox (AMT) [11] in the release of the version 1.1.0 available as a full package for download ([80] https://sourceforge.net/projects/amtoolbox/files/AMT1.x/amtoolbox-full-1.1.0.zip/download).

Acknowledgments

The authors would like to acknowledge the crucial contributions of their colleagues who participated in the development of the models presented here: John Culling, Sam Jelfs, Benjamin Collin, Thibaud Leclère, Jörg Buchholz, and Virginia Best. This work was conducted within the LabEx CeLyA (Lyon Acoustics Center, ANR-10-LABX-0060) and supported by the grants Speech2Ears (Fondation pour l'Audition) and ASH (PHC Danube, Grant Nos. 45268RE, APVV DS-FR-19-0025, WTZ MULT 07/2020).

References

1. R. Plomp: Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of azimuth of a single competing sound source (speech or noise). Acta Acustica united with Acustica 34 (1976) 200–211.
2. M.L. Hawley, R.Y. Litovsky, J.F. Culling: The benefit of binaural hearing in a cocktail party: effect of location and type of interferer. Journal of the Acoustical Society of America 115, 2 (2004) 833–843.
3. A.W. Bronkhorst, R. Plomp: The effect of head-induced interaural time and level differences on speech intelligibility in noise. Journal of the Acoustical Society of America 83, 4 (1988) 1508–1516.
4. N.I. Durlach: Binaural signal detection: Equalization and cancellation theory. In: J. Tobias, Ed. Foundations of Modern Auditory Theory, Vol. II, New York: Academic, 1972: 371–462.
5. J.F. Culling, M. Lavandier: Binaural unmasking and spatial release from masking. In: R.Y. Litovsky, M.J. Goupell, A.N. Popper, R.R. Fay, Eds. Binaural Hearing, Vol. 73 of Springer Handbook of Auditory Research, Switzerland: Springer Nature, 2021: 209–241.
6. J.F. Culling, K.I. Hodder, C.Y. Toh: Effects of reverberation on perceptual segregation of competing voices. Journal of the Acoustical Society of America 114, 5 (2003) 2871–2876.
7. R. Beutelmann, T. Brand: Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. Journal of the Acoustical Society of America 120, 1 (2006) 331–342.
8. M. Lavandier, J.F. Culling: Speech segregation in rooms: Monaural, binaural, and interacting effects of reverberation on target and interferer. Journal of the Acoustical Society of America 123, 4 (2008) 2237–2248.
9. M. Lavandier, V. Best: Modeling binaural speech understanding in complex situations: In: J. Blauert, J. Braasch, Eds. The technology of binaural understanding, Switzerland: Springer Nature, 2020: 547–578.
10. M. Lavandier, J.F. Culling: Prediction of binaural speech intelligibility against noise in rooms. Journal of the Acoustical Society of America 127, 1 (2010) 387–399.
11. P. Majdak, C. Hollomey, R. Baumgartner: AMT 1.x: A toolbox for reproducible research in auditory modeling. Acta Acustica 6 (2022) 19.
12. R. Wan, N.I. Durlach, H.S. Colburn: Application of an extended equalization- cancellation model to speech intelligibility with spatially distributed maskers. Journal of the Acoustical Society of America 128, 6 (2010) 3678–3690.
13. H. Levitt, L.R. Rabiner: Predicting binaural gain in intelligibility and release from masking for speech. Journal of the Acoustical Society of America 424 (1967) 820–829.
14. P.M. Zurek: Binaural advantages and directional effects in speech intelligibility. In: G. Studebaker, I. Hochberg, Eds. Acoustical factors affecting hearing aid performance, Needham Heights, MA: Allyn and Bacon, 1993: 255–276.
15. ANSI S3.5: Methods for calculation of the speech intelligibility index, American National Standards Institute, New York. 1997.
16. J.F. Culling, M.L. Hawley, R.Y. Litovsky: The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources. Journal of the Acoustical Society of America 116, 2 (2004) 1057–1065.
17. J.F. Culling, M.L. Hawley, R.Y. Litovsky: Erratum: The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources. Journal of the Acoustical Society of America 118, 1 (2005) 552.
18. N.I. Durlach: Equalization and cancellation theory of binaural masking-level differences. Journal of the Acoustical Society of America 35, 8 (1963) 1206–1218.
19. S. Jelfs, J.F. Culling, M. Lavandier: Revision and validation of a binaural model for speech intelligibility in noise. Hearing Research 275 (2011) 96–104.
20. M. Lavandier, S. Jelfs, J.F. Culling, A.J. Watkins, A.P. Raimond, S.J. Makin: Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources. Journal of the Acoustical Society of America 131, 1 (2012) 218–231.

21. K.S. Rhebergen, N.J. Versfeld: A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. Journal of the Acoustical Society of America 117, 4 (2005) 2181–2192.
22. J.F. Culling, S. Jelfs, M. Lavandier: An alternative perspective on multi-channel reproduction, in Reproduced Sound 2010, Proceedings of the Institute of Acoustics. 2010.
23. J.F. Culling, M. Lavandier, S. Jelfs: Predicting binaural speech intelligibility in architectural acoustics. In: J. Blauert, Ed. The technology of binaural listening, Berlin-Heidelberg-New York NY: Springer, 2013: 427–447.
24. T. Leclère, D. Thery, M. Lavandier, J.F. Culling: Speech intelligibility for target and masker with different spectra. In: P. van Dijk, D. Başkent, E. Gaudrain, E. de Kleine, A. Wagner, C. Lanting (Eds.), Physiology, psychoacoustics and cognition in normal and impaired hearing, Vol. 894, Springer, Advances in Experimental Medicine and Biology, 2016: 257–266.
25. J.M. Festen, R. Plomp: Effects of fluctuating noise and interfering speech on the speech- reception threshold for impaired and normal hearing. Journal of the Acoustical Society of America 88, 4 (1990) 1725–1736.
26. A.W. Bronkhorst, R. Plomp: Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing. Journal of the Acoustical Society of America 92, 6 (1992) 3132–3139.
27. M. Cooke: A glimpsing model of speech perception in noise. Journal of the Acoustical Society of America 119, 3 (2006) 1562–1573.
28. A.W. Bronkhorst, R. Plomp: A clinical test for the assessment of binaural speech perception in noise. Audiology 29 (1990) 275–285.
29. E.L.J. George, J.M. Festen, T. Houtgast: The combined effects of reverberation and nonstationary noise on sentence intelligibility. Journal of the Acoustical Society of America 124, 2 (2008) 1269–1277.
30. R. Beutelmann, T. Brand, B. Kollmeier: Revision, extension, and evaluation of a binaural speech intelligibility model. Journal of the Acoustical Society of America 127, 4 (2010) 2479–2497.
31. B. Collin, M. Lavandier: Binaural speech intelligibility in rooms with variations in spatial location of sources and modulation depth of noise interferers. Journal of the Acoustical Society of America 134, 2 (2013) 1146–1159.
32. T. Vicente, M. Lavandier: Further validation of a binaural model predicting speech intelligibility against envelope-modulated noises. Hearing Research 390 (2020) 107937.
33. J.F. Culling, Q. Summerfield: Measurements of the binaural temporal window using a detection task. Journal of the Acoustical Society of America 103, 6 (1998) 3540–3553.
34. D.W. Grantham, F.L. Wightman: Detectability of a pulsed tone in the presence of a masker with time-varying interaural correlation. Journal of the Acoustical Society of America 65, 6 (1979) 1509–1517.
35. J.F. Culling, E.R. Mansell: Speech intelligibility among modulated and spatially distributed noise sources. Journal of the Acoustical Society of America 133, 4 (2013) 2254–2261.
36. C.F. Hauth, T. Brand: Modeling sluggishness in binaural unmasking of speech for maskers with time-varying interaural phase differences. Trends in Hearing 22 (2018) 1–10.
37. J. Cubick, J.M. Buchholz, V. Best, M. Lavandier, T. Dau: Listening through hearing aids affects spatial perception and speech intelligibility in normal-hearing listeners. Journal of the Acoustical Society of America 144, 5 (2018) 2896–2905.
38. T. Houtgast, H.J.M. Steeneken: A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. Journal of the Acoustical Society of America 77, 3 (1985) 1069–1077.
39. M. Lavandier, J.F. Culling: Speech segregation in rooms: Effects of reverberation on both target and interferer. Journal of the Acoustical Society of America 122, 3 (2007) 1713–1723.
40. J.P. Moncur, D. Dirks: Binaural and monaural speech intelligibility in reverberation. Journal of Speech and Hearing Research 10 (1967) 186–195.
41. A.K. Nábělek, P.K. Robinson: Monaural and binaural speech perception in reverberation for listeners of various ages. Journal of the Acoustical Society of America 71, 5 (1982) 1242–1248.
42. T. Leclère, M. Lavandier, J.F. Culling: Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking and binaural de-reverberation. Journal of the Acoustical Society of America 137, 6 (2015) 3335–3345.
43. J.S. Bradley, H. Sato, M. Picard: On the importance of early reflections for speech in rooms. Journal of the Acoustical Society of America 113, 6 (2003) 3233–3244.
44. J.P.A. Lochner, J.F. Burger: The influence of reflections on auditorium acoustics. Journal of Sound and Vibration 1, 4 (1964) 426–454.
45. J.S. Bradley: Predictors of speech intelligibility in rooms. Journal of the Acoustical Society of America 80, 3 (1986) 837–845.
46. J.S. Bradley, R.D. Reich, S.G. Norcross: On the combined effects of signal-to-noise ratio and room acoustics on speech intelligibility. Journal of the Acoustical Society of America 106, 4 (1999) 1820–1828.
47. G.A. Soulodre, N. Popplewell, J.S. Bradley: Combined effects of early reflections and background noise on speech intelligibility. Journal of Sound and Vibration 135, 1 (1989) 123–133.
48. I. Arweiler, J.M. Buchholz: The influence of spectral characteristics of early reflections on speech intelligibility. Journal of the Acoustical Society of America 130, 2 (2011) 996–1005.
49. N. Roman, J. Woodruff: Speech intelligibility in reverberation with ideal binary masking: Effects of early reflections and signal-to-noise ratio threshold. Journal of the Acoustical Society of America 133, 3 (2013) 1707–1717.
50. A. Warzybok, J. Rennies, T. Brand, S. Doclo, B. Kollmeier: Effects of spatial and temporal integration of a single early reflection on speech intelligibility. Journal of the Acoustical Society of America 133, 1 (2013) 269–282.
51. A. Lindau, L. Kosanke, S. Weinzierl: Perceptual evaluation of model- and signal-based predictors of the mixing time in binaural room impulse responses. Journal of the Audio Engineering Society 60, 11 (2012) 887–898.
52. O. Kokabi, F. Brinkmann, S. Weinzierl: Segmentation of binaural room impulse responses for speech intelligibility prediction. Journal of the Acoustical Society of America 144, 5 (2018) 2793–2800.
53. J. Rennies, A. Warzybok, T. Brand, B. Kollmeier: Measurement and prediction of binaural-temporal integration of speech reflections. Trends in Hearing 23 (2019) 2331216519854267.
54. A.J. Watkins: Perceptual compensation for effects of reverberation in speech identification. Journal of the Acoustical Society of America 118, 1 (2005) 249–262.
55. E. Brandewie, P. Zahorik: Prior listening in rooms improves speech intelligibility. Journal of the Acoustical Society of America 128, 1 (2010) 291–299.

56. V. Best, E.R. Thompson, C.R. Mason, G. Kidd: An energetic limit on spatial release from masking. Journal of the Association for Research in Otolaryngology 14, 4 (2013) 603–610.
57. B. Rana, J.M. Buchholz: Effect of audibility on better-ear glimpsing as a function of frequency in normal-hearing and hearing-impaired listeners. Journal of the Acoustical Society of America 143 (2018) 2195–2206.
58. S. Santurette, T. Dau: Relating binaural pitch perception to the individual listener's auditory profile. Journal of the Acoustical Society of America 131, 4 (2012) 2968–2986.
59. H. Glyde, L. Hickson, S. Cameron, H. Dillon: Problems hearing in noise in older adults: a review of spatial processing disorder. Trends in Amplification 15, 3 (2011) 116–126.
60. T. Vicente, M. Lavandier, J.M. Buchholz: A binaural model implementing an internal noise to predict the effect of hearing impairment on speech intelligibility in non-stationary noises. Journal of the Acoustical Society of America 148, 5 (2020) 3305–3317.
61. B.C.J. Moore, B.R. Glasberg: A revised model of loudness perception applied to cochlear hearing loss. Hearing Research 188 (2004) 70–88.
62. I. Pieper, M. Mauermann, D. Oetting, B. Kollmeier, S.D. Ewert: Physiologically motivated individual loudness model for normal hearing and hearing impaired listeners. Journal of the Acoustical Society of America 144, 2 (2018) 917–930.
63. L.R. Bernstein, C. Trahiotis: Binaural signal detection, overall masking level, and masker interaural correlation: Revisiting the internal noise hypothesis, Journal of the Acoustical Society of America 124, 6 (2008) 3850–3860.
64. P.A. Wasiuk, M. Lavandier, E. Buss, J. Oleson, L. Calandruccio: The effect of fundamental frequency contour similarity on multi-talker listening in older and younger adults. Journal of the Acoustical Society of America 148, 6 (2020) 3527–3543.
65. T. Vicente, M. Lavandier, J.M. Buchholz: Modelling binaural unmasking and the intelligibility of speech in noise and reverberation for normal-hearing and hearing-impaired listeners. Journal of the Acoustical Society of America 150, 5 (2021) 3275–3287.
66. M. Lavandier, C.R. Mason, L.S. Baltzell, V. Best: Individual differences in speech intelligibility at a cocktail party: a modelling perspective. Journal of the Acoustical Society of America 150, 2 (2021) 1076–1087.
67. G. Kidd, H.S. Colburn: Informational masking in speech recognition. In: J. Middlebrooks, J. Simon, A.N. Popper, R. R. Fay, Eds. The Auditory System at the Cocktail Party, Springer Handbook of Auditory Research, Cham: Springer, 2017: 75–109.
68. A. de Cheveigné, S. McAdams, J. Laroche, M. Rosenberg: Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement. Journal of the Acoustical Society of America 97, 6 (1995) 3736–3748.
69. K. Steinmetzger, S. Rosen: The role of periodicity in perceiving speech in quiet and in background noise. Journal of the Acoustical Society of America 138, 6 (2015) 3586–3599.
70. M.L.D. Deroche, J.F. Culling, M. Chatterjee, C.J. Limb: Speech recognition against harmonic and inharmonic complexes: Spectral dips and periodicity. Journal of the Acoustical Society of America 135, 5 (2014) 2873–2884.
71. A. de Cheveigné, S. McAdams, C.M.H. Marin: Concurrent vowel identification II. Effects of phase, harmonicity, and task. Journal of the Acoustical Society of America 101, 5 (1997) 2848–2856.
72. T. Leclère, M. Lavandier, M.L.D. Deroche: The intelligibility of speech in a harmonic masker varying in fundamental frequency contour, broadband temporal envelope, and spatial location. Hearing Research 350 (2017) 1–10.
73. M.A. Stone, C. Füllgrabe, B.C.J. Moore: Notionally steady background noise acts primarily as a modulation masker of speech. Journal of the Acoustical Society of America 132, 1 (2012) 317–326.
74. M.A. Stone, C. Füllgrabe, R.C. Mackinnon, B.C.J. Moore: The importance for speech intelligibility of random fluctuations in steady background noise. Journal of the Acoustical Society of America 130, 5 (2011) 2874–2881.
75. K. Steinmetzger, J. Zaar, H. Relaño-Iborra, S. Rosen, T. Dau: Predicting the effects of periodicity on the intelligibility of masked speech: An evaluation of different modelling approaches and their limitations. Journal of the Acoustical Society of America 146, 4 (2019) 2562–2576.
76. L. Prud'homme, M. Lavandier, V. Best: A harmonic-cancellation-based model to predict speech intelligibility against a harmonic masker. Journal of the Acoustical Society of America 148, 5 (2020) 3246–3254.
77. J.F. Culling, M.A. Stone: Energetic masking and masking release. In: J. Middlebrooks, J. Simon, A.N. Popper, R.R. Fay, Eds. The Auditory System at the Cocktail Party, Vol. 60, Springer Handbook of Auditory Research, Cham: Springer, 2017: 41–73.
78. L.S. Baltzell, J. Swaminathan, A. Cho, M. Lavandier, V. Best: Binaural sensitivity and release from speech-on-speech masking in listeners with and without hearing loss. Journal of the Acoustical Society of America 147, 3 (2020) 1546–1561.
79. K.D. Kryter: Methods for the calculation and use of the Articulation Index. Journal of the Acoustical Society of America 34, 11 (1962) 1689–1697.
80. The AMT Team: The auditory modeling toolbox full package (version 1.1.0) [code]. (2021). https://sourceforge.net/projects/amtoolbox/files/AMT1.x/amtoolbox-full-1.1.0.zip/download.
81. J. Rennies, T. Brand, B. Kollmeier: Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet. Journal of the Acoustical Society of America 130, 5 (2011) 2999–3012.
82. J. Rennies, A. Warzybok, T. Brand, B. Kollmeier: Modeling the effects of a single reflection on binaural speech intelligibility. Journal of the Acoustical Society of America 135, 3 (2014) 1556–1567.
83. A. Bregman: Auditory scene analysis, the perceptual organization of sound, The MIT Press, Cambridge, MA, 1990.
84. B.C.J. Moore, H. Gockel: Properties of auditory stream formation, Philos. Trans. R. Soc. B. 367, 1591 (2012) 919–931.
85. M. David, M. Lavandier, N. Grimault, A.J. Oxenham: Sequential stream segregation of voiced and unvoiced speech sounds based on fundamental frequency. Hearing Research 344 (2017) 235–243.

Cite this article as: Lavandier M. Vicente T. & Prud'homme L. 2022. A series of SNR-based speech intelligibility models in the Auditory Modeling Toolbox. Acta Acustica, **6**, 20.