

Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Contents lists available at ScienceDirect

Hearing Research

journal homepage: www.elsevier.com/locate/heares

Research paper

Revision and validation of a binaural model for speech intelligibility in noise

Sam Jelfs^a, John F. Culling^{b,*}, Mathieu Lavandier^c^aWelsh School of Architecture, Cardiff University, Bute Building, King Edward VII Avenue, Cardiff CF10 3NB, UK^bSchool of Psychology, Cardiff University, Tower Building, Park Place, Cardiff, CF10 3AT, UK^cUniversité de Lyon, Ecole Nationale des Travaux Publics de l'Etat, Département Génie Civil et Bâtiment (C.N.R.S.), Rue M. Audin, 69518 Vaulx-en-Velin Cedex, France

ARTICLE INFO

Article history:

Received 27 May 2010

Received in revised form

1 December 2010

Accepted 6 December 2010

Available online 13 December 2010

ABSTRACT

Lavandier and Culling [Lavandier, M. and Culling, J. F. 2010. Prediction of binaural speech intelligibility against noise in rooms. *J. Acoust. Soc. Am.* 127, 387–399] demonstrated a method of predicting human speech reception thresholds for speech in combined noise and reverberation. An updated version of the model is presented, which is substantially more computationally efficient. The updated model makes similar predictions for the SRT data considered by Lavandier and Culling, which tested the model's ability to predict effects of binaural unmasking and room colouration. In addition, we show here that the model accurately predicts the effects of headshadow and reproduces a range of data sets from the literature, including situations with multiple interfering sounds in anechoic conditions.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Lavandier and Culling (2010) proposed a method of predicting Speech Reception Threshold (SRT) measurements for speech in combined noise and reverberation. Like previous models (vom Hovel, 1984; Zurek, 1993; Beutelmann and Brand, 2006; Beutelmann et al., 2010), the method was based on a theory of binaural unmasking, combined with a model of better-ear listening. An updated version of the model has been developed which is substantially more computationally efficient.

The method used by Lavandier and Culling (and also here) is illustrated schematically in Fig. 1. The first pathway calculates the expected binaural advantage due to binaural unmasking (Hirsh, 1948; Licklider, 1948) using Equalization-Cancellation theory (Durlach, 1963, 1972; Culling, 2007) to predict the Binaural Masking Level Difference (BMLD). The second path was designed to predict the benefits of better-ear listening. Lavandier and Culling did not include experiments to test whether the effects of headshadow on better-ear listening could be predicted, but some of their conditions were strongly affected by room colouration, caused by reverberation, and they demonstrated that the model's second path was essential to correctly predict these effects. Combined, the two paths should thus account for the two cues associated with spatial unmasking (Bronkhorst and Plomp, 1988), which is an established mechanism for the separation of competing sounds (Plomp, 1976; Hawley et al., 2004; Culling et al., 2004), but the accurate

prediction of headshadow effects has not yet been demonstrated. Moreover, Lavandier and Culling only tested the model in situations with one interfering sound. The effectiveness of the model for stimuli that include headshadow and multiple, spatially distributed interferers is demonstrated here.

1.1. Specifics of the Lavandier and Culling model

The model of Lavandier and Culling took as input speech-shaped noise, which had been convolved by Binaural Room Impulse Response (BRIR) recordings, to create reverberant speech-shaped-noise interferers. Within each frequency channel, the waveforms were then processed through the two paths of the model independently. Different peripheral frequency channels were simulated by passing the waveforms through a gammatone filterbank (Patterson et al., 1987, 1988) with two filters per Equivalent Rectangular Bandwidth (ERB) (Moore and Glasberg, 1983). To simulate binaural unmasking, four 320-ms sections of the filtered waveforms were extracted from the left and right channels and cross-correlated using utilities from the |WAVE software suite (Culling, 1996). From the cross-correlation function, the interaural coherence of the noise interferer, ρ_i , (the maximum of the cross-correlation function) and the interaural phases of both target ϕ_t and interferer ϕ_i were calculated. These were taken from the delays at which the maxima of the separately evaluated cross-correlation functions occurred. Delay values were multiplied by the filter centre frequency in radians/s, ω , to obtain phases. From these three variables, the BMLD was calculated using the method of Culling et al. (2004, 2005) using the formula given in Culling et al. (2005),

* Corresponding author. Tel.: +44 29 2087 4523; fax: +44 29 20874858.
E-mail address: Cullingj@cf.ac.uk (J.F. Culling).

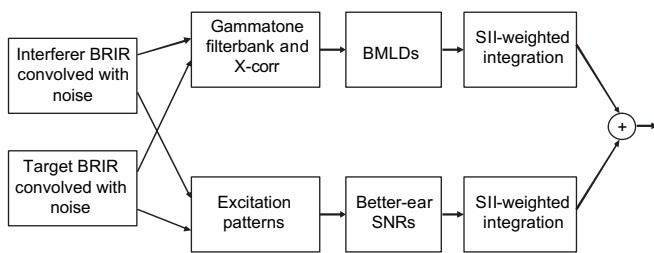


Fig. 1. Schematic illustration of the Lavandier and Culling (2010) model.

$$BMLD = 10 \log_{10} \left[\frac{k - \cos(\phi_i - \phi_t)}{k - \rho_i} \right] \quad (1)$$

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

where, $\sigma_\varepsilon = 0.000105$
 $\sigma_\delta = 0.25$

This equation was adapted from those of Durlach (1972) and uses the same values for the constants σ_δ and σ_ε . Also following Durlach's E-C theory, Lavandier and Culling assumed that binaural thresholds are never below their equivalent monaural thresholds, such that for any frequency channel where the formula returns a negative number the result is set to zero. To obtain the broadband binaural advantage for speech, the individual channel values were integrated across frequency using the Speech Intelligibility Index (SII) weighting method (ANSI, 1997). This approach was similar to that adopted by Levitt and Rabiner (1967) in their predictions of binaural unmasking for speech.

To compute the better-ear listening effect, Lavandier and Culling calculated the better-ear target-to-interferer ratio at each frequency. To do this they used the cochlear excitation patterns (Moore and Glasberg, 1983) for each section of the target and interferer waveforms, calculated between 0 and 33.25 ERBs (corresponding to 0–10 kHz) every 0.13 ERBs for both the left and the right ears. For each ear, the target-to-interferer ratio (TIR) at a given frequency was determined to be the difference between the target and interferer excitation patterns (in dB) at that frequency. The “better-ear” TIR at each frequency was then taken to be the greater of the left and right ear TIRs. The contribution of better-ear listening to spatial unmasking was taken to be the integration across frequency of these better-ear TIRs, using the SII weightings.

For both the binaural advantage and the better-ear TIR, the values were averaged over the four sections of the waveform and combined to give one single “effective” speech-to-noise ratio, minimising the stochastic effects introduced by convolving the BRIRs by random noise.

Since the output of the Lavandier and Culling model is an effective TIR, it does not directly predict SRT. However, the SRT is, by definition, a speech-to-noise ratio, so relative SRTs across different experimental conditions can be predicted by inverting the set of effective TIRs produced by the model and considering them either against a fixed reference condition or against the mean SRT of the data set. Unless one needs to model listeners with different receptive capacities or speech varying in intelligibility (differing in word frequency or the presence of syntactic semantic constraints), there is no requirement to calculate speech indices (AI, SII) or conduct index-to-intelligibility mapping (Beutelmann and Brand, 2006; Levitt and Rabiner, 1967) in order to validate the model. In Lavandier and Culling (2010), and in all of the analyses below, the observed and predicted SRTs were aligned by subtracting their average difference.

1.2. Revisions to the model

Two economies have been introduced to the computation in the revised model. First, the model operates directly upon BRIRs.

Second, the use of separately calculated excitation patterns has been replaced with further processing of the gammatone filter outputs.

For calculation of the BMLD, gammatone-filtered BRIRs can be directly cross-correlated, rather than first convolving the BRIRs with noise. This change not only saves the need to do the convolution, but it also produces precise non-stochastic results: there is no longer any need to average the measured parameters from several noise samples. If a large number of such parameters are averaged, their mean progressively approaches that found by directly cross-correlating the impulse responses. In order for this approach to be successful, however, it is important that the BRIRs are long enough for the gammatone filter to complete its response. In our experience, an impulse response of at least 1024 samples is needed for this reason.

For the calculation of the better-ear TIR, rather than generate separate excitation patterns, the same filtered BRIRs can be analysed further. The power of a waveform after convolution with an impulse response is proportional to the total energy of the impulse response, so the TIR at each ear can be accurately predicted from the energy ratios between the filtered impulse responses for target and interferer. Once again, this result is precise and non-stochastic, where previously it was an estimate.

Where target and interferer are in the same location, the output of the model will be 0 dB effective TIR (it is assumed, in the first instance, that target and interferer have equal power). If the locations differ only in azimuth, the output of the model will be a prediction of spatial unmasking, but if they vary also in distance, the effect of the attenuation with distance will also come into play. Where source locations differ, the effective TIR may also be influenced by room-acoustic effects, such as colouration.

Lavandier and Culling (2010) did not test their model's predictions of the effects of multiple interfering sources, but, had they done so, the different convolved noises would have been summed prior to cross-correlation and calculation of the excitation patterns. In order to achieve the same end directly from the impulse responses, a different approach is required. Instead, the different impulse responses must be concatenated (joined end-to-end). Concatenation has the effect of generating an averaged cross-correlation function (weighted according to the energy in each impulse response), and, of course, adding the energy of each contributing impulse response¹.

2. Validation of the revised model

For purposes of validation, the pattern of SRTs predicted by the model was compared with the pattern of SRTs observed in various experiments. The correspondence between prediction and observation was primarily assessed using scattergrams and correlation measurements, which illustrate the overall accuracy of the prediction, rather than their correspondence across the specific conditions involved in each experiment. Since correlation measures test for a linear relationship of any sort rather than specifically a 1:1 correspondence, the scattergrams are all displayed with a reference line of unit slope and passing through the origin. As noted above, because the output of the model is an effective TIR rather than an absolute prediction of threshold, SRTs were produced by inverting

¹ It may seem intuitively reasonable to add together the impulse responses, just as one would add together different interfering sounds. However, in the case of impulse responses, summing them together will result in spectral distortion due to interference. Concatenation is the appropriate approach when each interfering source is an independent noise; only if identical noise were to be used, should the impulse responses be summed.

Table 1

Goodness of fit between the model predictions and the observed data from various empirical studies. The goodness of fit is represented by the correlation between these two variables, the root-mean-squared error after their means are aligned and the size of offset required to bring the predicted and observed means into alignment.

Study	Correlation	rms error	Difference in means (predicted-observed)
Lavandier and Culling, 2010 (expt 1.)	0.91	0.4 dB	5.4 dB
Lavandier and Culling, 2010, (expt 2.)	0.98	0.3 dB	4.7 dB
Peissig and Kollmeier (1997)	0.98	1.2 dB	3.9 dB
Hawley et al., 2004, (monaural data)	0.99	0.4 dB	3.1 dB
Hawley et al., 2004, (binaural data)	0.99	0.6 dB	3.2 dB
Culling et al. (2004)	0.94	1.1 dB	3.1 dB
Bronkhorst and Plomp (1988)	0.86	1.5 dB	6.7 dB

the effective TIRs and then subtracting the mean difference between this set of numbers and the observed SRTs. Where SRTs were reported in the original studies (rather than just the magnitude of spatial unmasking) these differences in mean are presented in Table 1, along with the correlation between predicted and observed SRTs and the rms error of those predictions.

2.1. Lavandier and Culling (2010)

Lavandier and Culling measured SRTs for speech against a speech-spectrum noise using a 1-up/1-down adaptive threshold method (Plomp and Mimpen, 1979). Their target speech was sentences taken from the Harvard Sentence List (IEEE, 1969), and was always anechoic, whilst they varied the level of interaural coherence of the interferer. This was done through modelling virtual rooms using |WAVE, allowing them to vary the size of the room, the source distances within each room, the absorption coefficients of each of the room's materials and the azimuthal separation of the sources. Their room modelling did not incorporate a head, rather the "ears" of the listener were modelled as two omnidirectional microphones, 18 cm apart, and 1.5 m from the floor. Correlations of 0.95 and 0.97 were obtained between predicted and observed SRTs in their two experiments.

Having revised the model, it first was important to verify that the changes had not materially altered its established predictions. The conditions tested by Lavandier and Culling in their experiments were predicted using the same room impulse responses, but the revised model. However, after Lavandier and Culling convolved speech-shaped noise by these impulse responses, they also equalised the relative levels of speech and noise at each ear. Moreover, since the impulse responses from the |WAVE program do not have zero mean, the relative levels were evaluated from the power spectra for all frequency bins above 20 Hz. In order to simulate these features of the stimuli, the filter used to create their speech-shaped noise was applied to all the impulse responses and their relative energies above 20 Hz were obtained. These relative energies were then used to normalise the original (unfiltered) impulse responses before presenting them to the model.

The results of this modelling for both Lavandier and Culling's experiments are shown in Fig. 2, compared to the original data. For experiment 2 only the 15 conditions for which BRIRs were applied to the same sources were modelled. A 16th condition in that experiment involved interaurally uncorrelated noise, which was created by convolving the left- and right-ear impulse responses by independent sources. As the revised prediction model works

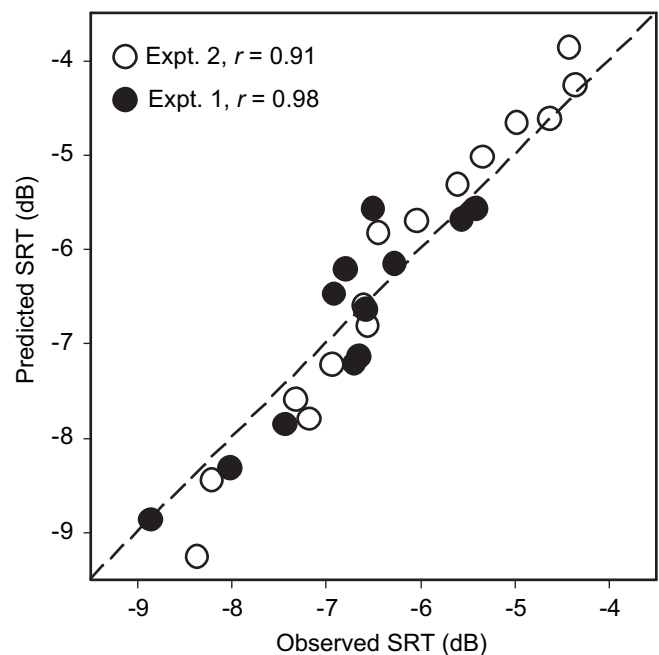


Fig. 2. Comparison of SRTs observed by Lavandier and Culling (2010) with the predictions of the revised model. Data and predictions for expt. 1 are shown with filled circles and those for expt. 2 with open circles. The dashed reference line is a line of unit slope passing through the origin and represents a 1:1 relationship between the predicted and observed SRTs.

directly from the BRIRs (a manoeuvre that assumes that the BRIRs would otherwise have been convolved with *identical* noise) it was unable to tackle this condition. The revised model predicted the data of Lavandier and Culling well; for experiment 1 the correlation coefficient between observed and predicted SRTs was 0.91 ($p < 0.0001$), while for experiment 2 it was 0.98 ($p < 0.0001$). These figures compare with 0.95 and 0.97 reported by Lavandier and Culling for the original method. The rms errors were 0.4 and 0.3 dB respectively.

2.2. Peissig and Kollmeier (1997)

Peissig and Kollmeier (1997) measured SRTs for speech with one, two, or three spatially separated noise interferers in anechoic conditions. They used a subjective adjustment method (Wesselkamp, 1994; Kollmeier and Wesselkamp, 1997) where listeners were allowed to alter the level of the target relative to the interferer to obtain a level that was subjectively judged to be 50% intelligible. The test material was two sentences taken from the Göttinger Satztest, a German sentence intelligibility test (Wesselkamp et al., 1992; Kollmeier and Wesselkamp, 1997). Virtual simulation of the different sound source azimuths was achieved using real-time convolution of the source sounds with the outer ear impulse responses for the respective angles, taken from Pösselt et al. (1986). For each of the one, two, or three interfering conditions the target was presented frontally (0° azimuth), while one interferer was positioned at a range of 17 azimuths in the horizontal plane. In the two-interferer condition a second fixed interferer was included at 105° azimuth, whilst in the three-interferer conditions two fixed interferers were included at 105° and 255° .

We predicted the resulting data using the revised method. None of the original test materials were available, so we employed generic materials. The different source azimuths were modelled using head-related impulse response (HRIRs) recordings from MIT

(Gardner and Martin, 1994). As in other cases considered here, the conditions simulated by Peissig and Kollmeier were anechoic, so the HRIRs could be used in place of BRIRs, because HRIRs are effectively BRIRs in an anechoic room. These HRIRs are 512 points long and sampled at 44.1 kHz. We added 1024 points of silence at the end of each HRIR in order to allow time for the gammatone filterbank's response to settle. Since the source locations varied only in azimuth, the output of the model was a predicted spatial unmasking for each condition. These were converted to relative SRTs by inversion, and aligned with the empirical data by subtracting the mean difference.

The results of the modelling can be seen in Fig. 3, compared to data scanned from their Figs. 2, 3 and 4. Different numbers of interferers are plotted with different symbols. Despite the use of generic materials, the correlation between the observed and the predicted SRTs is 0.98 ($p < 0.0001$) when comparing between all conditions. The rms error was 1.2 dB.

2.3. Hawley et al. (2004)

Hawley et al. (2004) measured SRTs using a variant of the Plomp and Mimpfen (1979) method using Harvard IEEE sentences. Again, SRTs were measured against one, two and three interferers. All sources were anechoic, with the target presented in front of the listener, at 0° azimuth, and interferers presented from one, two or three of a range of azimuths (−30, 0°, 30°, 60° and 90°). The virtual stimuli were created using HRIRs recorded from the HMS III acoustic manikin and distributed in the AUDIS database (Blauert et al., 1998). SRTs were measured for four different interfering source types: (1) other sentences spoken by the same voice as the target talker, (2) time-reversed sentences of the same talker, (3) speech-spectrum-shaped noise, and (4) speech-spectrum-shaped noise, modulated by the temporal envelopes of the sentence materials. SRTs were also measured for both binaural listening and monaural listening (left ear only).

Again, we predicted the data for the speech-shaped-noise interferers using HRIRs from the MIT database, the results of which

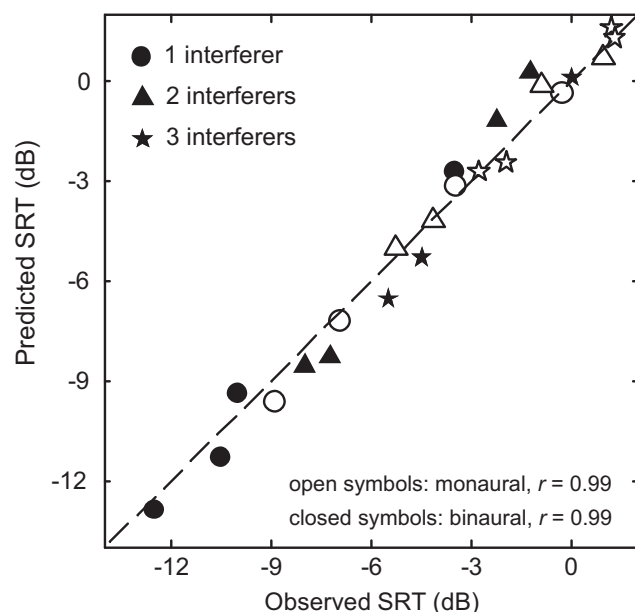


Fig. 4. As Fig. 3, but for the data of Hawley et al. (2004) with speech-shaped-noise interferers. Filled symbols are for binaural listening and open symbols are for monaural listening (left ear).

are shown in Fig. 4 compared to the original data. As with the Peissig and Kollmeier data, the spatial unmasking predictions from the model were inverted and aligned with the empirical data by subtracting the mean difference. The model accurately predicted variation in the observed data, with a correlation coefficient of 0.99 ($p < 0.0001$) when comparing across one, two, and three interfering conditions. The same correlation was observed for both monaural and binaural listening. The rms errors were 0.4 dB and 0.6 dB, respectively.

2.4. Bronkhorst and Plomp (1988)

Bronkhorst and Plomp (1988) measured SRTs using Plomp and Mimpfen (1979) method for speech with a single noise interferer at a range of azimuths (0°, 30°, 60°, 90°, 120°, 150°, and 180°). Bronkhorst and Plomp's experiment pre-dated the wide availability of HRIR data. Nonetheless, they sought to separate the individual spatial unmasking cues experimentally. They measured the power and phase spectra at each ear for noise sources presented from different directions around a KEMAR manikin in an anechoic chamber. Better-ear effects were then eliminated by taking a recording of noise presented at 0° azimuth and adding to the Fourier spectrum at each ear the appropriate frequency-dependent interaural delays found in their measurements (dT condition). Similarly, binaural unmasking was eliminated by shaping the power spectrum of that noise at each ear according to their measurements without adding the interaural delays (dL condition). Finally, the combination of both cues ("free-field" FF condition) was achieved simply by using the noise recordings from each azimuth.

As with the data of Peissig and Kollmeier and of Hawley et al., we predicted Bronkhorst and Plomp's data using the HRIRs from the MIT database. To predict the effects of better-ear listening (dL condition), only the values produced by the better-ear-listening pathway of the model were used, likewise, for the effects of binaural unmasking (dT condition), only the corresponding pathway was considered. Fig. 5 shows the predicted data as a scatter plot for the dT, dL, and FF conditions (different symbols), compared to data from Bronkhorst and Plomp's Table 1. The

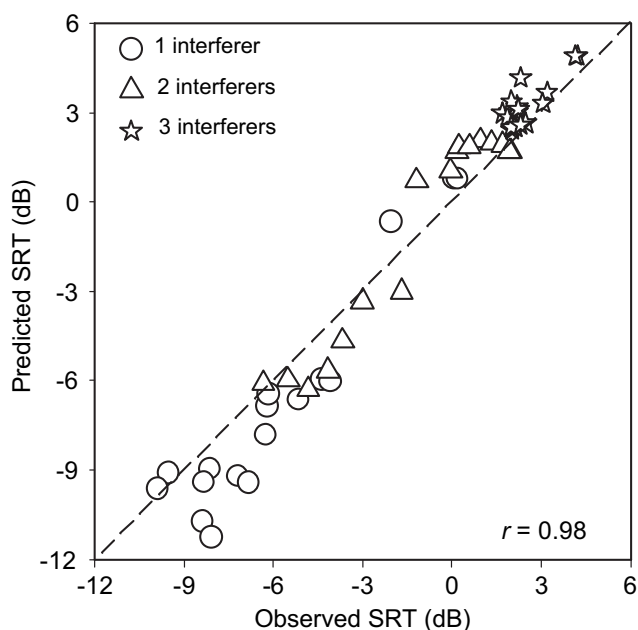


Fig. 3. As Fig. 2 but for the data of Peissig and Kollmeier (1997). Circles are for SRTs with one interferer, upright triangles are for SRTs with two interferers and stars are for SRTs with three interferers.

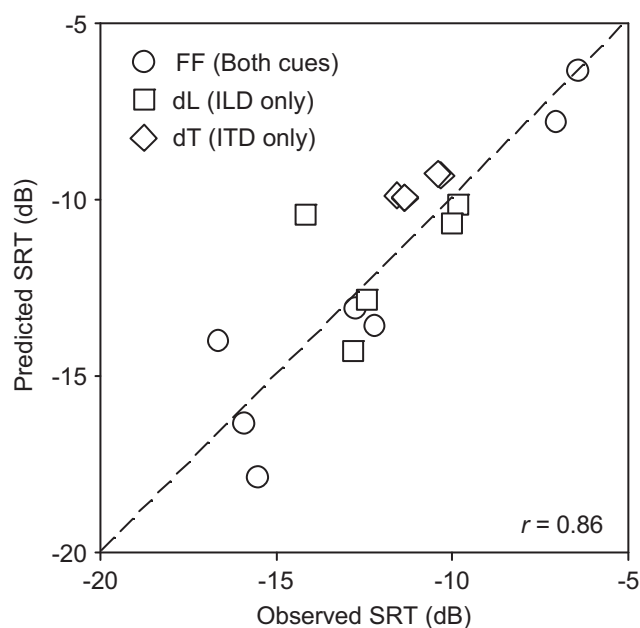


Fig. 5. As Figs. 2–4, but for the data of Bronkhorst and Plomp (1988). Circles are for the free-field condition in which a single noise interferer was presented from a range of azimuths, squares are for the dL condition in which interaural level cues were isolated, and diamonds are for the dT condition in which interaural temporal cues were isolated. The 0° and 180° conditions are plotted only as circles; Bronkhorst and Plomp did not repeat these angles in the dL and dT conditions, because the stimuli would have been identical.

correlation coefficient for all of the three cue types combined is 0.86 ($p < 0.0001$). The rms error was 1.5 dB.

This result is less convincing than for the preceding data sets. In order to gain an insight into the problems with this set of predictions, Fig. 6 shows the results from each condition as a function of azimuth. It can be seen that the main discrepancy occurs in the dL and FF conditions for an interfering noise at 90°. In this case, the model predicts that SRTs should be sharply elevated compared to adjacent azimuths of 60° and 120°, but the data show the 90° case to have an even lower SRT.

2.5. Culling et al. (2004)

Culling et al. (2004) measured SRTs for speech against three spatially located noise interferers using the same technique as Hawley et al. As in that companion study, all sources were generated by convolving the material with the relevant anechoic HRIRs from the HMS III acoustic manikin, from the AUDIS catalog (Blauert et al., 1998). However, like Bronkhorst and Plomp, they experimentally separated the two spatial-unmasking cues. In this case, the HRIRs from the HMS III manikin were manipulated. The phase or the amplitude spectra of the HRIRs were altered such that interaural differences in intensity or time delay were eliminated. These conditions were respectively termed ITD-only (equivalent to dT) and ILD-only (equivalent to dL). They used the same azimuth separations as those used by Hawley et al. for their three-interferer conditions.

The results of the SRT measurements were predicted using the same impulse responses as those used by Culling et al., having undergone the same processing methods. The predictions are shown in Fig. 7, separated into Both-cues (equivalent to FF), ILD-only (dL), and ITD-only (dT) conditions, and compared with the original data. The model predicted the observed data accurately,

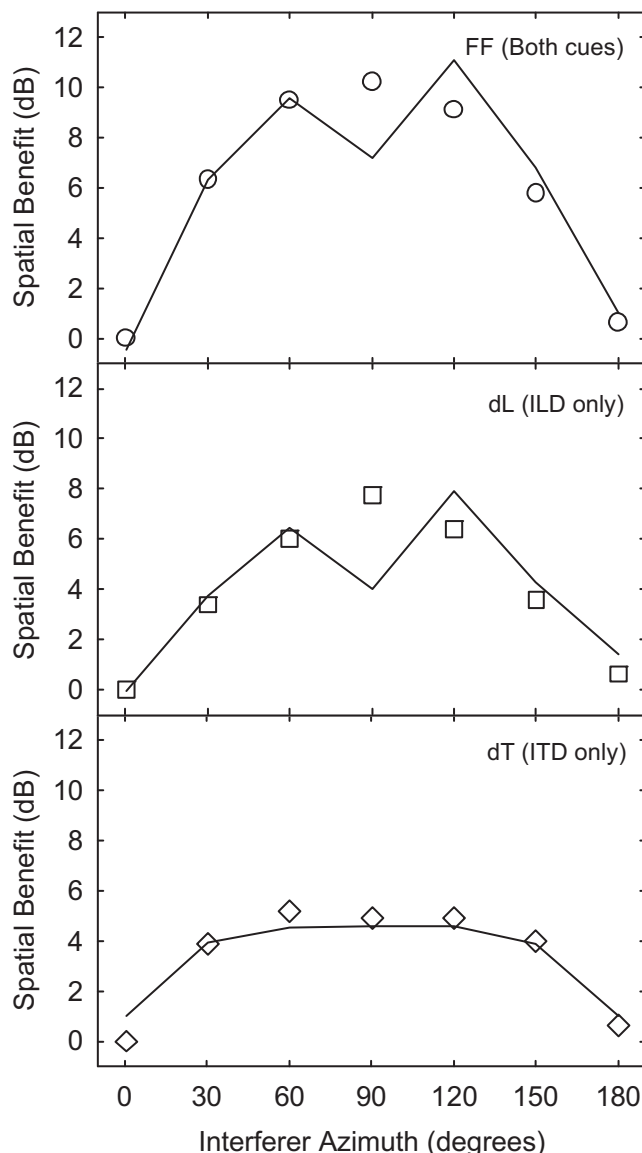


Fig. 6. The same data as on Fig. 5, but presented as a function of the azimuth of the interfering noise with different cue conditions on different panels. Prediction of the revised model are shown as lines.

with a correlation coefficient of 0.94 ($p < 0.0001$). The rms error was 1.1 dB.

3. Discussion

3.1. Performance of the model

Lavandier and Culling (2010) showed that their model was able to predict SRTs for anechoic target speech with a variety of reverberant single interferers. We used their data to validate the revised model, ensuring that it is able to take into account the importance of both azimuthal separation of the sources, as well as the importance of interferer interaural coherence. The revised model was still able to accurately predict their findings, whilst making significant savings with respect to the computational overhead. It should be noted that neither model involves the fitting of parameters; all the parameters used come from the literature. The frequency selectivity of the auditory system is taken from Moore and Glasberg

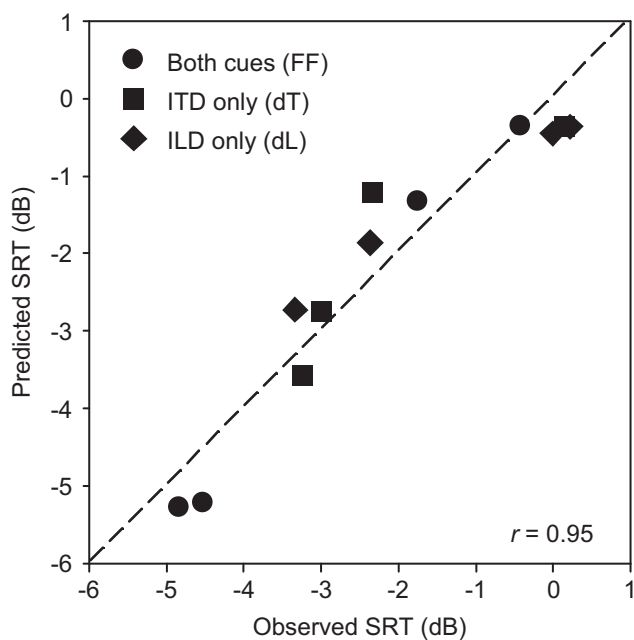


Fig. 7. As Fig. 5, but for the data of Culling et al. (2004). Filled circles are for the Both-cues condition in which a single noise interferer was presented from a range of azimuths, filled squares are for the ITD-only condition in which interaural temporal cues were isolated and, filled diamonds are for the ILD-only condition, in which interaural level cues were isolated.

(1983); the two “jitter” parameters of the E-C model are taken from Durlach (1972); the SII weightings are taken from the ANSI standard (ANSI, 1997); the acoustics of the head are taken from Gardner and Martin’s (1994) measurements from the KEMAR acoustic manikin.

Through modelling of the data of Peissig and Kollmeier and of Hawley et al., it was demonstrated that the model is also able to accurately predict anechoic listening situations with multiple spatially separated interferers, even when the HRIRs used in the model differ from those employed in the experiment. In the case of Peissig and Kollmeier, the model also accurately predicted the data set even though they used a different form of SRT measurement in a non-English language. The correlation of 0.99 for the monaural data from Hawley et al. also shows accurate modelling of the effects of headshadow, though not necessarily of better-ear listening which involves a choice of ears.

Hawley et al. (2004) discussed the potential for a “BE + BU” model (the additive combination of better-ear listening and binaural unmasking) to account for their data. In particular, they puzzled over the fact that binaural unmasking was robust when interfering sources were spatially distributed. Culling et al. (2004) continued that explanatory effort, by attempting to use E-C theory to predict the effects of binaural unmasking for the continuous noise case. They pointed out that when there are multiple interferers with different interaural time delays, the interaural coherence is not reduced at all frequencies, but becomes frequency dependent. Thus the binaural unmasking effect will fluctuate across frequency, but will still exist. The present manuscript represents a conclusion of that effort. The success of the model indicates that better-ear listening and binaural unmasking, combined in an additive way, are sufficient to account for the variation in SRTs across different spatial arrangements of continuous noise maskers.

The modelling of Bronkhorst and Plomp (1988) and Culling et al. (2004), for which better-ear listening and binaural unmasking

effects were tested separately, shows how well the two components of the model operate independently. A good correlation was observed between the predictions and the data of Culling et al. (2004), which also demonstrates that the model can correctly predict the effects of multiple interferers on each of these cues. However, the model was less able to predict the results of Bronkhorst and Plomp. Looking at Fig. 6, the points that were predicted badly are for the FF and dL conditions with the interferer at 90° to the listener. At this azimuth, the model predicted a drop in intelligibility (a higher SRT), which is not evident in the data. The model predicted this elevated SRT, because the underlying HRIRs show a relatively small difference in interaural level for sound sources at 90° azimuth. Consequently, the better-ear listening effect is predicted to be smaller for an interferer in that direction.

For an interferer at 90° one might expect the contralateral ear to be in a deep headshadow, but Gardner and Martin’s measurements indicate that it receives a remarkably undiminished sound level from the interferer. A high sound level received at the contralateral ear can be understood in terms of how sounds travel around the head. The path length of sounds travelling around different sides of an approximately spherical obstacle, like the head, are roughly equal, resulting in constructive interference on the far side, which raises the sound pressure level, and reduces the effectiveness of head shadowing (Duda and Martens, 1998). This effect was first observed acoustically by Lord Rayleigh (1880) and before that (1818) in optics, where it is referred to as the “Arago Spot” (or Poisson Spot). The underlying physics was first described by Babinet (1837). Elevated thresholds for a masker at 90° are clearly evident in the data from Peissig and Kollmeier (1997), (Fig. 2) and also account for the fact that Hawley et al.’s 90°, 90°, 90° condition yielded higher thresholds than the 30°, 60°, 90° condition, despite the binaural unmasking effect being somewhat compromised in the latter case. From this perspective, it seems surprising that Bronkhorst and Plomp did not observe poorer performance for an interferer at 90° than at, say, 60°, and the result was investigated further.

Fig. 8 compares the frequency response of each ear to sound sources at different azimuths derived from Gardner and Martin’s HRTFs (thin lines) with Bronkhorst and Plomp’s measurements. Bronkhorst and Plomp’s measurements are shown in their Fig. 2 as difference spectra for each ear, in which the head’s frequency response at a given angle is shown relative to that of a sound source presented from in front of the listener. These data have been scanned from their paper for inclusion in Fig. 8. Comparable plots were generated from Gardner and Martin’s HRIRs, by calculating the power spectrum in 1/6th oct. bands of the HRIRs for each ear and for sources at each azimuth and subtracting from these the power spectrum at the same ear for a frontal source.

It is evident that while the different plots are, for the most part, very consistent, Bronkhorst and Plomp’s measurements from the contralateral ear at 90° do not correspond well with the spectra from Martin and Gardner’s HRIRs. There is a large dip in the spectrum at around 3 kHz (indicated by an arrow in Fig. 8), which extends about 10 dB lower than any comparable feature from the Gardner and Martin measurements. Since Bronkhorst and Plomp’s stimuli were based on these measurements, it seems very likely that the differences are responsible for the mismatch between the observed and predicted SRTs in this case. It therefore appears that equivalent measurements from the same acoustic manikin have yielded substantially different results, possibly as a result of differences in spatial alignment.

3.2. Wider application of the model

All of the anechoic experiments modelled in this study involved continuous speech-shaped-noise interferers presented at equal

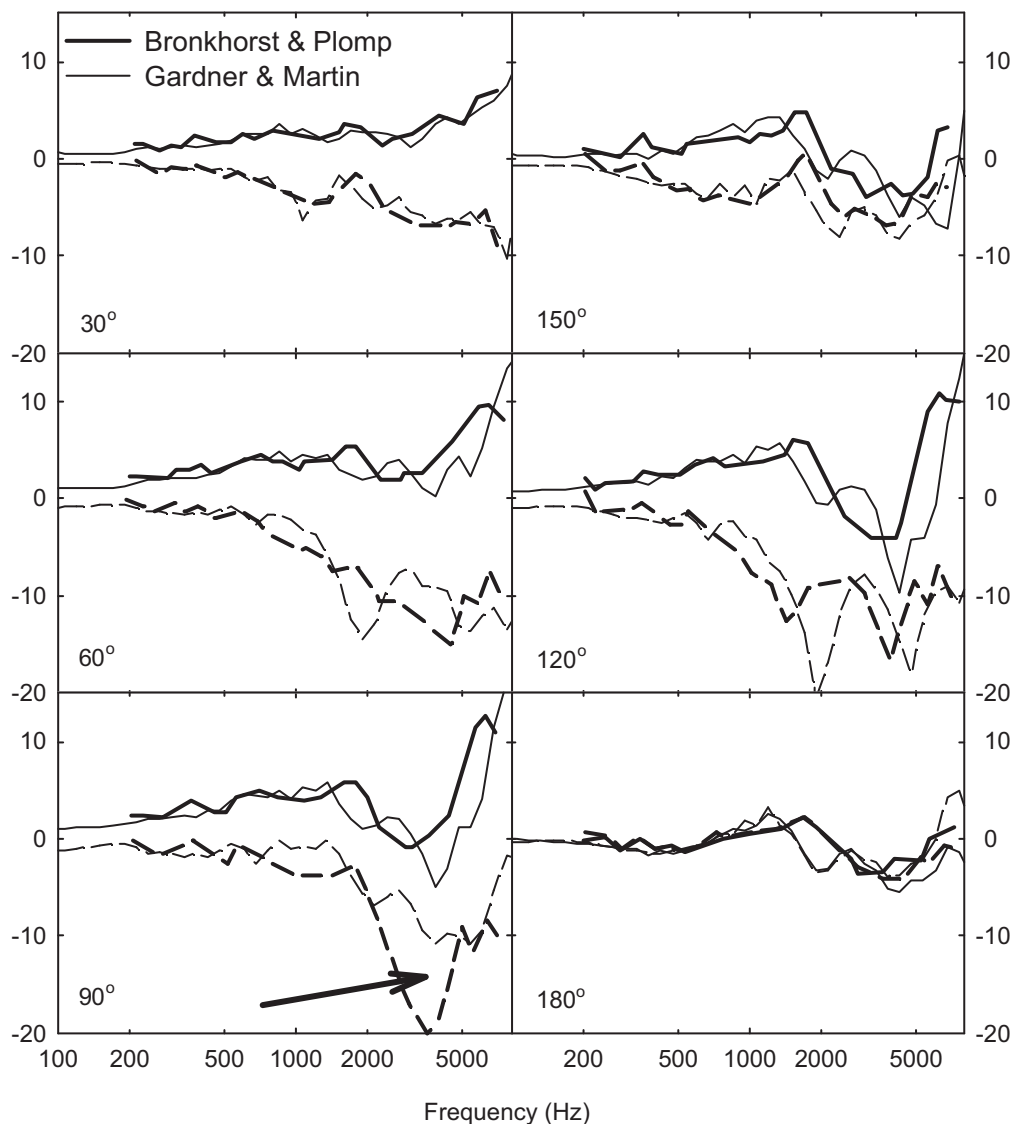


Fig. 8. Comparison of the differential frequency responses from Fig. 2 of Bronkhorst and Plomp (1988) (thick lines) with equivalent spectra generated from the HRIRs of Gardner and Martin (1994) (thin lines). Each line is the spectrum in 1/6th-octave bands received at a given ear for a given source direction, less the spectrum at the same ear for a source directly in front. Solid lines are spectra for the ear ipsilateral to the sound source and dashed lines are spectra for the ear contralateral to the sound source. The arrow on the 90° panel indicates an area of major discrepancy.

distances and equal sound levels, often simulated using HRIRs. They were also, effectively, at the same distance as the source of the target speech. Applying the model to variations on these constraints would be straightforward. Noises presented from different distances can be modelled by scaling their respective HRIRs in accordance with the inverse-square law prior to concatenation. In principle, continuous noise with a different source spectrum can easily be modelled by filtering each BRIR according to the difference spectrum between target and masker. Although we have not directly validated these manoeuvres, they follow logically from the effects of simple physical changes to the sound levels and interaural cross-correlations which would be received at the ear. Moreover, the better-ear-listening component of the model successfully exploits differences in the speech and noise spectra introduced to the BRIRs, by room colouration and headshadow. Since the processing of spectral differences from these sources of variation has already been validated, we see no reason why processing the BRIRs in this way should not also result in accurate prediction of the effect of different source spectra.

For the purpose of modelling performance in experiments with normally-hearing listeners in anechoic conditions, HRIRs from the Kemar (or similar) manikin are sufficient. Reverberant environments can be simulated using ray-tracing software in which BRIRs are constructed from the addition of many such HRIRs (selected according to the direction of incidence for each sound ray and delayed and scaled according to the ray's path length). For some applications it will be appropriate to use different HRIRs. Users of cochlear implants and hearing aids will have microphone positions which differ from those of an acoustic manikin, and recordings of in-situ HRIRs for these microphone positions will be needed. Such measurements could also capture the effects of any directional characteristics these microphones may possess.

3.3. Assumptions of the model

While some aspects of the model were strongly motivated by empirical data, other aspects were chosen more arbitrarily. The model assumes that each frequency channel operates entirely

independently and also that binaural unmasking and better-ear listening operate independently of each other.

At least three sources of experimental evidence support the suggestion that frequency channels are independent in binaural unmasking. First, Akeroyd (2004) found that detection of a complex tone was unaffected if each component had a different ITD. Second, Edmonds and Culling (2005a) showed that SRTs were unaffected if the target speech had different ITDs in different frequency bands. Third, Beutelmann et al. (2009) measured SRTs where the interaural phases of target speech and masking noise were modulated as a function of log frequency; a channel-independent model was required to produce a good fit to the results, but using a relatively broad channel bandwidth equivalent to 2.3 ERBs (Moore and Glasberg, 1983). In each case, the most important parameter was the difference in interaural phase between target and masking sound within each frequency channel.

The channel-independence of better-ear listening is less well supported. In fact, Edmonds and Culling (2006) found that, for a speech interferer, SRTs were substantially elevated when high and low frequency bands of the target and interfering speech were switched between the ears at 1500 Hz. This result suggests that listeners are obliged to listen to the same ear at all frequencies. This obligation may be associated with the listener's use of perceived position, which requires integration of binaural cues across frequency, in order to overcome informational masking (Shinn-Cunningham et al., 2005; Kidd et al., 2005). We are not aware of any similar experiments using noise, which would ultimately decide this matter, but the performance of the model in different circumstances provides some indication that channel-independence may be appropriate. Accurate predictions in anechoic conditions are probably not informative, because headshadow effects, which dominate the better-ear listening effect in anechoic conditions, tend to favour the same ear at all frequencies. However, the model was also successful at predicting the effect of room colouration in the Lavandier and Culling (2010) data, and here, where headshadow was not included in the stimuli, the relative levels of target and masker would have varied less systematically at each ear. This is currently the only evidence in favour of channel-independent better-ear listening.

The assumption that better-ear listening and binaural unmasking are independent and additive has some support from Edmonds and Culling (2005b), who showed that SRTs are unaffected when realistic interaural level differences and interaural time delays are placed in opposition rather than consistent with each other. However, Wan et al. (in press) have developed a model similar to ours in which the results of the two processes are not added together. Instead, their model selects the process which provides the better spatial unmasking effect at each frequency and uses only that value. One might call it a "BE or BU" model. Wan et al. have shown that their model can also make very accurate predictions of Hawley et al.'s data. The fact that either assumption works well may reflect the fact that binaural unmasking and better-ear listening tend to operate in different frequency regions (low frequencies for unmasking and high frequencies for better-ear listening), such that when they are added together, one of them is always negligible anyway. Once again, the matter can only be decided by a specifically designed experiment which ensures that both processes are active in the same frequency region.

3.4. Limitations of the model

The current version of the model is limited to cases involving continuous noise and anechoic or relatively close target sources. Moreover, it has only been tested in situations involving speech-shaped noise in anechoic conditions or in artificial room reverberation.

It would be desirable to extend the model so that it can deal with the effects of modulated and/or periodic interfering sounds such as competing speech. Both interferer modulation (de Laat and Plomp, 1983; Hawley et al., 2004) and interferer periodicity (Summers and Leek, 1998; Hawley et al., 2004) facilitate lower SRTs. These effects appear to be additional to those of binaural hearing. Indeed, Hawley et al. found evidence that there was a facilitative interaction between spatial unmasking and interferer periodicity; the spatial unmasking effect was bigger when speech or reversed-speech interferers were used. If the model could take into account these effects then it would be much closer to being able to predict performance in a real cocktail-party situation.

Lavandier and Culling (2010) tested the model using anechoic target speech. This decision was partly motivated by the need to test the effectiveness of the binaural unmasking component of the model. Since parameters of the interferer were critical to the unmasking component, only the reverberation of the interferer was manipulated. The experiments were thus conducted as though the target speech source was very close and so had a very high direct-to-reverberant ratio. A second reason to avoid reverberant speech was that reverberation is known to affect speech intelligibility due (in part) to temporal smearing effects, although such effects only occur at higher levels of reverberation than required to affect binaural unmasking (Lavandier and Culling, 2007, 2008). In order to accurately model speech reception in very reverberant rooms or in situations where the target speech is distant and so has a low direct-to-reverberant ratio, it will be necessary to incorporate some means of modelling these effects. Van Wijngaarden and Drullman (2008) have recently introduced a binaural version of the speech transmission index (Houtgast and Steeneken, 1985), which is able to make such predictions, but because this model interprets all modulations in the stimulus (monaural and binaural) as evidence of modulation from target speech, the model seems to be structurally incapable of accommodating modulation of the interferer.

Because the model is based on human hearing, real-room reverberation, including natural absorption characteristics and diffuse reflections should not, in principle, produce additional problems, but it would nonetheless be desirable to validate the model with real-room BRIRs and with more complex room shapes than the simple rectangular and specularly reflecting rooms with which it has been tested so far.

4. Conclusions

The new version of the model, while being much more computationally efficient, produces similar results to those reported by Lavandier and Culling (2010). Indeed, since the output is now non-stochastic, one should expect the results to be slightly more accurate.

By predicting a number of other data sets from the literature it has been possible to validate the model's components of binaural unmasking and better-ear listening both in isolation and in combination, and also in situations that have one, two, or three anechoic interferers.

It remains to validate the model in real rooms and refine it so that it accounts for 1) the temporal smearing effect of reverberation on the target speech 2) modulated interferers and 3) periodic interferers.

Acknowledgements

The authors would like to thank the Associate Editor, Birger Kollmeier and one anonymous reviewer for their valuable comments on earlier versions of this paper.

References

- Akeroyd, M.A., 2004. The across frequency independence of equalization of interaural time delay in the equalization-cancellation model of binaural unmasking. *J. Acoust. Soc. Am.* 116, 1135–1148.
- ANSI, 1997. S3.5. Methods for the Calculation of the Speech Intelligibility Index. American National Standards Institute, New York.
- Babinet, M., 1837. Mémoires d'optique météorologique. *C. R. Acad. Sci.* 4, 638–648.
- Beutelmann, R., Brand, T., 2006. Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.* 120, 331–342.
- Beutelmann, R., Brand, T., Kollmeier, B., 2009. Prediction of binaural speech intelligibility with frequency-dependent interaural phase differences. *J. Acoust. Soc. Am.* 126, 1359–1368.
- Beutelmann, R., Brand, T., Kollmeier, B., 2010. Revision, extension and evaluation of a binaural speech intelligibility model. *J. Acoust. Soc. Am.* 127, 2479–2497.
- Blauert, J., Brueggen, M., Bronkhorst, A.W., Drullman, R., Reynaud, G., Pellieux, L., Kriebler, W., Sottok, R., 1998. The AUDIS catalog of human HRTFs (A). *J. Acoust. Soc. Am.* 103, 3082.
- Bronkhorst, A.W., Plomp, R., 1988. The effect of head-induced interaural time and level differences on speech intelligibility in noise. *J. Acoust. Soc. Am.* 83, 1508–1516.
- Culling, J.F., 1996. Signal-processing software for teaching and research in psychoacoustic under unix and x-windows. *Beh. Res. Meth., Inst. Comp.* 28, 376–382.
- Culling, J.F., 2007. Evidence specifically favoring the equalization-cancellation theory of binaural unmasking. *J. Acoust. Soc. Am.* 122, 2803–2813.
- Culling, J.F., Hawley, M.L., Litovsky, R.Y., 2004. The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources. *J. Acoust. Soc. Am.* 116, 1057–1065.
- Culling, J.F., Hawley, M.L., Litovsky, R.Y., 2005. Erratum: the role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources. *J. Acoust. Soc. Am.* 118, 552.
- de Laat, J.A.P.M., Plomp, R., 1983. The reception threshold of interrupted speech. In: Kinke, R., Hartman, R. (Eds.), *Hearing: Physiological Bases and Psychophysics*. Springer, pp. 359–363.
- Duda, R.O., Martens, W.L., 1998. Range dependence of the response of a spherical head model. *J. Acoust. Soc. Am.* 104, 3048–3058.
- Durlach, N.I., 1963. Equalization and cancellation theory of binaural masking-level differences. *J. Acoust. Soc. Am.* 35, 1206–1218.
- Durlach, N.I., 1972. Binaural signal detection: equalization and cancellation theory. In: Tobias, J. (Ed.), *Foundations of Modern Auditory Theory*, vol. 2. Academic, New York, pp. 371–462.
- Edmonds, B.A., Culling, J.F., 2005a. The spatial unmasking of speech: evidence for within-channel processing of interaural time delay. *J. Acoust. Soc. Am.* 117, 3069–3078.
- Edmonds, B.A., Culling, J.F., 2005b. Effect of conflicting head-related time and level differences on spatial unmasking of speech. *Act. Acust. u. Acust.* 91, 546–553.
- Edmonds, B.A., Culling, J.F., 2006. The spatial unmasking of speech: evidence for better ear listening. *J. Acoust. Soc. Am.* 120, 1539–1545.
- Gardner, B., Martin, K., 1994. HRTF Measurements of a Kemar Dummy-head Microphone. Technical report. MIT Media Lab.
- Hawley, M.L., Litovsky, R.Y., Culling, J.F., 2004. The benefit of binaural hearing in a cocktail party: effect of location and type of interferer. *J. Acoust. Soc. Am.* 115, 833–843.
- Hirsh, I.J., 1948. The influence of interaural phase on interaural summation and inhibition. *J. Acoust. Soc. Am.* 20, 536–544.
- Houtgast, T., Steeneken, H.J.M., 1985. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J. Acoust. Soc. Am.* 77, 1069–1077.
- IEEE, 1969. IEEE recommended practice for speech quality measurements. *IEEE Trans. Aud. Electro.* 17, 227–246.
- Kidd Jr., G., Mason, C.R., Brughera, A., Hartmann, W.M., 2005. The role of reverberation in release from masking due to spatial separation of sources for speech identification. *Act. Acust. u. Acust.* 91, 526–535.
- Kollmeier, B., Wesselkamp, M., 1997. Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. *J. Acoust. Soc. Am.* 102, 2412–2421.
- Lavandier, M., Culling, J.F., 2007. Speech segregation in rooms: effects of reverberation on both target and interferer. *J. Acoust. Soc. Am.* 122, 1713–1723.
- Lavandier, M., Culling, J.F., 2008. Speech segregation in rooms: monaural, binaural and interacting effects of reverberation on target and interferer. *J. Acoust. Soc. Am.* 123, 2237–2248.
- Lavandier, M., Culling, J.F., 2010. Prediction of binaural speech intelligibility against noise in rooms. *J. Acoust. Soc. Am.* 127, 387–399.
- Levitt, H., Rabiner, L.R., 1967. Predicting binaural gain in intelligibility and release from masking for speech. *J. Acoust. Soc. Am.* 42, 820–829.
- Licklider, J.C.R., 1948. The influence of interaural phase relations upon the masking of speech by white noise. *J. Acoust. Soc. Am.* 20, 150–159.
- Moore, B.C.J., Glasberg, B.R., 1983. Suggested formulae for calculating auditory filter bandwidths and excitation patterns. *J. Acoust. Soc. Am.* 74, 750–753.
- Patterson, R., Nimmo-Smith, I., Holdsworth, J., Rice, P., 1987. An efficient auditory filterbank based on the gammatone function. In: *Institute of Acoustics Speech Group on Auditory Modelling*. Royal Signal Research Establishment.
- Patterson, R., Nimmo-Smith, I., Holdsworth, J., Rice, P., 1988. Spiral vos final report, part a: The auditory filterbank, cambridge electronic design, contract report (APU 2341).
- Peissig, J., Kollmeier, B., 1997. Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners. *J. Acoust. Soc. Am.* 101, 1660–1670.
- Plomp, R., Mimpen, A., 1979. Improving the reliability of testing the speech-reception threshold for sentences. *Audiology* 18, 43–52.
- Plomp, R., 1976. Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of azimuth of a single competing sound source (speech or noise). *Acustica* 34, 200–211.
- Pösselt, C., Schrter, J., Opitz, M., Divenji, P., Blauert, J., 1986. Generation of binaural signals for research and home entertainment. *Proceedings of the 12th International Congress on Acoustics*. Toronto. 1: B1–6.
- Rayleigh, Lord, 1880. The acoustical shadow of a circular disk. *Phil. Mag.* 9, 278–283.
- Shinn-Cunningham, B.G., Ihlefeld, A., Satyavarta, Larson, E., 2005. Bottom-up and top-down influences on spatial unmasking. *Act. Acust. u. Acust.* 91, 967–979.
- Summers, V., Leek, M.R., 1998. Masking of tones and speech by Schroeder-phase harmonic complexes in normally hearing and hearing-impaired listeners. *Hear. Res.* 118, 139–150.
- van Wijngaarden, S.J., Drullman, R., 2008. Binaural intelligibility prediction based on the speech transmission index. *J. Acoust. Soc. Am.* 123, 4514–4523.
- von Hövel, H., 1984. Zur Bedeutung der Übertragungseigenschaften des Außenohrs sowie des binauralen Hörsystems bei gestörter Sprachübertragung (On the importance of the transmission properties of the outer ear and the binaural auditory system in disturbed speech transmission), Ph.D. thesis, RWTH, Aachen.
- Wan, R., Durlach, N. I., Colburn, H. S., in press. Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers. *J. Acoust. Soc. Am.*
- Wesselkamp, M., 1994. Messung und Modellierung der Verständlichkeit von Sprache, PhD thesis.
- Wesselkamp, M., Kliem, K., Kollmeier, B., 1992. Erstellung eines optimierten satztestes in deutscher sprache. In: Kollmeier, B. (Ed.), *Moderne Verfahren der Sprachaudiometrie*. Median-Verlag, Heidelberg, pp. 330–343.
- Zurek, P., 1993. Binaural advantages and directional effects in speech intelligibility. In: Studebaker, G., Hochberg, I. (Eds.), *Acoustical Factors Affecting Hearing Aid Performance*. Allyn and Bacon, Needham Heights, MA., pp. 255–276.