Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking, and binaural de-reverberation

Thibaud Leclère, Mathieu Lavandier, John F. Culling

Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking, and binaural de-reverberation

Thibaud Leclère and Mathieu LavandierJohn F. CullingELP

Citation: The Journal of the Acoustical Society of America **137**, 3335 (2015); doi: 10.1121/1.4921028
View online: http://dx.doi.org/10.1121/1.4921028
View Table of Contents: http://asa.scitation.org/toc/jas/137/6
Published by the Acoustical Society of America

Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking, and binaural de-reverberation

Thibaud Leclère[a] and Mathieu Lavandier
Université de Lyon, École Nationale des Travaux Publics de l'État, Laboratoire Génie Civil et Bâtiment, Rue M. Audin, 69518 Vaulx-en-Velin Cedex, France

John F. Culling
School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff CF10 3AT, United Kingdom

(Received 28 July 2014; revised 15 April 2015; accepted 20 April 2015)

Room acoustic indicators of intelligibility have focused on the effects of temporal smearing of speech by reverberation and masking by diffuse ambient noise. In the presence of a discrete noise source, these indicators neglect the binaural listener's ability to separate target speech from noise. Lavandier and Culling [**(2010)**. J. Acoust. Soc. Am. **127**, 387–399] proposed a model that incorporates this ability but neglects the temporal smearing of speech, so that predictions hold for near-field targets. An extended model based on useful-to-detrimental (U/D) ratios is presented here that accounts for temporal smearing, spatial unmasking, and binaural de-reverberation in reverberant environments. The influence of the model parameters was tested by comparing the model predictions with speech reception thresholds measured in three experiments from the literature. Accurate predictions were obtained by adjusting the parameters to each room. Room-independent parameters did not lead to similar performances, suggesting that a single U/D model cannot be generalized to any room. Despite this limitation, the model framework allows to propose a unified interpretation of spatial unmasking, temporal smearing, and binaural de-reverberation. © *2015 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4921028]

[ELP] Pages: 3335–3345

I. INTRODUCTION

Speech intelligibility is impaired in noisy rooms by both noise and reverberation. The speech signal is mixed with delayed versions of itself reflected by room boundaries: the speech can be smeared and self-masked (Bradley, 1986; Houtgast and Steeneken, 1985). In the presence of discrete noise sources, a listener is able to partly separate target speech from masking noise using the binaural system. This ability is impaired by reverberation (Beutelmann and Brand, 2006; Culling *et al.*, 2003; Plomp, 1976). The corresponding loss of intelligibility appears at lower levels of reverberation, and thus occurs more readily, than the loss of intelligibility associated with the smearing of speech (Lavandier and Culling, 2008). The aim of the present study was to propose and validate a model predicting these multiple effects.

Architectural acoustic indicators of intelligibility have focused on the effects of temporal smearing of speech and masking by diffuse ambient noise. The speech transmission index (STI) measures the reduction of amplitude modulation in the speech signal due to reverberation and noise (Houtgast and Steeneken, 1985). The useful-to-detrimental (U/D) ratio computes a signal-to-noise ratio (SNR) in which the early reflections of the target are regarded as useful and as the "signal" because they reinforce the direct sound (Bradley *et al.*, 2003), while the late reflections are regarded as detrimental and effectively a part of the noise (Bradley *et al.*, 1999; Bradley, 1986; Lochner and Burger, 1964). These monaural indicators neglect the listener's ability to separate target speech from interfering sounds using the binaural system as well as the susceptibility of this ability to reverberation.

In the presence of discrete noise sources, masking is less efficient when the target and noise sources are on different bearings (Hawley *et al.*, 2004; Plomp, 1976). This spatial release from masking is based on two mechanisms (Bronkhorst and Plomp, 1988): better-ear listening and binaural unmasking, which rely on interaural level and time differences (ILDs and ITDs), respectively. Target and interferers at different locations often produce different ILDs so that one ear usually offers a better SNR than the other, and listeners can attend to the ear offering the better ratio. Differences in the ITD generated by target and interferer facilitate binaural unmasking in which the auditory system is able to "cancel" to some extent the noise generated by the interferer [equalization-cancellation (EC) theory; Durlach, 1972], thus improving the internal SNR. Both processes are affected by reverberation. Sound reflections traveling around the listener reduce the acoustic shadowing by the head (Plomp, 1976) and impair binaural unmasking mainly by decorrelating the interfering noise at the two ears (Lavandier and Culling, 2008).

Beutelmann and Brand (2006) implemented this binaural ability into a model of speech intelligibility. Simulated stimuli at the ears are processed through a gammatone filterbank and an EC stage, then re-synthesized, and the speech

[a]Electronic mail: thibaud.leclere@entpe.fr

intelligibility index (SII) method is used to evaluate intelligibility (ANSI, 1997). For each frequency band of the gammatone filterbank, the EC stage directly implements a mechanism based on EC theory, testing different delays and attenuations for the signals at the ears and choosing those maximizing the SNR. Lavandier and Culling (2010) developed a prediction model also based on EC theory, but the better-ear listening and binaural unmasking are computed separately. The direct implementation of cancellation is replaced by a predictive equation, extending the models of Levitt and Rabiner (1967) and Zurek (1993). Binaural unmasking prediction and better-ear target-to-interferer ratio are added and weighted across frequency with the SII-importance band coefficients. Like in the model of Beutelmann and Brand (2006), the prediction method is based on the signals in the room, requiring averaging across signals to produce reliable predictions. Beutelmann *et al.* (2010) revised their original model by improving the computational EC stage with an analytical expression instead of using probabilistic methods. The model of Lavandier and Culling (2010) was also revised by directly applying the model to binaural room impulse responses (BRIRs) instead of signals, thus producing non-stochastic predictions (Lavandier *et al.,* 2012; Jelfs *et al.,* 2011). Like the model of Beutelmann and Brand (2006), the model of Wan *et al.* (2010) uses a direct implementation of an EC process but with time-varying jitters in time and amplitude and monaural pathways in addition to the binaural pathway. All these binaural models neglect the temporal smearing of speech by reverberation, so their predictions only hold for near-field targets with a high direct-to-reverberant (D/R) ratio.

Van Wijngaarden and Drullman (2008) introduced a binaural version of the STI. This approach makes the assumption that the target is the only source of modulation at the listener's ears, so that it does not offer any opportunity for extension to modulated noise (Collin and Lavandier, 2013; Beutelmann *et al.,* 2010) or speech interferers. In these cases, the modulation is coming from both target and interferer. Rennies *et al.* (2011) extended the model of Beutelmann *et al.* (2010) to take the smearing effect of reverberation into account using three alternatives: the modulation transfer function (MTF), the definition factor (D_{te}, ISO, 1997), and the U/D ratio. In the first two approaches, spatial unmasking and temporal smearing are processed separately: the SNRs obtained with their binaural model applied to the entire speech and noise signals are corrected *a posteriori* by either measuring the MTF or D_{te} of the target room impulse response. In the third approach, this impulse response is split into early and late parts that are convolved with the speech signal to create an "early speech" signal and a "late speech" signal. The prediction process is then similar to that of Beutelmann *et al.* (2010) except that the original target signal is replaced by the early speech and the late speech is added to the interferer, so that the detrimental influence of late reflections is taken into account before the binaural process. Rennies *et al.* (2014) tested these three approaches on the data of Warzybok *et al.* (2013) that involved a frontal target smeared by a single reflection. They introduced a weighting function to separate early and late reflections within the impulse response (with the D_{te} and U/D extensions). These modelings allowed them to retain the U/D approach as the most suitable to account for the temporal smearing of speech.

The present study aimed to test the U/D approach to extend the validity of a different binaural model framework (Lavandier and Culling, 2010). In the literature, U/D models are based on a wide range of values/methods to separate early and late reflections (Rennies *et al.,* 2014; Rennies *et al.,* 2011; Bradley *et al.,* 2003; Soulodre *et al.,* 1989; Bradley, 1986; Lochner and Burger, 1964). So, this study further investigated the influence of the early/late separation (see Sec. II B), using realistic reverberation from different rooms.

None of the binaural models presented in the preceding text have ever been shown to predict the "squelching" effect of binaural hearing. In the literature, the term "binaural squelch" has been used to describe the general advantage of binaural hearing over monaural hearing (Koenig, 1950) or the binaural advantage when better-ear listening has been taken into account (Bronkhorst and Plomp, 1988). However, this last advantage is also sometimes referred to as "binaural unmasking" or "binaural interaction." To avoid any ambiguity, the term "binaural de-reverberation" will be preferred to binaural squelch here. It will refer hereafter to the benefit from binaural listening compared to diotic/monaural listening in reverberation even in the absence of an interfering source. This benefit has been shown to slightly improve intelligibility for reverberant speech in quiet (Nábělek and Robinson, 1982; Moncur and Dirks, 1967). Such a small but significant binaural advantage was also measured by Lavandier and Culling (2008) in the presence of a noise interferer. Binaural speech led to lower thresholds than diotic speech. Because binaural unmasking from the noise was probably not affected by the target listening mode in this configuration, the authors concluded that the result could be explained by the binaural de-reverberation observed in quiet.

An integrated model is proposed here to account for speech transmission (and temporal smearing), spatial unmasking from noise interferers, and binaural de-reverberation as defined in the preceding text. The predictions were compared with speech reception thresholds (SRTs, level of the target compared to that of the interferer for 50% intelligibility) measured in three experiments from the literature (Rennies *et al.,* 2011; Lavandier and Culling, 2008), in which spatial unmasking and target smearing were both simultaneously involved. Two versions of the model were tested: a room-dependent (RD) model the parameters of which were adjusted in each room and a room-independent (RI) model with fixed parameters across rooms. The RI model was tested on a fourth dataset that involved several rooms not used to define its parameters (van Wijngaarden and Drullman, 2008).

II. THE INTEGRATED MODEL

A. Model structure

Because the U/D approach requires the target BRIR as input, the present study extends the model of Lavandier and Culling (2010) in its implementation based on the BRIRs

measured between the sources and listener positions (referred to as "old model" in this paper; Jelfs *et al.*, 2011; Lavandier *et al.*, 2012) rather than the last version proposed by Collin and Lavandier (2013) that is not applied to BRIRs but to the signals within short-time frames. The target BRIR is first separated into an early and a late part (see Sec. II B for details). The early part constitutes the useful component. The late part is combined with the BRIRs of the interferers to form the detrimental component. These BRIRs are concatenated rather than added to preserve phase information and avoid constructive/destructive interference (Jelfs *et al.*, 2011). The binaural model is then applied to the useful and detrimental components in the same way as it was previously applied to the target and interferer BRIRs. The detailed implementation of the old model is not described here, but it can be summarized by three steps: (1) gammatone filtering, (2) computation of the better-ear listening and binaural unmasking, (3) SII weightings (ANSI, 1997). Better-ear listening is estimated from the U/D energy ratios computed as a function of frequency at each ear, selecting the ear for which the ratio is higher. Binaural unmasking is estimated from the binaural masking level difference (BMLD) computed using the interaural phase differences of the useful and detrimental parts and the interaural coherence of the detrimental part (Lavandier *et al.*, 2012, Eq. 1.2). The resulting better-ear U/D ratios and BMLDs (in decibels) are SII weighted, integrated across frequency and summed to provide a broadband binaural U/D ratio.

To be compared with SRTs, which are by definition SNRs, binaural U/D ratios are inverted, so that high ratios correspond to low thresholds. Differences in inverted ratios can be directly compared to SRT differences, or a reference is chosen for the comparison. The reference here was the averaged SRT across conditions for each experiment.

B. Early/late separation parameters

Useful and detrimental signals are obtained by splitting the target BRIR into early and late parts. This separation uses two temporal weighting windows: the early and the late windows that isolate the early and late parts, respectively, by multiplying the original impulse response by the window in the time domain. Here, early and late windows are always defined to be complementary, such that their sum is always 1 (Fig. 1).

Before the early/late separation, the direct sound was defined as the earliest sound at the ears. A recursive algorithm was applied to each BRIR channel (left and right) to locate the direct sound, and then the earlier of the two was taken as the unique direct arrival time of the BRIR. The algorithm found the first sample, which is at least 25% greater than all previous samples in the BRIR channel. This algorithm was used because taking the maximum value or the first non-zero sample in the BRIR could induce biases in the direct sound arrival time (if a combination of reflections is stronger than the direct sound or if some ambient noise is recorded before the impulse).

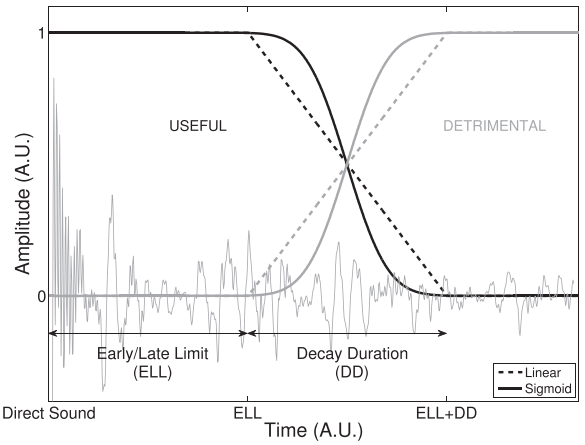

FIG. 1. Illustration of the temporal weighting windows tested in the present study. Black curves represent the early windows, whereas the gray curves represent the late windows. Samples in the impulse response are either considered as fully useful [before the early/late limit (ELL)], fully detrimental [beyond ELL + decay duration (DD)], or partially useful (during DD). The rectangular window is a linear window with a null DD and ELL as a unique parameter.

The rectangular window is the most usual way to split an impulse response into early and late parts. The early part is defined as the original impulse response until a temporal limit, beyond this limit, the samples of the window are set to zero. This early/late limit (ELL) is relative to the direct sound and is the only parameter required for the rectangular window. Despite the simplicity of this window, the frontier between useful and detrimental is very sharp and thus two reflections can be considered very differently even if they are separated with only few samples. Warzybok *et al.* (2013) highlighted this limitation in the presence of a single reflection. The work of Lochner and Burger (1964) showed that only a part of the energy of early reflections can be considered as "useful" regarding speech intelligibility. Rennies *et al.* (2014) also tested a linearly decaying window to separate early and late reflections. Two window shapes with a progressive weighting of reflections across time were thus tested: the linear window and the sigmoid[1] window (see Fig. 1), which both have a decay duration (DD) parameter in addition to ELL. These temporal parameters are here defined differently than in Rennies *et al.* (2014). ELL defines the duration of the flat part of the window, whereas DD is the duration of the decrease starting from one at ELL and ending at zero at ELL + DD (Fig. 1). With these definitions, a rectangular window is a linear window with DD = 0 ms.

Three parameters were thus tested concerning the separation of early and late parts of the target BRIR: ELL, DD, and window shape.

III. VALIDATION OF THE ROOM-DEPENDENT MODEL AND DEFINITION OF THE ROOM-INDEPENDENT MODEL

A. Data from the literature

The model predictions were compared to SRTs measured using headphones in three experiments (Rennies *et al.*, 2011; Lavandier and Culling, 2008) with one target source

(connected speech) in competition with one interferer source (speech-spectrum noise). The three modeled experiments are briefly presented to describe the effects which need to be predicted by the proposed model: spatial unmasking, temporal smearing, and binaural de-reverberation. More details are available in the original publications.

1. Temporal smearing and spatial unmasking

In their experiment 1 (referred to as RBK in the following), Rennies *et al.* (2011) measured SRTs across 12 conditions in a virtual room. The reverberation level was varied by moving the listener away from the fixed frontal target (0.5, 1.5, 3.5, and 13 m). For each distance, the single interferer source was placed either frontally, at 22.5°, or at 90° to the right of the listener. The distances between listener and each source were generally the same for all listener positions. Because both the azimuth of the interferer source and the reverberation level on the target varied across conditions, both spatial unmasking and temporal smearing were observed in the results.

In experiment 3 (referred to as LC3 in the following) of Lavandier and Culling (2008), the listener was facing a target and an interferer source spatially separated at fixed positions (65° to the left and right of the listener's head) in a virtual room whose absorption coefficients were set to four values: 1 (anechoic room), 0.7, 0.5, and 0.2. The reverberation level was varied across conditions, independently for target and interferer, such that intelligibility was disrupted by both the smearing of target speech and the reduction of binaural unmasking due to reverberation on the interferer.

2. Binaural de-reverberation

In their experiment 4 (referred to as LC4 in the following), Lavandier and Culling (2008) simulated the sources and listener at fixed positions in a virtual room (slightly wider than in LC3). The interferer source was located at 65° on the right of the listener's head while the target was straight ahead. The absorption coefficient of the room boundaries was fixed to 0.5 for the interferer, while two coefficients (1 and 0.2) were tested for the target. The interferer was always binaural, whereas the target was either binaural or diotic. SRTs increased when the target was reverberant rather than anechoic (temporal smearing), but this deleterious effect of reverberation was reduced when the target was binaural rather than diotic (see Fig. 6). This reduction illustrates binaural de-reverberation as it is defined in this paper: in the presence of a reverberant target, SRTs are lower under binaural listening conditions compared to diotic conditions.

B. Model parameters and performance criteria

As discussed in Sec. II B, three model parameters have to be defined for a given early/late separation: ELL, DD, and window shape. In the literature, this separation process has often used the equivalent of a rectangular window with ELL as the unique parameter and its value changed quite significantly across studies. An early/late limit of 50 ms ("Rect_{50}") has been used very commonly (Roman and Woodruff, 2013; Arweiler and Buchholz, 2011; Bradley *et al.*, 2003; Soulodre *et al.*, 1989), but other studies also used a limit of 35 ms (Bradley, 1986), 80 ms (Bradley, 1986), or 100 ms (Rennies *et al.*, 2011; Lochner and Burger, 1964). Because of the wide range of ELLs reported in the literature, the present study carried out a systematic test on the three model parameters to determine their role in reverberant speech recognition. Twenty-one ELL values were tested (from 0 to 100 ms each 5 ms), along with 21 DD values (from 0 to 100 ms each 5 ms) and two window shapes (linear and sigmoid). The rectangular window predictions were obtained from those of the linear window with DD = 0 ms.

Model predictions and experimental data were compared for each model setup. Prediction performance was assessed using the correlation coefficient (r), the mean absolute error ($\bar{\varepsilon}$), and the largest error ($\varepsilon_{\max}$) across conditions between data and predictions for each of the three experiments mentioned in the preceding text.

C. Results

Figure 2 presents the mean absolute prediction error across conditions as a function of ELL for the rectangular window. For the three experiments, the prediction error is first reduced with increasing ELL, it reaches a minimum and then increases for longer ELLs (even if not plotted here, the error increased for ELLs above 100 ms for the data of RBK). For RBK and LC3, involving temporal smearing and spatial unmasking, the prediction error is small over a broad range of ELLs. For an ELL between 40 and 200 ms for RBK and between 25 and 95 ms for LC3, the mean error is less than 1 dB. For LC4 involving binaural de-reverberation, the same mean error is reached for ELLs between 20 and 60 ms. Because the de-reverberation effect is only about 1 dB, the range of ELLs leading to good predictions of binaural de-reverberation is much narrower (30–40 ms) for LC4.

Figure 3 presents contour plots for RBK, LC3 and LC4 showing the prediction error as a function of ELL and DD with a linear window. In addition to the contour lines, a

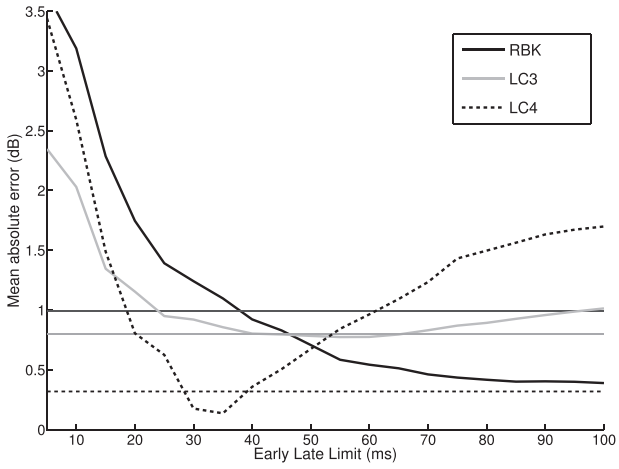

FIG. 2. Mean absolute error between measurements and model predictions for each experiment as a function of ELL for the rectangular window. The mean absolute errors of the room-independent (RI) model are plotted as horizontal lines.

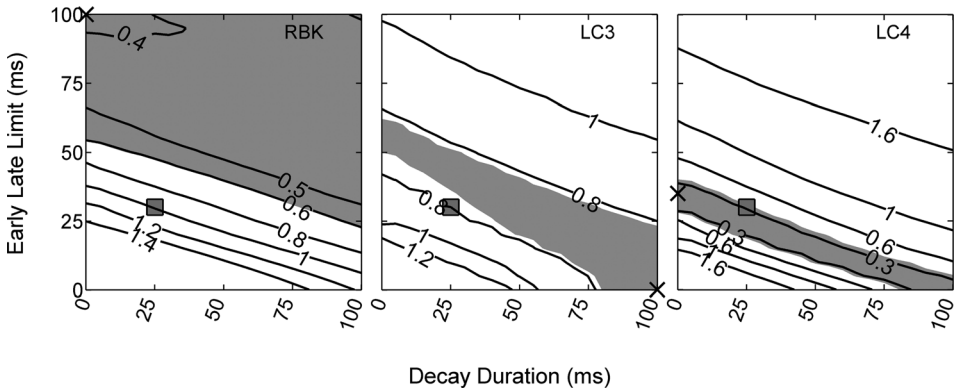

FIG. 3. Contour plots of the mean absolute error between measurements and model predictions as a function of ELL and DD for each experiment. The gray area represents the minimal error zone. The black cross indicates the smallest prediction error among all predictions. The gray square represents the error of the RI model (ELL = 30 ms and DD = 25 ms).

"best" point (black cross) and a minimum area (gray zone) are plotted. The best point represents the pair of parameters which leads to the smallest mean absolute error[2] ($\varepsilon_{\min}$).

The minimum area is the zone between the levels $\varepsilon_{\min}$ and $\varepsilon_{\min} + 0.05 \times (\varepsilon_{\max} - \varepsilon_{\min})$. In this area, the prediction error is close to its minimum, within 5% of the spread of prediction errors. For each experiment, the influence of ELL on the prediction error follows the same pattern as in Fig. 2 for the rectangular window. The differences across experiments mainly concern the gradient (along ELL and DD) of the mean error and consequently, the size of the area where the mean absolute error was minimized.

The results obtained with the sigmoid window were very similar to those obtained with the linear window. On average across experiments, the correlation coefficient between the mean absolute error obtained with the linear and sigmoid windows was 0.99. On average across ELL and DD values, the differences of mean absolute errors were 0.05 dB (RBK), 0.01 dB (LC3), and 0.07 dB (LC4). The present study thus focused on the linear window (which is simpler to implement).

The three minimum areas obtained with RBK, LC3, and LC4 did not clearly overlap, and the best performances were obtained for very different values of ELL and DD across experiments. These three sets of data did not lead to a unique and optimal value of the window parameters, suggesting that the best performance of the model could be room-dependent (RD): the window parameters of the proposed model have to be adjusted differently in each experiment to yield the best performance. To propose a room-independent (RI) model with a fixed window, a pair of parameters was chosen with a will to keep the binaural de-reverberation well predicted because it presents the smallest minimum area (the two other experiments should be more robust to the compromise). This pair of RI parameters is presented as a gray square on each contour plot (ELL = 30 ms, DD = 25 ms).

Figures 4–6 compare the measured SRTs to the RD and RI model predictions for RBK, LC3 and LC4, respectively. The predictions of the old model (Lavandier *et al.*, 2012; Jelfs *et al.*, 2011), without splitting the target BRIR, are also plotted. The predictions obtained with the RD model accurately fit experimental data, especially for RBK and LC4. A recurrent discrepancy occurred for the anechoic target in LC3. The RI model is less accurate than the RD model even though it does predict the trends associated with temporal smearing, spatial unmasking and binaural de-reverberation. The old model led to very poor performances by considering the entire reverberant target speech as useful.

The performances of three model configurations are compared in Table I: RD, RI (ELL = 30 ms, DD = 25 ms) and Rect_{50} (rectangular window with ELL = 50 ms commonly used in the literature). The best performance is achieved by the RD model according to r, $\bar{\varepsilon}$, and $\varepsilon_{\max}$ in the three experiments. The RI and Rect_{50} models predict well the trends of temporal smearing and spatial unmasking in reverberation as indicated by the high correlations obtained, but with less accuracy than the RD model (larger errors). Prediction accuracy is improved when the early/late parameters are adjusted to each room and only the trends are predicted with fixed parameters.

D. Discussion

For each dataset, the model performance initially improved as soon as ELL or DD increased. This result confirms the usefulness of early reflections for speech intelligibility in rooms (Arweiler and Buchholz, 2011; Bradley *et al.*, 2003; Lochner and Burger, 1964). Performance decreased when ELL or DD became too long, highlighting

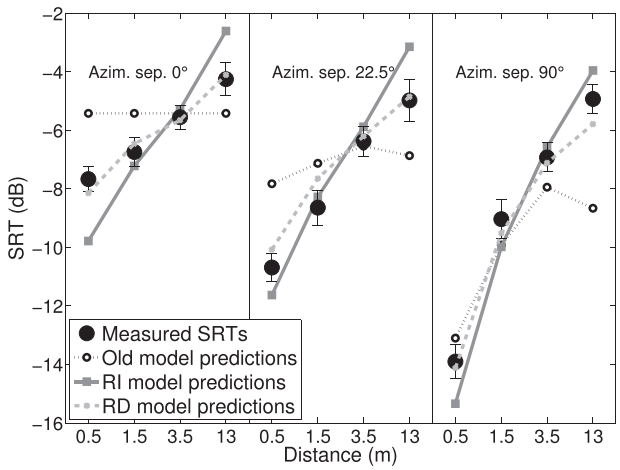

FIG. 4. Mean SRTs (black circles) with standard errors across listeners measured by Rennies *et al.* (2011, RBK) as a function of target-to-listener distance and azimuth separation (Azim. sep.). Predictions are plotted for the room-dependent model (crosses; ELL = 100 ms and DD = 0 ms), the room-independent model (squares; *ELL* = 30 ms, *DD* = 25 ms) and the old model (dotted line; without splitting the target BRIR into early and late parts).

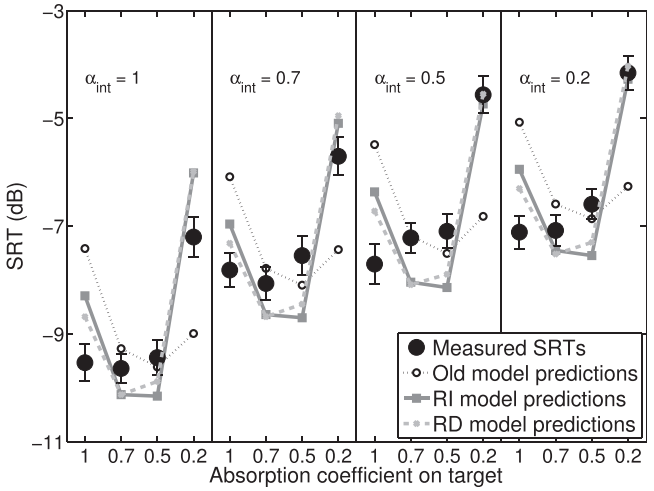

FIG. 5. Mean SRTs (black circles) with standard errors across listeners measured by Lavandier and Culling (2008, LC3) as a function of the absorption coefficient used for the target and interferer (α_{int}). Predictions are plotted for the room-dependent model (crosses; ELL = 0 ms and DD = 100 ms), room-independent model (squares; *ELL* = 30 ms, *DD* = 25 ms) and the old model (dotted line; without splitting the target BRIR into early and late parts).

the detrimental effect of the late reflections on speech intelligibility.

In RBK and LC3, reverberation disrupted intelligibility by reducing the spatial masking release and by temporally smearing the target speech. The RD model accurately predicted these two effects with a similar level of performance as previous models in the literature (Rennies *et al.*, 2014; Lavandier *et al.*, 2012; Jelfs *et al.*, 2011; Rennies *et al.*, 2011; Beutelmann and Brand, 2006): $r > 0.9$, $\bar{\varepsilon} < 1$ dB and $\varepsilon_{max} < 1.5$ dB. A noticeable discrepancy of about 1 dB recurrently occurred for the anechoic target condition in LC3. It only concerned one BRIR, which was tested against four

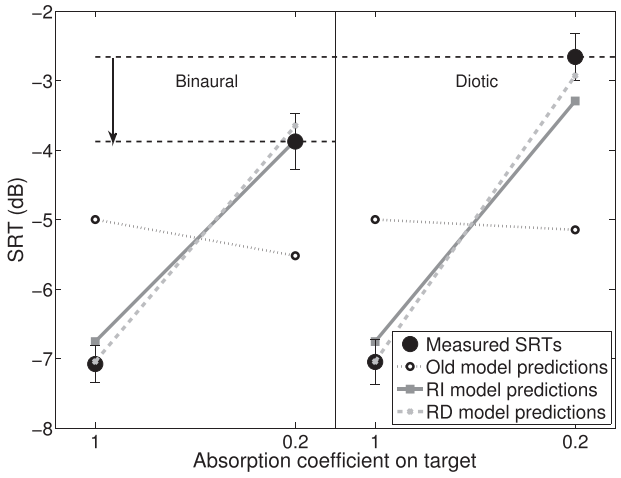

FIG. 6. Mean SRTs (black circles) with standard errors across listeners measured by Lavandier and Culling (2008, LC4) as a function of the absorption coefficient and listening mode (binaural/diotic) used for the target. Predictions are plotted for the room-dependent model (crosses; ELL = 35 ms and DD = 0 ms), room-independent model (squares; ELL = 30 ms, DD = 25 ms) and the old model (dotted line; without splitting the target BRIR into early and late parts). In the presence of a reverberant target, the benefit between binaural and diotic conditions illustrated by an arrow corresponds to the binaural dereverberation effect.

TABLE I. Prediction performance for each experiment for different model setups: room-dependent (different model parameters for each experiment, see Figs. 4–6), room-independent (ELL = 30 ms, DD = 25 ms, linear window), and "$Rect_{50}$" (ELL = 50 ms, rectangular window). Performance is assessed using the correlation coefficient (r), the mean absolute error ($\bar{\varepsilon}$ in dB) and the largest absolute error (ε_{max} in dB) between data and predictions.

	RD			RI			$Rect_{50}$		
Experiment	r	$\bar{\varepsilon}$	ε_{max}	r	$\bar{\varepsilon}$	ε_{max}	r	$\bar{\varepsilon}$	ε_{max}
RBK	0.98	0.4	1	0.97	1	2.1	0.98	0.7	1.7
LC3	0.90	0.7	1.2	0.86	0.8	1.3	0.83	0.8	1.5
LC4	0.99	0.1	0.3	0.99	0.3	0.6	0.99	0.7	1

different maskers. According to the model predictions (even in its old version), the SRT decrease between the anechoic target and the moderately reverberant ones would be due to coloration, which influenced the better-ear component of the model. Listeners did not seem to have taken any advantage of this coloration.

The monaural STI and U/D ratio cannot predict spatial unmasking or the reduction of spatial unmasking caused by reverberation. The binaural model of Lavandier and Culling (2010) can predict these two effects: it predicts the decrease of SRT with increasing azimuth separation of sources (at fixed distances) and also the reduction of this spatial unmasking advantage with increasing source distance in Fig. 4 (see also the prediction of the SRT increase with increasing reverberation for the interferer, at fixed reverberation levels for the target in Fig. 5). However, this old model does not predict the temporal smearing of speech, as represented by the predicted SRTs remaining constant with increasing target distance in the first panel of Fig. 4 (colocated source condition). Splitting the target BRIR into a useful and detrimental parts facilitated an extension of the model prediction ability to reverberant targets, while keeping accurate predictions for spatial unmasking.

Rennies *et al.* (2011) modeled their data by extending their binaural speech intelligibility model (BSIM) in three different ways: MTF, D_{te}, and U/D. In the models using MTF or D_{te}, the binaural model is applied to the entire speech signal including the late reverberant part, and the binaurally improved SNRs are corrected afterwards to take into account the temporal smearing of the target speech. As in the model proposed here, the U/D extension computes the early and late parts of the target before applying the binaural model to the useful (early target) and detrimental (late target + interferer) components. They observed similar levels of performance with the U/D and D_{te} models, whereas the MTF approach induced a larger bias. Three ELL values (50, 80, and 100 ms) were tested with a rectangular window for the U/D and D_{te} extensions. The ELL of 100 ms gave the best predictions for both models: $r = 0.98$, $\rho_S = 0.95$ (Spearman's rank correlation), and $RMSE = 1.4$ dB (root mean square error) for U/D and $r = 0.98$, $\rho_S = 0.97$ and $RMSE = 1.1$ dB for D_{te}. The model proposed here yielded its best predictions ($r = 0.98$, $\rho_S = 0.97$, and $RMSE = 0.48$ dB) on the same data with the same window (rectangular with ELL of 100 ms). Rennies *et al.* (2014) tested the three

approaches proposed by Rennies *et al.* (2011) on the data of Warzybok *et al.* (2013) in which reverberation was limited to a single reflection. In addition, they tested two temporal window shapes for the U/D and D_{te} versions: a rectangular window with ELL = 100 ms and a window equivalent to our linear window with ELL = 0 ms and DD = 200 ms. They observed the best performance with the U/D approach and a linear window, reaching a similar level of performance as the model proposed here: $r = 0.97$ and an $RMSE = 0.9$ dB across three noise conditions (diffuse, located at 0° or at 135°). They also tested six ELLs, four DDs, and four window shapes in the case of a frontal reflection with a colocated or separated noise source.

The present study focused on the U/D approach and investigated the influence of each model parameter, extending the tests conducted by Rennies *et al.* (2014): all combinations of window parameters have been tested, and this was done in three different rooms with realistic reverberation. The conclusions of the present study were consistent with those of Rennies *et al.* (2014) concerning the shape of the window, indicating a minor influence of using either a linear or a sigmoid window. A clearer understanding of the influence of ELL and DD is also provided by Fig. 3, which revealed that ELL and DD can be adjusted to reach a given level of performance. Predictions obtained with a rectangular window (DD = 0 ms) can also be of the same accuracy as those obtained with a linear window (DD > 0 ms) as long as a different ELL is used. Thus the parameter values required to reach a given prediction error are not unique, several window configurations can provide the same performance.

Previous studies (Roman and Woodruff, 2013; Arweiler and Buchholz, 2011; Bradley *et al.*, 2003; Soulodre *et al.*, 1989) often used a $Rect_{50}$ window to separate early from late reflections in an impulse response. Early-to-late energy ratios (or clarity) are usually computed using a 50 ms limit for speech and an 80 ms limit for music (ISO, 1997). Warzybok *et al.* (2013) highlighted the limitation of a rectangular window in presence of a single reflection. In this extreme case, such a window is clearly not suitable because the reflection is considered either as fully useful or fully detrimental. Conversely, in the presence of more realistic reflection patterns, the present study showed that the $Rect_{50}$ window yielded similar correlations to the RI or RD models but with larger errors (Table I). The present work does not question previous uses of this window, but it is pointed out here that the prediction is limited to an approximation of the temporal smearing effect. The $Rect_{50}$ window does not appear suitable to predict binaural de-reverberation (LC4). An ELL of 35 ms (previously used by Bradley, 1986) rather than 50 ms led to a better performance for predicting this effect.

The systematic tests of the model parameters on RBK, LC3 and LC4 highlighted that the parameters giving the best prediction are room-dependent. This dependence could partially explain the wide range of ELL reported in the literature. Fixing the window parameters across experiments did not lead to satisfactory predictions: the RI model defined by these three experiments could predict the trends of temporal smearing and binaural de-reverberation but less accurately than the RD model. This would suggest that the U/D approach might not be sufficient to describe speech perception in rooms.

The validity of the RI model and its ability to describe the trends of speech transmission independently from the room was further tested on a fourth dataset, which was not used to define its parameters. It involved temporal smearing and spatial unmasking in different rooms.

IV. ROOM-INDEPENDENT MODEL VALIDITY

A. Experimental data

Van Wijngaarden and Drullman (2008) measured consonant-vowel-consonant (CVC) scores (which use simple nonsense words embedded in carrier sentences) instead of SRTs to measure speech intelligibility in 39 conditions. Among these 39 conditions, only 24 were modeled here ([1–5; 8–12; 15–18; 22–24; 27–31; 35; 38]), excluding the conditions in quiet (they present an infinite SNR and CVC scores conversion into SRTs is possible only with finite SNRs; see Sec. IV B) and the conditions in which noise was not convolved by a BRIR (because the proposed model requires BRIRs as inputs). Intelligibility scores were measured at different SNRs (−6, −3, 0, 3, and 6 dB) using headphones by simulating a target masked by a discrete speech-shaped noise in four listening environments: anechoic room, listening room, classroom and cathedral [see Table I of van Wijngaarden and Drullman (2008) for a detailed description of the conditions].

B. Scores transformation

To be compared to the model predictions, the experimental CVC scores were first transformed into SRTs according to the psychometric function proposed by (Brand and Kollmeier, 2002, Eq. 1) which has the SRT and its slope at SRT as parameters. The slope can be deduced from the conditions that only differ in SNR. Such conditions should share the same psychometric function and SRT. Eight pairs of such conditions were identified (1/8, 2/9, 3/10, 4/11, 5/12, 15/17, 16/18, and 35/38). For each pair, the two SRTs obtained by transforming the CVC score with the psychometric function should be equal. It was not the case in practice because experimental errors occurred during the measurement. A unique slope value (9.68%/dB) was then determined with a least-square method such that it minimized this experimental error across the eight pairs. The score-to-SRT transformation was then applied to all modeled conditions using the same slope value.

Sixteen transformed SRTs (averages of each eight pairs and eight singles) were compared to the predictions obtained with the RI model (linear window, ELL = 30 ms, DD = 25 ms).

C. Results

Figure 7 presents the transformed SRTs and the predictions from both the RI model and the old model (without splitting the target BRIR) for the 16 conditions considered. The different panels refer to the tested rooms (anechoic

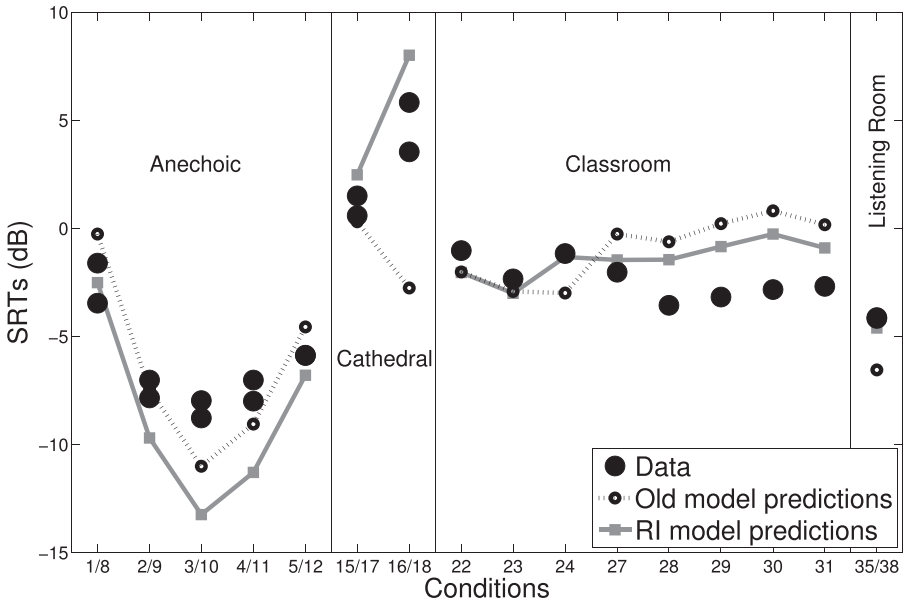

FIG. 7. Transformed SRTs (black circles) from CVC scores measured by van Wijngaarden and Drullman (2008) in four rooms: anechoic room, cathedral, classroom, and listening room. Predictions are plotted for the room-independent model (squares; ELL = 30 ms, DD = 25 ms) and for the old model (dotted line; without splitting the target BRIR into early and late parts). The condition numbers are labeled as they appear in the Table I from van Wijngaarden and Drullman (2008) except for the re-assigned conditions (see footnote 3).

room, cathedral, classroom, and listening room). The abscissa refers to the condition index taken from Table I of van Wijngaarden and Drullman (2008). According to this table, spatial unmasking occurred in the anechoic conditions[3] (among conditions 1–5) as well as in the classroom (among conditions 27–31). Temporal smearing of speech occurred in the cathedral conditions (between condition 15 and 16) as well as in the classroom (between condition 23 and 24). No binaural de-reverberation was highlighted in any condition.

For each room, the RI model defined in Sec. III only described the trends of the transformed SRTs with a limited accuracy. By first averaging the transformed SRTs of each of the eight pairs, the correlation coefficient between experimental data and model predictions was $r = 0.96$, the mean absolute error over the 16 conditions was $\bar{\varepsilon} = 1.77$ dB, and the largest error was $\varepsilon_{\max} = 4.87$ dB. The old model predicted less accurately this experimental dataset ($r = 0.65$, $\bar{\varepsilon} = 2.27$ dB, and $\varepsilon_{\max} = 7.43$ dB). The prediction errors were even larger than with the RI model in some conditions. In particular, the old model did not predict the deleterious effect of temporal smearing (conditions 15/16 and 23/24).

D. Discussion

The performance of the RI model for the experimental data from van Wijngaarden and Drullman (2008) was less accurate than the modeling of the other three experiments even though the trends of the different effects are described (resulting in a good correlation). Four rooms were tested in this experiment, which is the reason why it appeared suitable to test the RI model. Even if this model described the main trends in the data, it failed to accurately predict intelligibility in all conditions. This might indicate an inherent limitation of this model. The observed discrepancies across rooms confirm the room dependence of the window parameters. Some sources of variability in the experimental and modeling processes might have also affected the model performance. First, only seven listeners participated in the experiment, which contained 39 conditions, and the variability in the experimental data was not presented in the results. The transformation of the CVC scores into SRTs implied a fitting of the psychometric function slope (s_{50}), assuming it only depends on speech material. This fitting process prevents any direct comparison between data and predictions as performed with the three other experiments.

The predictions obtained with the old model did not fit to the experimental data. The largest errors occurred in presence of temporal smearing, while predictions were similar to the RI model for high D/R ratios. For instance, very accurate predictions were reached in the anechoic conditions (the offset between the two models being only due to the fact the predictions are compared to the data by fitting the averaged SRT across all 16 conditions, this average being different for the two models). The entire target BRIR is considered as useful and the detrimental part only consists of the noise BRIR, so that the two models are identical.

Van Wijngaarden and Drullman (2008) modeled their data by applying a binaural STI model using interaural correlograms from modulation transfer functions on the left and right ears. Since they compared their model to the STI reference curve instead of measuring its goodness of fit to the data, a direct comparison of performance is not possible.

V. GENERAL DISCUSSION

A. Limitations of the U/D approach

In the four experimental datasets used in the present study, the predictions of the RI model were always limited to the trends of the effects. Adjustments on the early/late separation parameters were needed to yield accurate predictions. Unlike previous studies (Rennies *et al.*, 2014; Rennies *et al.*, 2011), the U/D approach was tested here in different rooms. It was thus able to highlight this room dependence, which might constitute a fundamental limitation to the U/D approach to predict speech intelligibility in rooms. The current version of the model cannot be used to make *a priori* predictions in different rooms. The early-late separation might depend on other parameters that are not taken into

account in the current version of the model proposed here. To obtain both prediction accuracy and room independence for the model, the early/late separation could be determined by modeling other perceptual mechanisms. For instance, previous studies showed that listeners are able to adapt to room acoustics thanks to prior exposure (Brandewie and Zahorik, 2010; Watkins, 2005). The proposed RI model would be improved by including this adaptation ability, which might be related to room acoustics parameters: do listeners adapt to the particular BRIR or to the room as a whole? The separation between early/useful and late/detrimental parts of speech might also depend on the speech rate. The direct sound of a pronounced word can overlap with the reflection of the previous word depending on how fast the words are spoken, illustrating how a reflection can be regarded as useful or detrimental depending on the speech rate. To account for this effect, the early-late separation could be made dependent on the frequency modulation in each frequency band, but this implementation would not be easy in the present model framework.

Room dependence appears to be a relevant aspect of speech intelligibility modeling. This suggests that other approaches (MTF, D_{te}) should be considered as potential candidates to account for temporal smearing and tested across different rooms. Even if Rennies *et al.* (2011) implemented and compared the performance of these approaches, their room-independence should be investigated.

B. Unified interpretation of spatial unmasking, temporal smearing, and binaural de-reverberation

By adding the late target to the interferer to constitute the detrimental input of the binaural process, the proposed model provides an interpretation of temporal smearing in terms of self-masking of the target induced by late reflections in the room. The late target is an additional masker, treated like any other interfering source by the model. Its effect appears at high levels of reverberation (Lavandier and Culling, 2008) because the late target needs to be sufficiently energetic to become a non-negligible new source of interference.

The RD model predicted correctly the effect of binaural de-reverberation in a narrow range of ELLs. According to the model, this ability to benefit from binaural listening in reverberant environments can be understood simply in terms of binaural unmasking of the early target against the late target. This interpretation is compatible with both the EC theory (Durlach, 1972) and the U/D ratio concept (Lochner and Burger, 1964). In diotic listening, early and late targets do not have any interaural phase differences, so cancellation is impossible and there is no binaural unmasking. For binaural targets, reverberation spreads part of the late energy to different interaural phases from that of the early target, so that the EC mechanism can eliminate a part of this late target (its coherence determining the level of cancellation). It should be noted that early and late targets might have different ILDs so that better-ear listening could also contribute to de-reverberation, which would then involve the two components of spatial unmasking.

The interpretation of de-reverberation in terms of binaural unmasking is also consistent with the signal-processing technique proposed by Allen *et al.* (1977) to remove reverberation from speech signals. It consists in decomposing in frequency bands the signals from two microphones placed in the room and weighting the different frequency bands according to the cross-correlation of the two signals in each band, before synthesizing the composite de-reverberated signal. Based on the hypothesis that the early signal is more correlated than the late signal at the two microphones, the weighting process aims at re-synthesizing only the coherent early part of the signal. The binaural system processes a similar cancellation of the late signal, but this cancellation is based on differences of interaural phase difference between early and late targets rather than on coherence. The low coherence of the late reverberated target is a limitation for the binaural system, which prevents the EC mechanism from cancelling the late target perfectly. This limitation could explain why Allen's signal-processing technique was found to perform better than the binaural system.

Libbey and Rogers (2004) interpreted binaural de-reverberation as binaural overlap-masking release with reverberation acting as masking noise. They compared the ability to unmask reverberation and reverberation-like noise. The benefit of binaural listening was reduced with reverberation-like noise compared to reverberation. This could be explained by the fact that reverberation-like noises were constructed by randomizing the reverberation phases leading to uncorrelated noise. In contrast, reverberation is not totally uncorrelated, and it is its correlated part that can be unmasked by the binaural system. Thus the difference of performance did not necessarily reveal that two mechanisms were involved but rather that a unique mechanism (spatial unmasking) behaved differently to different levels of correlation (as predicted by the proposed model).

Warzybok *et al.* (2013) investigated the influence of a single delayed reflection on frontal target speech masked by discrete noise. Their main findings are in good agreement with the conceptual interpretation of the proposed model. First, they observed no influence of the delay of a frontal speech reflection on spatial unmasking. Such a reflection cannot be unmasked because it has the same interaural phase as the target whatever the delay is, resulting in no BMLD. Second, the detrimental effect of long delays on a frontal reflection was reduced by separating the reflection from the target direction. Because a late reflection is regarded as a masker, unmasking is easier as soon as target and reflection are spatially separated. Third, in the presence of a discrete noise, the late reflection was less detrimental when it arrived from the same hemisphere as the noise than when it arrived from the opposite hemisphere. The binaural unmasking process in the present model is applied to the detrimental component (late speech + noise sources), which could be more coherent (so easier to cancel) when the masking sources come from the same spatial region.

Arweiler *et al.* (2013) investigated the integration of early reflections for improving speech intelligibility. Participants listened (monaurally or binaurally) to a frontal target (in anechoic or with early reflections) masked by a

speech-shaped noise (diffuse or located at 90° on the right). Because no advantage was observed between the monaural and binaural conditions, the authors concluded that the integration process of early reflections with the direct sound "appears to be monaural for both the directional and the diffuse masker," which, at first, does not seem in agreement with the concept of the binaural model proposed here. This model might, however, explain why no binaural effect was observed concerning the early/late integration process. First, late reflections were not involved, so that their binaural effect on interferer coherence could not be observed. Then early reflections influence target interaural phase difference, but when the difference in interaural phase difference between target and interferer is large (which was the case in this study), the interaural phase difference of each source has little effect if any on binaural unmasking (Lavandier and Culling, 2010). So the early/late integration was reduced to its monaural component in the particular conditions tested, and this study fits in the framework of the proposed binaural model.

VI. CONCLUSION

A model computing binaural U/D ratios was proposed to simultaneously account for temporal smearing, spatial unmasking, and binaural de-reverberation in reverberant environments. It combines a binaural model predicting spatial unmasking of a near-field target from multiple discrete noise interferers and a U/D decomposition taking into account the temporal smearing effect of reverberation on speech transmission. The early/late limit and decay duration used in the U/D separation both contribute to the model accuracy, but, it has been shown that these two parameters can be adjusted to reach a given prediction error, so that there is no unique way of defining early and late parts. The best model performance was achieved by adjusting the early/late separation for each experiment, leading to a room-dependent model. A room-independent model with fixed parameters was proposed, but it always predicted the trends of the temporal smearing with less accuracy than the room-dependent model. This result suggests that a fixed early/late separation might not be sufficient to predict speech intelligibility in rooms jeopardizing the generalization of the U/D approach to any room. However, the present modeling showed a unified interpretation of temporal smearing, spatial unmasking, and binaural de-reverberation in terms of masking of early target (useful) by late target (detrimental) combined with unmasking by the binaural system. Temporal smearing during speech transmission is just masking from a particular interferer: the late target. Binaural de-reverberation is simply spatial unmasking of this particular interferer (or spatial un-self-masking of the target).

ACKNOWLEDGMENTS

The authors are grateful to Jan Rennies and Rob Drullman for providing the BRIRs of their study. This work was performed within the LabEX CeLyA of Université de Lyon (ANR-10-LABX-0060)/(ANR-11-IDEX-0007).

[1]The sigmoid window was defined as $\Phi(t) = 0.5 \times (1 + \mathrm{erf}(t - \mu/\sigma\sqrt{2}))$ with σ and μ defined such that $\Phi(\mathrm{ELL}) = 0.999$ and $\Phi(\mathrm{ELL} + \mathrm{DD}) = 0.001$.

[2]Because RBK and LC3 presented best performance for border values, the systematic tests have been purchased further than 100 ms for ELL (for RBK) and DD (for LC3). The best performance for RBK was still reached at ELL = 100 ms, while it was reached again at DD = 145 ms for LC3 with the same mean error as with DD = 100 ms ($\bar{\varepsilon} = 0.6$).

[3]Based on previous data on spatial unmasking in anechoic conditions (Beutelmann and Brand, 2006; Plomp, 1976; Hawley *et al.*, 2004), we strongly suspect that labels have been switched among the anechoic conditions. This would explain some odd results: for instance, the target is more unmasked when the masker is located at 30° rather than 60° (conditions 3 and 4 or 10 and 11). We then decided to re-assign the labels of the conditions by conserving logical scores regarding spatial unmasking (except for 0°, the azimuth labels have just been shifted one rank upward such that the conditions 4/9, 3/10, 2/11 and 5/12 correspond to the azimuth 30°, 60°, 90° and 150°, respectively).

Allen, J. B., Berkley, D. A., and Blauert, J. (**1977**). "Multimicrophone signal-processing technique to remove room reverberation from speech signals," J. Acoust. Soc. Am. **62**, 912–915.

ANSI (**1997**). S3.5, *American National Standard Methods for Calculation of the Speech Intelligibility Index* (Acoustical Society of America, New York).

Arweiler, I., and Buchholz, J. M. (**2011**). "The influence of spectral characteristics of early reflections on speech intelligibility," J. Acoust. Soc. Am. **130**, 996–1005.

Arweiler, I., Buchholz, J. M., and Dau, T. (**2013**). "The influence of masker type on early reflection processing and speech intelligibility," J. Acoust. Soc. Am. **133**, 13–16.

Beutelmann, R., and 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., and Kollmeier, B. (**2010**). "Revision, extension, and evaluation of a binaural speech intelligibility model," J. Acoust. Soc. Am. **127**, 2479–2497.

Bradley, J. S. (**1986**). "Predictors of speech intelligibility in rooms," J. Acoust. Soc. Am. **80**, 837–845.

Bradley, J. S., Reich, R. D., and Norcross, S. G. (**1999**). "On the combined effects of signal-to-noise ratio and room acoustics on speech intelligibility," J. Acoust. Soc. Am. **106**, 1820–1828.

Bradley, J. S., Sato, H., and Picard, M. (**2003**). "On the importance of early reflections for speech in rooms," J. Acoust. Soc. Am. **113**, 3233–3244.

Brand, T., and Kollmeier, B. (**2002**). "Efficient adaptive procedures for thresholds and concurrent slope estimates for psychophysics and speech intelligibility tests," J. Acoust. Soc. Am. **111**, 2801–2810.

Brandewie, E., and Zahorik, P. (**2010**). "Prior listening in rooms improves speech intelligibility," J. Acoust. Soc. Am. **128**, 291–299.

Bronkhorst, A. W., and 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.

Collin, B., and Lavandier, M. (**2013**). "Binaural speech intelligibility in rooms with variations in spatial location of sources and modulation depth of noise interferers," J. Acoust. Soc. Am. **134**, 1146–1159.

Culling, J. F., Hodder, K. I., and Toh, C. Y. (**2003**). "Effects of reverberation on perceptual segregation of competing voices," J. Acoust. Soc. Am. **114**, 2871–2876.

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

Hawley, M. L., Litovsky, R. Y., and 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.

Houtgast, T., and 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.

ISO (**1997**). 3382:1997, *Acoustics—measurement of the reverberation time of rooms with reference to other acoustical parameters* (International Organization for Standardization, Geneva, Switzerland).

Jelfs, S., Culling, J. F., and Lavandier, M. (**2011**). "Revision and validation of a binaural model for speech intelligibility in noise," Hear. Res. **275**, 96–104.

Koenig, W. (**1950**). "Subjective effects in binaural hearing," J. Acoust. Soc. Am. **22**, 61–62.
Lavandier, M., and 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., and Culling, J. F. (**2010**). "Prediction of binaural speech intelligibility against noise in rooms," J. Acoust. Soc. Am. **127**, 387–399.
Lavandier, M., Jelfs, S., Culling, J. F., Watkins, A. J., Raimond, A. P., and Makin, S. J. (**2012**). "Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources," J. Acoust. Soc. Am. **131**, 218–231.
Levitt, H., and Rabiner, L. R. (**1967**). "Predicting binaural gain in intelligibility and release from masking for speech," J. Acoust. Soc. Am. **42**, 820–829.
Libbey, B., and Rogers, P. H. (**2004**). "The effect of overlap-masking on binaural reverberant word intelligibility," J. Acoust. Soc. Am. **116**, 3141–3151.
Lochner, J. P. A., and Burger, J. F. (**1964**). "The influence of reflections on auditorium acoustics," J. Sound. Vib. **1**, 426–454.
Moncur, J. P., and Dirks, D. (**1967**). "Binaural and monaural speech intelligibility in reverberation," J. Speech Hear. Res. **10**, 186–195.
Nábělek, A. K., and Robinson, P. K. (**1982**). "Monaural and binaural speech perception in reverberation for listeners of various ages," J. Acoust. Soc. Am. **71**, 1242–1248.
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.
Rennies, J., Brand, T., and Kollmeier, B. (**2011**). "Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet," J. Acoust. Soc. Am. **130**, 2999–3012.
Rennies, J., Warzybok, A., Brand, T., and Kollmeier, B. (**2014**). "Modeling the effects of a single reflection on binaural speech intelligibility," J. Acoust. Soc. Am. **135**, 1556–1567.
Roman, N., and Woodruff, J. (**2013**). "Speech intelligibility in reverberation with ideal binary masking: Effects of early reflections and signal-to-noise ratio threshold," J. Acoust. Soc. Am. **133**, 1707–1717.
Soulodre, G. A., Popplewell, N., and Bradley, J. S. (**1989**). "Combined effects of early reflections and background noise on speech intelligibility," J. Sound Vib. **135**, 123–133.
van Wijngaarden, S. J., and Drullman, R. (**2008**). "Binaural intelligibility prediction based on the speech transmission index," J. Acoust. Soc. Am. **123**, 4514–4523.
Wan, R., Durlach, N. I., and Colburn, H. S. (**2010**). "Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers," J. Acoust. Soc. Am. **128**, 3678–3690.
Warzybok, A., Rennies, J., Brand, T., Doclo, S., and Kollmeier, B. (**2013**). "Effects of spatial and temporal integration of a single early reflection on speech intelligibility," J. Acoust. Soc. Am. **133**, 269–282.
Watkins, A. J. (**2005**). "Perceptual compensation for effects of reverberation in speech identification," J. Acoust. Soc. Am. **118**, 249–262.
Zurek, P. M. (**1993**). "Binaural advantages and directional effects in speech intelligibility," in *Acoustical Factors Affecting Hearing Aid Performance*, edited by G. Studebaker and I. Hochberg (Allyn and Bacon, Needham Heights, MA), pp. 255–276.