

Speech intelligibility in reverberation with ideal binary masking: Effects of early reflections and signal-to-noise ratio threshold

Nicoleta Roman^{a)}

Department of Computer Science and Engineering, The Ohio State University at Lima, Lima, Ohio 45804

John Woodruff

Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210

(Received 18 April 2012; revised 10 December 2012; accepted 14 January 2013)

Ideal binary masking is a signal processing technique that separates a desired signal from a mixture by retaining only the time-frequency units where the signal-to-noise ratio (SNR) exceeds a predetermined threshold. In reverberant conditions there are multiple possible definitions of the ideal binary mask in that one may choose to treat the target early reflections as either desired signal or noise. The ideal binary mask may therefore be parameterized by the *reflection boundary*, a predetermined division point between early and late reflections. Another important parameter is the local SNR threshold used in labeling the time-frequency units as either target or background. Two experiments were designed to assess the impact of these two parameters on speech intelligibility with ideal binary masking for normal-hearing listeners in reverberant conditions. Experiment 1 shows that in order to achieve intelligibility improvements only the early reflections should be preserved by the binary mask. Moreover, it shows that the effective SNR should be accounted for when deciding the local threshold optimal range. Experiment 2 shows that with long reverberation times, intelligibility improvements are only obtained when the reflection boundary is 100 ms or less. Also, the experiment suggests that binary masking can be used for dereverberation.

© 2013 Acoustical Society of America. [http://dx.doi.org/10.1121/1.4789895]

PACS number(s): 43.72.Qr [CYE]

Pages: 1707–1717

I. INTRODUCTION

It is well known that the intelligibility of speech in natural environments is influenced by both background noise and room reverberation (for a review on the topic see [Assmann and Summerfield, 2004](#)). In the absence of background noise, reverberation alone can degrade speech perception. Several factors contribute to the influence of reverberation on intelligibility including: room reverberation time, source distance, and the characteristics of both early reflections and late reverberation (for a review see [Boothroyd, 2004](#)). [George et al. \(2010\)](#) report that the speech reception reverberation threshold—the reverberation time at which the listener understands only 50% of the presented sentences in the absence of noise—is on the order of 2 s. The combined effect of reverberation and background noise on speech intelligibility has been well documented for both normal-hearing and hearing-impaired listeners with both stationary and non-stationary interferences and under both monaural and binaural presentations (see, for example, [Duquesnoy and Plomp, 1980](#); [Bradley et al., 1999](#); [Lavandier and Culling, 2008](#); [George et al., 2008, 2010](#)).

Speech enhancement, speech separation and dereverberation are active areas of research designed to mitigate the influence of these environmental factors. Computational auditory scene analysis (CASA) is an approach to speech separation that is inspired by studies of human speech perception

(for a review see [Wang and Brown, 2006](#)). The ideal time-frequency binary mask (IBM) has been proposed as a performance upper bound for the separation capacity of CASA algorithms ([Wang, 2005](#)). Given a two-dimensional time-frequency (T-F) decomposition of an acoustic mixture composed of target speech and background noise, the mask uses *a priori* information and selects the T-F units where the signal-to-noise ratio (SNR) exceeds a predetermined local threshold, called the local criterion (LC), and cancels the other units ([Wang and Brown, 2006](#)). A series of studies have investigated the effects of IBM processing on speech intelligibility with anechoic speech in background noise and found that the IBM produces significant improvements for both normal-hearing and hearing-impaired listeners ([Brungart et al., 2006](#); [Li and Loizou, 2008](#); [Brungart et al., 2009](#); [Wang et al., 2009](#); [Kjems et al., 2009](#)). These studies have both validated the IBM as a computational goal of separation algorithms, and alongside other recent work ([Loizou and Kim, 2011](#)), have helped to illustrate why measures of algorithm performance based on signal-to-noise ratio (SNR) ([Quackenbush et al., 1988](#); [Greenberg et al., 1993](#); [ANSI, 1997](#)) do not correlate well with the intelligibility of processed speech ([Hu and Loizou, 2008](#); [Taal et al., 2011b](#)). Specifically, consistent with recent objective intelligibility measures designed for nonlinearly processed speech (e.g., [Goldsworthy and Greenberg, 2004](#); [Taal et al., 2011a](#)), it has been shown that an LC below 0 dB can lead to improved intelligibility by better restoring the spectro-temporal modulation pattern of target speech ([Brungart et al., 2006](#); [Li and Loizou, 2008](#); [Wang et al., 2009](#); [Kjems et al., 2009](#)); in

^{a)}Author to whom correspondence should be addressed. Electronic mail: roman.45@osu.edu

spite of the fact that such masks are suboptimal in terms of SNR (Li and Wang, 2009). Further, analysis conducted by systematically introducing mask errors has shown that false alarm errors (wrongly labeling a T-F unit as 1) are more detrimental to intelligibility than miss errors (wrongly labeling a T-F unit as 0) (Li and Loizou, 2008). Such findings provide important insight for algorithm design and evaluation.

A recent investigation considered how best to extend the IBM definition to conditions with both noise and reverberation (Roman and Woodruff, 2011). The detrimental effect of reverberation on speech perception is largely the result of late arriving reflections, which essentially act as a masking noise, while it has been shown that early reflections are integrated by the auditory system to increase the loudness of direct sound (Wallach *et al.*, 1949; Lochner and Burger, 1964; Bradley *et al.*, 2003). Motivated by this, Roman and Woodruff (2011) compared the intelligibility of masked speech for normal-hearing listeners using three alternative IBM definitions: the mask obtained when the desired signal corresponds to the direct target sound (IBM-DS), the mask obtained when the desired signal is the direct target sound plus the early reflections (IBM-ER), and the mask obtained when the desired signal is the reverberant target (IBM-R). This study found that while all three masks lowered speech reception thresholds (the SNR at which sentence recognition is 50%) relative to the unprocessed mixtures for a moderate reverberation time (T_{60} equal to 0.4 s), only IBM-ER significantly improved intelligibility with a longer reverberation time (T_{60} equal to 0.8 s). This result was consistent for both stationary and non-stationary additive noise. Recent experiments show that the IBM-DS mask, referred to as the ideal reverberation mask, can lead to substantial intelligibility gains for cochlear implant listeners in reverberant conditions (Kokkinakis *et al.*, 2011; Hazrati and Loizou, 2012).

The main scope of the current study is to provide a more detailed analysis of IBM processing for reverberant signals. Experiment 1 extends the analysis in Roman and Woodruff (2011) for the condition with speech corrupted by reverberation and stationary noise. In our previous study, speech reception thresholds (SRTs) were measured for the three IBM definitions using a fixed LC of -6 dB. For a given input SNR, previous work has shown that the IBM improves intelligibility over a range of LC values (Brungart *et al.*, 2006, 2009; Kjems *et al.*, 2009). When LC is too low, too many T-F units are retained to improve over unprocessed mixtures (under-attenuation of noise), whereas when LC is too high, masks become too sparse (over-attenuation of speech). Assuming a linear filterbank, a shift in the input SNR creates a corresponding shift in the plateau region of optimal LC values (Kjems *et al.*, 2009). This can be understood by recognizing that co-varying input SNR and LC produces precisely the same IBM. For reverberant signals, the characterization of target reflections as either part of the desired signal or as noise significantly affects the effective input SNR of a mixture. While the -6 dB LC used in Roman and Woodruff (2011) is supported by the studies of Brungart *et al.* (2006, 2009) and Wang *et al.* (2009), it is possible that the relationship between LC and the effective SNR favored the IBM-ER definition. In other words, it is possible that the poor perform-

ance of the IBM-DS mask was due to overattenuation of target speech energy, whereas poor performance of the IBM-R may have been the result of under-attenuation of noise energy. In this experiment we compare the same IBM definitions over a range of SNR thresholds. To ensure that we consider an appropriate range of thresholds for each mask definition, we employ the concept of relative criterion (RC) (Kjems *et al.*, 2009), which is motivated by the shift in optimal LC range discussed above. We let RC equal to LC minus the effective SNR of the mixture. For a fixed RC, changes to the input SNR (i.e., due to additive noise) have no effect on the IBM. For changes in effective SNR due to different reflection boundary settings, a fixed RC does not ensure precisely the same IBM, but ensures that the mask density (balance between 1 and 0s) is similar across IBM definitions. The use of RC allows us to study changes in mask patterns due to different definitions of the desired signal and eliminates the confounding influence of effective SNR.

Whereas experiment 1 and Roman and Woodruff (2011) consider a single division point between early and late target reflections, or *reflection boundary*, in experiment 2 we consider a range of possible values to better identify the point in time at which target reflections should be characterized as noise. While the 50 ms reflection boundary used in experiment 1 is motivated by prior work (Bradley *et al.*, 2003), we consider reflection boundaries up to 200 ms in this experiment. The 200 ms boundary, which has been shown to be relevant to the perception of musical sounds (Hidaka *et al.*, 1995), serves as a reasonable upper bound on the time at which reflections are no longer beneficial for speech perception. Further, as the range of effective reflection boundaries may be affected by reverberation time, we consider four acoustical settings with increasing reverberation levels. To both highlight performance differences due to reflection boundary and to study the use of binary masking as a dereverberation technique, no additive noise is included in the case of long reverberation times.

In what follows, Sec. II provides a formal definition of the IBM parameterized by both LC and reflection boundary. Section III describes the first experiment designed to directly extend the analysis performed in Roman and Woodruff (2011). Section IV describes a second experiment to compare alternative reflection boundary settings and to assess the effectiveness of ideal binary masking for speech dereverberation. Finally, Sec. V provides further discussion and concludes the paper.

II. IDEAL BINARY MASKING

The binary T-F masking approach to sound separation is motivated in part by the masking phenomena, according to which a louder sound renders a weaker one inaudible within the same critical band (Moore, 2003). Given a T-F signal representation, the IBM is generated by calculating the SNR of a (predetermined) desired signal relative to a residual (noise) signal in individual T-F units, and setting the mask values to either 1 or 0 depending on whether or not the SNR value for a T-F unit exceeds a predefined threshold, the LC. The IBM has been extensively utilized in the CASA field in

the case of additive distortions where the desired signal corresponds to anechoic target speech and the residual signal corresponds to background noise (for a review, see Wang and Brown, 2006). Ideal binary T-F masks have proven to provide highly effective front ends for robust speech recognition (Cooke *et al.*, 2001; Roman *et al.*, 2003; Srinivasan *et al.*, 2006; Srinivasan and Wang, 2010) as well as to improve speech intelligibility for both normal-hearing and hearing-impaired listeners (Brungart *et al.*, 2006; Li and Loizou, 2008; Brungart *et al.*, 2009; Wang *et al.*, 2009; Kjems *et al.*, 2009). While in practice it is necessary to estimate the IBM from the mixture, the binary masking approach allows the use of supervised learning algorithms to perform classification, which are proving to be powerful alternatives to more traditional signal enhancement methods (Roman *et al.*, 2003; Kim *et al.*, 2009; Han and Wang, 2011).

In reverberant environments, CASA systems have generally treated the reverberant target speech as the desired signal to define the IBM, with the perspective that the goal of separation is to remove additive interference rather than perform dereverberation of the target signal (see, for example, Roman *et al.*, 2006; Pedersen *et al.*, 2008; Woodruff and Wang, 2010). Alternatively, dereverberation and speech separation systems often treat the direct sound (anechoic) target speech as the desired signal under the assumption that both noise and reverberation degrade signal quality (e.g., Miyoshi and Kaneda, 1988; Buchner *et al.*, 2004; Mandel *et al.*, 2010). In the former case, all reverberation due to the target speech source is treated as desirable, while in the latter, all reverberation is treated as background noise. Given the beneficial effect of early reflections on speech perception, Roman and Woodruff (2011) investigated treating the direct sound plus early reflections of the target speech as the desired signal in the definition of the IBM. Note that this characterization of desirable and undesirable signal components is consistent with the so-called useful-to-detrimental ratio (Bradley *et al.*, 1999). As discussed above, some of the performance differences between IBM definitions observed in Roman and Woodruff (2011) may have been due to the relationship between the -6 dB LC used and the effective SNR for each definition. To account for the systematic shift in effective SNR as the reflection boundary is changed, in this study we measure speech intelligibility as a function of RC, or the difference between the LC and the effective SNR. RC was defined as the difference between the LC and the mixture SNR in Kjems *et al.* (2009) in the context of anechoic mixtures. Again, use of RC accounts for the systematic shift in optimal LC range due to a shift in the input SNR, such that performance curves across different mixture SNR levels align well when results are displayed as a function of RC (Kjems *et al.*, 2009). As an extension to the anechoic case, we expect to see that different reflection boundaries, and thus different effective SNRs, produce similar performance patterns across a range of RC values. The conditions for which this extension is valid in reverberation will be studied extensively in the next sections.

In order to make the discussion more precise, we now introduce notations to denote the concepts described above and specify the processing details used to compute IBMs. We first define the mixture signal as

$$y(t) = h(t) * s(t) + n(t), \quad (1)$$

where $s(t)$ denotes the (anechoic) target speech signal, $h(t)$ denotes the room impulse response between the source location and microphone, and $n(t)$ denotes a background noise signal. We define the *desired* signal as

$$d_b(t) = h_b(t) * s(t), \quad (2)$$

where $h_b(t)$ denotes the part of $h(t)$ up to reflection boundary b . The *residual* signal is then

$$r_b(t) = y(t) - d_b(t). \quad (3)$$

The effective SNR for a given reflection boundary is calculated as

$$\text{SNR}_b = 10 \log_{10} \left(\frac{\sum_t d_b(t)^2}{\sum_t r_b(t)^2} \right). \quad (4)$$

IBMs are computed from the *cochleagram* representation, which is commonly used in the CASA field (Wang and Brown, 2006). Specifically, we analyze signals using a 64-channel gammatone filterbank with center frequencies from 50 to 8000 Hz equally spaced on the equivalent rectangular bandwidth scale. The response of each filter is then divided using 20 ms rectangular frames with 10 ms overlap into a set of T-F units (see also Wang and Brown, 2006). The cochleagram then corresponds to a two-dimensional response energy computed across all the T-F units. We let $D_b(\tau, f)$ and $R_b(\tau, f)$ denote the energy (in dB) of the desired and residual signals, respectively, in time frame τ and frequency channel f , calculated with reflection boundary b . The IBM is then

$$\text{IBM}_b(\tau, f) = \begin{cases} 1, & \text{if } D_b(\tau, f) - R_b(\tau, f) > \text{LC}_b, \\ 0, & \text{otherwise,} \end{cases} \quad (5)$$

where $\text{LC}_b = \text{RC} + \text{SNR}_b$ is the local SNR threshold expressed in dB. Finally, masks are applied to the mixture signals in a synthesis step in order to generate the waveform stimuli (Wang and Brown, 2006).

With this definition, the IBM is a function of both RC and reflection boundary parameters. At the limits, IBM_0 is the direct sound based mask while IBM_∞ represents the reverberant target based mask. Note that if no additional background noise is present, the residual signal contains only the late reflections whereas for noisy conditions the residual signal is the sum of late reflections and background noise.

Figure 1 illustrates ideal binary mask outputs for a mixture of speech and speech shaped noise. Figure 1(a) shows the cochleagram of the mixture. The acoustical condition corresponds to 0.8 s reverberation time and 3 dB input SNR (as described in Sec. IV). Three different masks, IBM_5 , IBM_{50} , and IBM_∞ , are shown for an RC value of -9 dB in Figs. 1(e)–1(g), respectively. The corresponding resynthesized signals are then displayed in Figs. 1(h)–1(j). While IBM_5 and IBM_{50} are clearly similar to each other and preserve much of the speech structure, a noticeable degradation is observed for the reverberant mask.

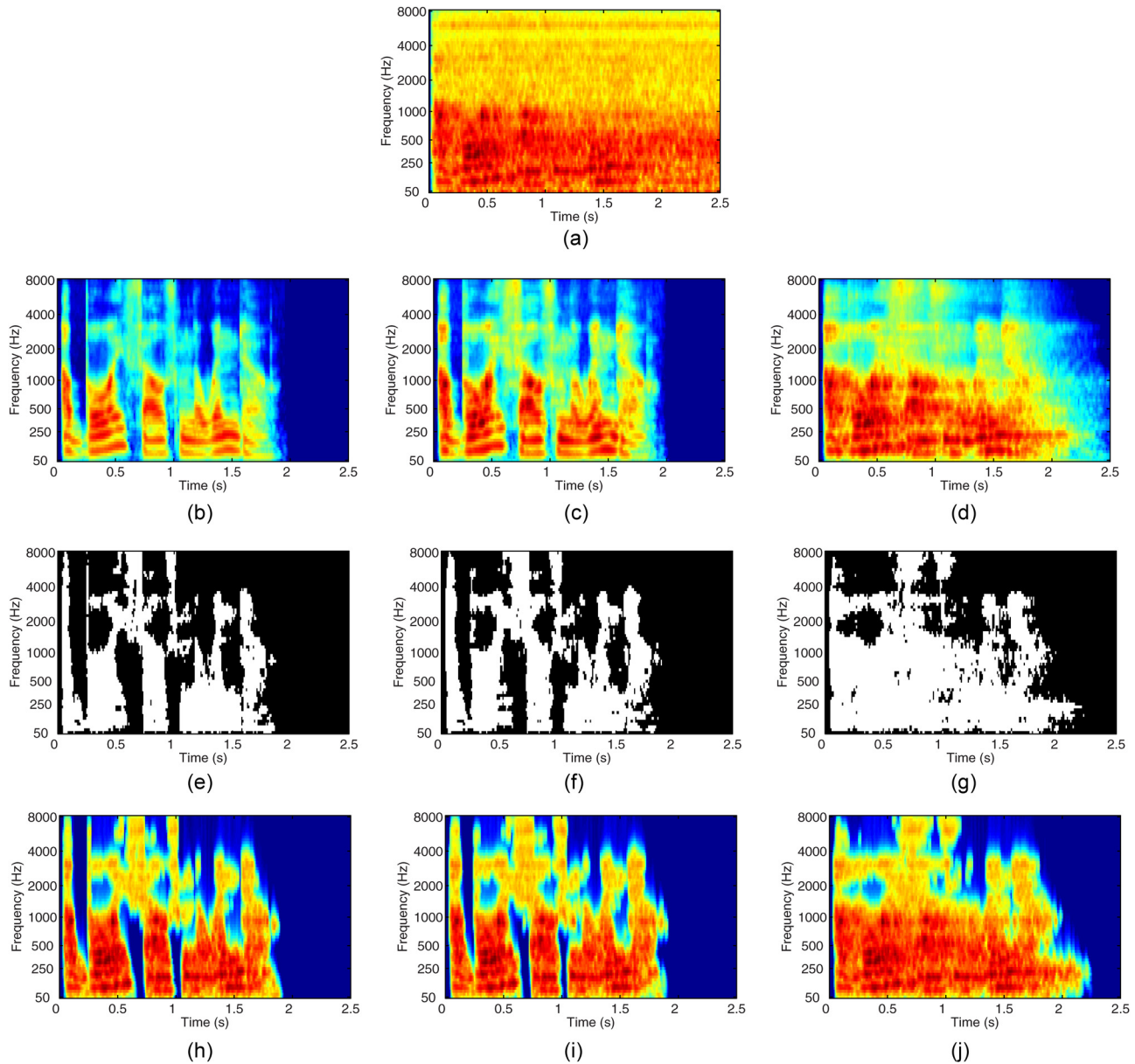


FIG. 1. (Color online) Illustration of ideal binary masks with different reflection boundaries. The acoustical condition corresponds to 0.8 s reverberation time and 3 dB input SNR (as described in Sec. IV). The binary masks are obtained for $RC = -9$ dB. Panels (a), (b), (c), and (d) show cochleagrams of reverberant mixture, the direct sound of target speech, the direct sound plus the early reflections of target speech, and the reverberant target speech, respectively. Panels (e), (f), and (g) show IBM_5 , IBM_{50} , and IBM_{∞} , respectively. Panels (h), (i), and (j) show cochleagrams of the resynthesized signals corresponding to the masks shown in (e), (f), and (g), respectively.

III. EXPERIMENT 1: THE EFFECT OF SNR THRESHOLD ON IDEAL BINARY MASKING

This experiment was designed to extend the analysis in Roman and Woodruff (2011) by comparing the performance of the three proposed masks across a range of RC values.

A. Method

1. Stimuli

Speech intelligibility in different conditions was evaluated by measuring the percentage of correctly identified sentences. The target speech signals used all contain the same

male speaker reading individual sentences from the HINT corpus (Nilsson *et al.*, 1994). The HINT corpus consists of 25 lists, each containing 10 sentences that are phonetically balanced and equated for naturalness, difficulty, length, and reliability (Nilsson *et al.*, 1994). Sentences within each list follow a predictable subject-verb-object syntactic structure, and range in length between four and seven words. Monaural signals were recorded in a controlled environment and digitized at 22.05 kHz with 16-bit quantization.

Following the setup employed in Roman and Woodruff (2011), impulse responses that simulate room reverberation were generated using the image method of Allen and Berkley (1979). A small rectangular room was simulated with

dimensions of $15\text{ m} \times 13\text{ m} \times 3.3\text{ m}$. The reflective characteristics of the room surfaces were set to be frequency-independent and to be the same at each surface so that a single parameter controlled the T_{60} time. The emitting source and the microphone were then fixed at $[9.5, 11, 1.2\text{ m}]$ and $[9.5, 7, 1.2\text{ m}]$, respectively. The 4 m distance between the emitting source and the microphone ensured that the reverberant sound field was predominant (Nabelek and Robinson, 1982). Monaural impulse responses were generated assuming an omni-directional microphone. We use this impulse response model to analyze relationships between different mask definitions and to generate performance profiles as a function of SNR threshold; the methods used, however, ignore frequency-dependent variations in sound absorption, interaction between reflections and do not provide listeners with all of the cues that would be available when listening in real environment.

Additionally, speech shaped noise (SSN), generated by filtering white noise with the long-term average spectrum of the male talker used as target speech, was used as an interference. Two acoustic conditions were tested: $T_{60} = 0.4\text{ s}$ with -2 dB SNR and $T_{60} = 0.8\text{ s}$ with -1 dB SNR, both of which correspond to the 50% SRT determined in (Roman and Woodruff, 2011). Both target and interference were convolved with the same monaural impulse response. Multiple copies of the target speech utterances were concatenated before convolving with the impulse response and the last copy was used as the reverberant signal. This ensured that the resulting signal was masked by reverberation throughout the presentation. The root-mean-square (RMS) level of the reverberant target signal was normalized to match the RMS level of a 64 dB SPL white noise signal across all sentence trials. The SSN interference was scaled and added to the reverberant target speech to obtain the desired mixture SNR.

IBMs were calculated for each reflection boundary and RC value, for each of the two mixture conditions tested in the experiment. As discussed in Sec. II, an IBM identifies the T-F units in a given mixture where the difference of the desired signal energy and the residual signal energy exceed a local SNR threshold, which is equal to the sum of the given RC value and the computed effective SNR. Note that for the same condition, the effective SNR of different sentences may vary slightly. All IBMs were computed as described in Section II and applied to the mixture signals in a synthesis step to generate the corresponding waveform stimuli (Wang and Brown, 2006). Waveform stimuli for unprocessed conditions were also generated by passing a signal through the same analysis and resynthesis process using an all-1 binary mask. This ensured that the same bandwidth constraints imposed by the filterbank were present in all stimuli.

2. Subjects

Fourteen normal-hearing, native American English speaking listeners participated in this experiment. Their ages varied between 21 and 33 with an average of 24. The subjects were paid for their participation. Although their audiograms were not evaluated, the subjects reported that they are unaware of any hearing problems.

3. Procedure

The pool of subjects was divided in two such that each listener participated in only one reverberant condition. For each such condition, IBM_0 , IBM_{50} , and IBM_{∞} were computed for the following range of RC values: -30 , -15 , -9 , -6 , -3 , 0 , and 6 dB . Finally, each alternative IBM was applied to the mixture signals and waveform stimuli were generated. Two unprocessed conditions were added: the unprocessed reverberant speech signal (i.e., no SSN interference added) and the unprocessed mixture signal.

The evaluation consisted of two phases: A training phase followed by accuracy level testing in the twenty-three different test conditions. Each evaluation took around 60 min. A break after the first half an hour was administered and the subject was instructed that additional breaks were possible as needed. The subject was allowed enough time to repeat the sentence or to guess the content. While the subject was seated in a sound attenuating booth, an operator interacted with a MATLAB program on a PC to report if the sentence was repeated correctly. A sentence was considered correct if all keywords were correct. The only substitutions allowed were a/the, an/the, is/was, are/were, has/had, and have/had. The stimuli were preprocessed and presented to the listener diotically through Sennheiser HD 280 Pro headphones, thus creating a midline intracranial sound image for both the target and the interference. As stated above, the RMS level of the reverberant speech signal was fixed across all trials. The training session, which familiarized the listener with the test procedure, was performed with clean sentences using one held out list of the HINT database. All listeners obtained 100% recognition on this training data. For each test condition, one unique list from the HINT database was used and intelligibility was measured as the percentage of correctly repeated sentences out of the ten sentences contained in the list. To facilitate subsequent statistical analysis of the results, the sequence of test conditions for each subject and the unique HINT list used for each of the twenty-three test conditions were randomized for each listener separately.

B. Results

Figure 2 shows the average percentage of correctly identified sentences for both test conditions. The top panel corresponds to the condition with $T_{60} = 0.4\text{ s}$ and -2 dB SNR, while the bottom panel shows results for the condition with $T_{60} = 0.8\text{ s}$ and -1 dB SNR. The results are given as functions of RC for each of the following three mask types: IBM_0 , IBM_{50} , and IBM_{∞} . Each dot on the graph corresponds to the average score of seven listeners. The results for the unprocessed conditions, which correspond to the intelligibility of the reverberant signal alone (UNP-R) and the unprocessed mixture (UNP) are also added to the left of the figures.

The unprocessed condition UNP resulted in a 41% accuracy level for the $T_{60} = 0.4\text{ s}$ condition and 42% accuracy level for the $T_{60} = 0.8\text{ s}$ condition, which are only slightly lower than the 50% accuracy level expected (Roman and Woodruff, 2011). The unprocessed condition corresponding to the reverberant speech signal without background noise, UNP-R, resulted in 100% accuracy for the $T_{60} = 0.4\text{ s}$ and

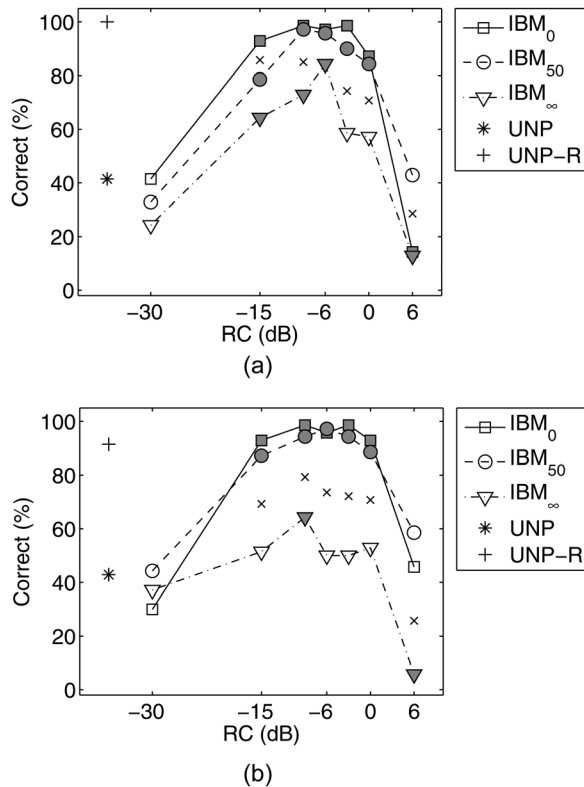


FIG. 2. Percentage of correctly identified sentences for ideal binary mask outputs as functions of the relative criterion. The acoustical condition in (a) is $T_{60}=0.4$ s, $\text{SNR}=-2$ dB while the condition in (b) is $T_{60}=0.8$ s, $\text{SNR}=-1$ dB. The results are given for IBM_0 , IBM_{50} , IBM_{∞} , the unprocessed mixture condition (UNP) and the unprocessed reverberant speech condition (UNP-R). The filled symbols indicate scores that are significantly different from the unprocessed mixture condition, and “x” indicates a significant difference between scores obtained for adjacent reflection boundaries with the same RC.

91.4% accuracy in the $T_{60}=0.8$ s, which shows only a slight degradation relative to anechoic speech for the longer reverberation time. This result is in accordance with the degradation in performance observed in the experiments of Nabelek and Robinson (1982).

The effective SNR varies greatly across the three different binary mask definitions. For $T_{60}=0.4$ s, the average effective SNRs are roughly -8 , -3 , and -2 dB for IBM_0 , IBM_{50} , and IBM_{∞} , respectively, while the effective SNRs are -12 , -5 , and -1 dB for $T_{60}=0.8$ s. Measuring performance as a function of RC rather than LC aligns the curves obtained for the different reflection boundaries, as predicted. Note that as the RC value increases the IBMs retain fewer and fewer T-F units causing the drop in performance observed on the right side of the curves. As the RC value decreases the IBMs retain more and more T-F units until the condition becomes equivalent to the unprocessed condition. In order to illustrate the effect of effective SNR on IBM processing, we also display in Fig. 3 the performance as a function of LC ($\text{RC}+\text{effective SNR}$).

To investigate the similarities between the different binary mask conditions, we compared subject mean percentages across all reflection boundaries and RC values using a *post hoc* protected Fisher’s LSD test using the results presented in Fig. 2. The test was applied separately for each mixture con-

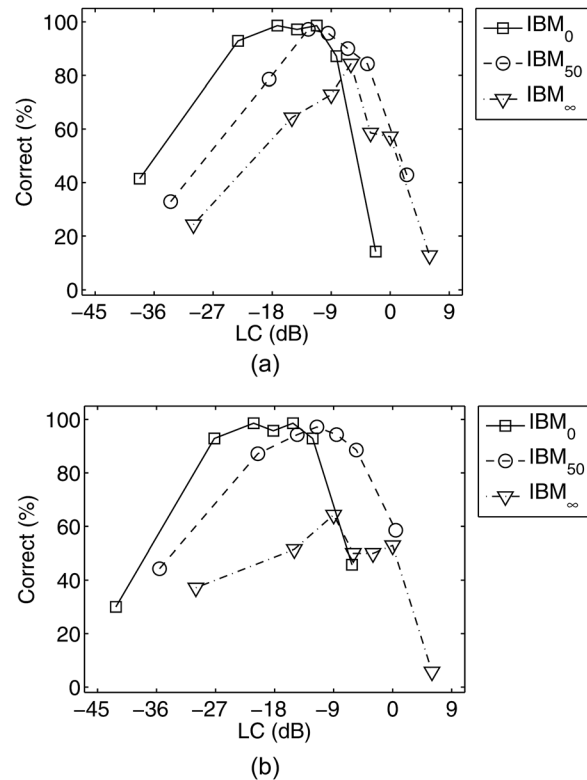


FIG. 3. Percentage of correctly identified sentences for the ideal binary mask outputs shown in Fig. 2 as functions of local criterion.

dition and subject mean percentages were transformed using the rationalized arcsine transform (Studebaker, 1985). The symbol “x” in the figures indicates significant pairwise differences ($\alpha=0.05$) between two adjacent reflection boundaries computed for the same RC value.

The analysis shows that while all the curves in the $T_{60}=0.4$ s have plateau regions where the performance is significantly better than the unprocessed condition, the gains in the IBM_{∞} condition are significantly lower than those obtained using IBM_0 and IBM_{50} masks except for $\text{RC}=-9$. Moreover, the IBM_0 and IBM_{50} curves are statistically similar at all RC values. In the $T_{60}=0.8$ s, while a similar performance pattern is observed for the IBM_0 and IBM_{50} masks, the curve for IBM_{∞} has maximum values similar to the unprocessed condition (slightly elevated at $\text{RC}=-9$ dB). Note that the performance of the IBM_{∞} condition cannot be explained by the distortion due to reverberation in the individual T-F units retained by the binary mask since listeners achieved significantly better performance in the UNP-R condition. The main observation is that the direct sound IBM and the 50 ms early reflections IBM both produce ceiling performances whereas the performance of reverberant IBM decreases as reverberation time increases.

For $T_{60}=0.4$ s, a two-way analysis of variance (ANOVA) on the rationalized arcsine transformed subject mean percentages from all conditions revealed that the two main effects of reflection boundary and RC were significant [$F(2, 126)=30.88$, $p<0.001$; $F(6, 126)=58.82$, $p<0.001$] but that there was no significant interaction. The results for $T_{60}=0.8$ s revealed that the effects of reflection

boundary and RC were significant [$F(2, 126) = 106.74$, $p < 0.001$; $F(6, 126) = 48.35$, $p < 0.001$]. There was also a significant interaction between the reflection boundary and RC [$F(12, 126) = 4.15$, $p < 0.001$].

C. Discussion

The results of this experiment show that both the IBM₀ and IBM₅₀ definitions lead to ceiling intelligibility scores, whereas the IBM_∞ definition does not. The poor performance of the IBM_∞ masks, particularly for the condition with $T_{60} = 0.8$ s, is consistent with Roman and Woodruff (2011). A key finding in this experiment is that the intelligibility of mixtures processed using IBM_∞ is significantly lower than the intelligibility of unprocessed reverberant speech without additive noise. This suggests that binary masking may not be capable of effectively restoring the perception of a reverberant target signal by removing additive noise and, consistent with Roman and Woodruff (2011), that IBM_∞ is a poor choice for measuring performance of a separation algorithm. Further, by testing over a large range of RC values we show that the outcome of the experiment in Roman and Woodruff (2011), was not the result of a suboptimal LC.

On the other hand, the high intelligibility scores achieved using the IBM₀ masks show that the reason for the poor performance reported in Roman and Woodruff (2011) was due to the use of LC equal to -6 dB. In Fig. 3, where results are plotted as a function of LC, we see that -6 dB is to the right of the plateau region for the IBM₀ definition. While the ceiling performance of the IBM₀ mask seems to suggest that it is not necessary to include early reflections in the IBM definition, a closer inspection of the masked signals shows the contrary. In the range of RC values for which the IBM₀ and IBM₅₀ definitions produce highly intelligible signals (roughly -15 – 0 dB), the two definitions essentially retain both direct sound and early reflections and attenuate late reflections and additive interference; however each mask accomplishes this in a different way. For IBM₅₀, early reflections are directly included in the definition of the desired signal. While early reflections are not included in the desired signal for the IBM₀ definition, the shift to lower LC values due to the lower effective SNR (relative to IBM₅₀) indirectly captures the energy of early target reflections. As listed in Table I, the percentage of direct sound and early reflections (up to 50 ms) energy retained is similar for both

TABLE I. Percentage of retained signal energy for IBM₀ and IBM₅₀ shown as functions of RC. “DS+ER” denotes the combination of direct sound and early reflections (up to 50 ms), while “LR+Noise” denotes the combination of late reflections (beyond 50 ms) and additive noise.

		RC				
		−15	−9	−6	−3	0
IBM ₀	DS+ER	99.0	97.4	95.4	91.1	81.2
	LR+Noise	52.8	42.3	35.5	27.8	19.1
IBM ₅₀	DS+ER	99.7	98.6	96.9	93.7	87.8
	LR+Noise	55.7	43.7	36.4	28.5	20.3

mask definitions. The percentages are computed for the second condition considered above ($T_{60} = 0.8$ s and -1 dB SNR) using the whole HINT database. A finer analysis with the IBM₀ definition shows that, given the frame lengths and filterbank used in this study, it is not possible to substantially attenuate early reflections energy while retaining direct sound energy. Thus, we contend that directly accounting for early reflections in the definition of the desired signal using a reflection boundary motivated by existing work is the more appealing of the two alternatives.

IV. EXPERIMENT 2: THE EFFECT OF REFLECTION BOUNDARY ON IDEAL BINARY MASKING

The results of experiment 1 show that the ideal binary masking technique can be utilized effectively for speech separation in reverberant conditions. Specifically, the direct-sound mask and the early reflection based mask are statistically similar given that a shift in the local SNR threshold is applied to account for the effective SNR, whereas the reverberant target mask is suboptimal. With this outcome in mind we expect there to be a range of functional reflection boundaries from 0 ms (direct sound) to an upper bound value. Experiment 2 is designed to explore this range by testing reflection boundaries up to 200 ms. To highlight differences due to reflection boundary values, and because the functional range may depend on the reverberation time, we consider a variety of reverberant conditions in this experiment. Three conditions consider reverberation times long enough to degrade intelligibility without additive noise and thus serve to test whether binary masking can be used effectively for speech dereverberation.

A. Method

The procedures used in the second experiment were nearly identical to those used in the first experiment. As in experiment 1, the HINT corpus was used for the target speech to evaluate the accuracy level of different conditions. The synthetic room impulse responses however have been generated as the multiplication of a white noise signal with an exponential envelope decay, as used previously in other studies (see, for example, George *et al.*, 2010). This approach again ignores frequency-dependent absorption and further, generates a greater density of reflections during the early time period than would be measured in real environments. However, we expect the model to be capable of revealing performance trends as a function of reflection boundary and threshold parameters. Specifically, we let

$$h(t) = v(t)e^{-6.91t/T_{60}}, \quad (6)$$

where $v(t)$ is the noise signal and the time constant for the envelope decay is equal to $T_{60}/6.91$ (Kuttruff, 2000). Impulse responses were truncated to have length equal to T_{60} . This methodology provides the following advantages over the image method simulation: (1) the parameter space is reduced to T_{60} and (2) it allows studying the impact of long reverberation times. Four T_{60} times were investigated: 0.8, 2, 3, and 30 s. The larger reverberation times (2 s and above) are

expected to have a significant effect on speech intelligibility (George *et al.*, 2010), while the moderate reverberation with $T_{60}=0.8$ s results in only a slight intelligibility degradation in the absence of noise (see the UNP-R condition from experiment 1). In order to create conditions with substantially degraded intelligibility so that the effect of IBM processing is significant, additive noise is required in the $T_{60}=0.8$ s condition. Specifically, we add speech-shaped noise to achieve a 3 dB input SNR for the $T_{60}=0.8$ s condition. Both the noisy condition tested here, $T_{60}=0.8$ s and 3 dB input SNR, and the reverberant only condition, $T_{60}=2$ s, correspond to a 50% speech reception reverberation threshold (SRRT) as investigated in George *et al.* (2010). The condition with $T_{60}=3$ s serves to exaggerate the differences due to reflection boundary. The condition with $T_{60}=30$ s was considered as it corresponds to an extreme condition where the unprocessed signal resembles a stationary background noise. This condition was motivated by the binary noise gating experiments in Wang *et al.* (2009) and Kjems *et al.* (2009), which found that the spectro-temporal modulation information carried by the binary mask pattern was sufficient for speech recognition. This experiment essentially replicates the prior studies for the reverberant case. This allows for direct analysis of the degradation in mask pattern due to inclusion of target reflections in the desired signal.

A total of 28 native American English speaking listeners with normal-hearing that did not participate in the previous experiments participated in this experiment. Their ages varied between 19 and 32 with an average of 22. As before, the subjects were paid for their participation and they reported that they are unaware of any hearing problems. Moreover, they obtained 100% accuracy on the training data.

The pool of subjects was divided such that each listener participated in only one reverberation condition (i.e., $T_{60}=0.8, 2, 3$, or 30 s). For each reverberation time, the following values for the reflection boundary were considered: 5, 50, 100, and 200 ms. The 50 ms value corresponds to the early reflection boundary suggested in literature for speech perception (Bradley *et al.*, 2003). Both 100 and 200 ms values were chosen as they have been reported in the acoustic literature to be division points beyond which late reflections produce no benefits for speech and music (Barron, 1993; Hidaka *et al.*, 1995). Note that with exponentially decaying synthetic impulse responses, one is not able to directly compute the direct sound based IBM, thus IBMs obtained for a 5 ms reflection boundary have been used here instead. At 5 and 50 ms reflection boundary values the following RC values have been tested: $-30, -15, -9, -6, -3, 0$, and 6 dB. Note that the design includes points spaced more closely between -9 and 0 dB, as this encompasses a plateau region with near perfect intelligibility in the case of speech mixed with a speech-shaped masker (Kjems *et al.*, 2009). For the 100 ms reflection boundary, the following RC values have been tested: $-15, -9, -6, -3$, and 0 dB. This is a subset of the set used in the 5 and 50 ms conditions. Limited by the size of the HINT database, we use only four RC levels for the 200 ms condition: $-15, -9, -6$, and 0 dB. Consequently, there are a total of 24 conditions tested for each reverberation time: one unprocessed condition, seven conditions for

IBM₅, seven conditions for IBM₅₀, five conditions for IBM₁₀₀, and four conditions for IBM₂₀₀.

B. Results

Figure 4 shows the percentage of correctly identified sentences for the conditions tested in experiment 2. The four panels correspond to the four acoustical conditions: $T_{60}=0.8$ s and 3 dB input SNR, $T_{60}=2$ s, $T_{60}=3$ s and $T_{60}=30$ s, respectively. In each panel, the results are given as functions of RC for each one of the following four mask types IBM₅, IBM₅₀, IBM₁₀₀, and IBM₂₀₀. Each dot on the graph corresponds to the average score of seven listeners. The results for the unprocessed condition are also added in each panel to the left of the RC curves.

The results in the unprocessed condition for the first two reverberant settings, $T_{60}=0.8$ s with 3 dB input SNR and $T_{60}=2$ s, resulted in 56 and 55 % accuracy level, respectively, which are only slightly higher than the 50% level expected (see George *et al.*, 2010). As predicted, the performance drops as reverberation increases and an accuracy level of 36% is observed in the $T_{60}=3$ s condition. Moreover, for the $T_{60}=30$ s listeners report that noise is presented instead of speech.

To investigate the similarities between the different binary mask definitions, we compared subject mean percentages across all reflection boundaries and RC values using a *post hoc* protected Fisher's LSD test. Again, the test was applied separately for each mixture condition and subject mean percentages were transformed using the rationalized arcsine transform. This analysis shows that all the curves for reflection boundaries in the range 5–100 ms have plateau regions where the performance is significantly better than the unprocessed condition. For the 0.8, 2, and 3 s T_{60} times, the 5 and 50 ms curves are statistically similar at all RC values. While the performance of the 100 ms curves is generally lower compared to the 5 and 50 ms curves, there are still several points which are not statistically different. Specifically, the similarities occur at RC values $-15, -9$, and -6 dB in the $T_{60}=0.8$ s condition, at RC values -9 and -6 dB in the $T_{60}=2$ s condition and at RC value -6 dB in the $T_{60}=3$ s condition. As expected, the differences due to reflection boundaries are increased as the reverberation time is increased. The curves for 200 ms show maximum values similar to the unprocessed condition at large reverberation times, but we observe that for $T_{60}=0.8$ s, the 200 ms curve does show significantly elevated scores when compared with the unprocessed condition (though is still generally suboptimal). The ceiling performance achievable by IBM processing tends to be slightly higher in the less extreme conditions (i.e., for $T_{60}=0.8$ s and $T_{60}=2$ s, which both correspond to 50% SRRT). However, consistent with (Wang *et al.*, 2009; Kjems *et al.*, 2009), the IBM is able to reconstruct the speech in the extreme case of $T_{60}=30$ s, which essentially corresponds to masked speech-shaped noise, and produce an accuracy level of 80%. Note that in this case the differences in plateau regions are statistically significant across all four reflection boundaries. As in experiment 1, the effective SNR varies as a

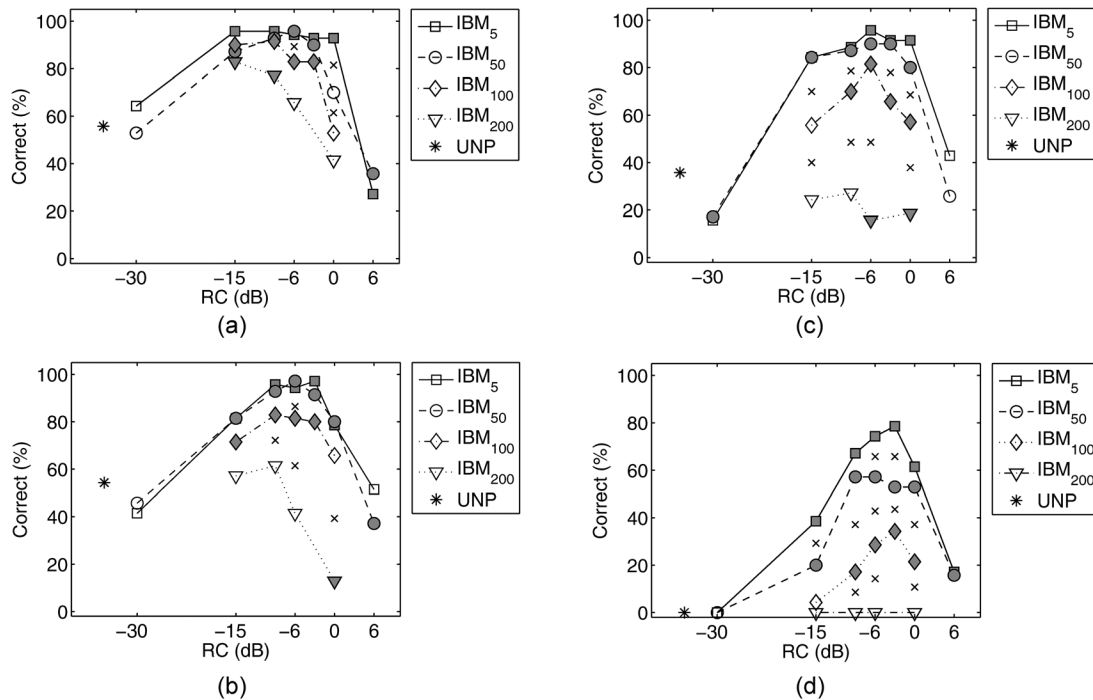


FIG. 4. Percentage of correctly identified sentences for ideal binary mask outputs as functions of reflection boundaries. Four reflection boundaries are used (5, 50, 100, and 200 ms) in four different acoustical settings: (a) $T_{60} = 0.8$ s and SNR = 3 dB, (b) $T_{60} = 2$ s, (c) $T_{60} = 3$ s, and (d) $T_{60} = 30$ s. The results are presented as functions of the relative criterion. Results for the unprocessed conditions are shown on the left of each plot. The filled symbols indicate scores that are significantly different from the unprocessed condition, and “x” indicates a significant difference between scores obtained for adjacent reflection boundaries with the same RC.

function of reflection boundary and T_{60} . We show the effective SNR for the experiment 2 conditions in Table II.

A three-way ANOVA on the rationalized arcsine transformed subject mean percentages from all conditions revealed that all three main effects of acoustical condition, reflection boundary and RC value were significant [$F(3, 552) = 303.10$, $p < 0.001$; $F(3, 552) = 155.24$, $p < 0.001$; $F(6, 552) = 166.54$, $p < 0.001$]. There was also a significant interaction between the condition tested and the reflection boundary [$F(9, 552) = 13.04$, $p < 0.001$] as well as between the condition tested and the RC value [$F(18, 552) = 7.80$, $p < 0.001$].

C. Discussion

The results from this experiment generally confirm the expectation developed based on the outcome of experiment 1; that there is a functional range of reflection boundaries from 0 ms to an upper bound value for which IBMs that improve intelligibility can be defined. For the three conditions with physically plausible T_{60} times (0.8, 2, and 3 s) we see that, consistent with experiment 1, the main difference

between the 5 and 50 ms masks is a shift in the optimal range of LC values (between roughly 10.5 and 11.5 dB, depending on the condition). We also see that IBMs that use reflection boundaries of 100 ms and larger are suboptimal. In longer reverberation times, masks defined with reflection boundary of 200 ms do not improve intelligibility relative to the unprocessed signals. These results largely agree with existing work on the effectiveness of early reflections to improve speech intelligibility in reverberation, which set the limit on beneficial reflections in the 50 to 100 ms range (Bradley *et al.*, 2003; Barron, 1993).

An interesting result in this experiment is the performance profile in the $T_{60} = 30$ s condition, which as noted above, serves as a follow up to Wang *et al.* (2009) and Kjems *et al.* (2009) for reverberant signals. As the unprocessed signal becomes essentially a stationary speech-shaped noise, any informative speech cues are carried by the mask patterns. We see that increasing the reflection boundary to include more target reflections in the desired signal gradually degrades performance, but nevertheless, IBMs for reflection boundaries up to 100 ms induce some recognition capability.

V. CONCLUDING REMARKS

It is well known that human speech perception is affected by environmental factors such as room reverberation and background noise. The ideal binary mask has been proposed as a performance upper bound for binary time-frequency masking methods that seek to improve speech intelligibility in adverse conditions. Although IBM processing has been shown to improve intelligibility for both normal-hearing and hearing-impaired listeners across a variety of conditions with

TABLE II. Effective SNR (in dB) for each reverberant condition and reflection boundary tested in experiment 2.

	Condition			
	0.8 s/3 dB	2 s	3 s	30 s
IBM ₅	-12.6	-15.2	-16.7	-26.8
IBM ₅₀	-2.1	-3.8	-5.9	-16.4
IBM ₁₀₀	0.8	0	-2.3	-13.3
IBM ₂₀₀	2.6	4.7	1.8	-10.2

anechoic speech and background noise, IBM processing has not been extensively studied in reverberant conditions. The experiments presented in this paper have been designed to both analyze whether binary masking can be used effectively for the purpose of improving intelligibility of reverberant speech, and also to analyze how the IBM should be defined in reverberant conditions. Motivated by the observation that early reflections are beneficial for speech intelligibility (Bradley *et al.*, 2003), we propose to parameterize the IBM by a division point between early and late reflections called the reflection boundary. While the experiments presented here refer to normal-hearing listeners with an optimal average age (Nabelek and Robinson, 1982), we expect to observe that our results generalize well to other listener categories.

Experiment 1 expands the analysis in Roman and Woodruff (2011). IBMs for the direct sound, the early reflections, and the reverberant target speech as desired signal are compared across a range of RC values. We note that the direct sound based mask and the early reflection based mask are statistically similar and achieve ceiling intelligibility scores compared to the reverberant binary mask that performs close to the unprocessed condition. The main difference from the study of Roman and Woodruff (2011) consists in accounting for the shift in effective SNR across the different binary mask definitions by considering the relative criterion—difference of local criterion and effective SNR—rather than the local criterion. Consistent with the previous findings, our results show that early target reflections should be preserved by the IBM, but that it is possible to accomplish this by either lowering the SNR threshold to account for treating early reflections as detrimental (as is done by the direct sound IBM), or by directly including early reflections in the definition of the desired signal. Given that early reflections are known to be beneficial rather than detrimental to intelligibility, we contend that increasing the reflection boundary to a reasonable value (e.g., 50–100 ms) is more consistent with human speech perception and, therefore, the more conceptually appealing of the two alternatives.

Experiment 2 tests a range of reflection boundaries for the IBM definition across different reverberation times. These tests show that the intelligibility of IBM-processed signals generally degrades with increasing reverberation time for reflection boundaries of 100 ms and beyond. Again, this result agrees with existing work that shows that late reflections are detrimental to speech perception by acting as masking noise (Bradley *et al.*, 2003). Several conditions in these experiments support previous work (Roman and Woodruff, 2011; Hazrati and Loizou, 2012) in establishing that binary masking can effectively overcome the combined effects of reverberation and additive noise to improve intelligibility. Finally, the conditions with $T_{60} = 2$ s and $T_{60} = 3$ s in experiment 2 help to establish that binary masking is capable of acting as a dereverberation strategy, which may be an appealing alternative to more complex strategies based on inverse filtering (Kokkinakis *et al.*, 2011).

The impulse responses used in experiments 1 and 2 are only approximately realistic. We expect room characteristics that are ignored in our simulations (e.g., frequency-dependent absorption, more complex reflection patterns) to have an

impact on the intelligibility of reverberant speech, and on the relative performance between different binary mask definitions. Further, we expect that overall performance would improve to some extent with binaural presentation.

ACKNOWLEDGMENTS

The authors thank two anonymous reviewers for their constructive criticisms and suggestions. The authors also thank DeLiang Wang for his contributions to the discussions underlying this paper and Fan-Gang Zeng for his help with the HINT corpus.

- Allen, J. B., and Berkley, D. A. (1979). "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.* **65**, 943–950.
- ANSI (1997). S3.5, *American National Standard: Methods for the Calculation of the Speech Intelligibility Index* (American National Standards Institute, New York).
- Assmann, P., and Summerfield, Q. (2004). "The perception of speech under adverse conditions," in *Speech Processing in the Auditory System*, edited by S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. R. Fay (Springer, New York), pp. 231–308.
- Barron, M. (1993). *Auditorium Acoustics and Architectural Design* (Spon Press/Taylor and Francis, London), p. 18.
- Boothroyd, A. (2004). "Room acoustics and speech perception," *Semin. Hear.* **25**, 155–166.
- 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.
- Brungart, D., Chang, P. S., Simpson, B. D., and Wang, D. L. (2006). "Isolating the energetic component of speech-on-speech masking with an ideal binary time-frequency mask," *J. Acoust. Soc. Am.* **120**, 4007–4018.
- Brungart, D., Chang, P. S., Simpson, B. D., and Wang, D. L. (2009). "Multitalker speech perception with ideal time-frequency segregation: Effects of voice characteristics and number of talkers," *J. Acoust. Soc. Am.* **125**, 4006–4022.
- Buchner, H., Aichner, R., and Kellermann, W. (2004). "Blind source separation for convolutive mixtures: A unified treatment," in *Audio Signal Processing for Next-Generation Multimedia Communications Systems*, edited by Y. Huang and J. Benesty (Kluwer Academic, Dordrecht), pp. 255–294.
- Cooke, M. P., Green, P., Josifovski, L., and Vizinho, A. (2001). "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Commun.* **34**, 267–285.
- Duquesnoy, A. J., and Plomp, R. (1980). "Effect of reverberation and noise on the intelligibility of sentences in cases of presbycusis," *J. Acoust. Soc. Am.* **68**, 537–544.
- George, E. L. J., Festen, J. M., and Houtgast, T. (2008). "The combined effects of reverberation and nonstationary noise on sentence intelligibility," *J. Acoust. Soc. Am.* **124**, 1269–1277.
- George, E. L. J., Goverts, S. T., Festen, J. M., and Houtgast, T. (2010). "Measuring the effects of reverberation and noise on sentence intelligibility for hearing-impaired listeners," *J. Speech Lang. Hear. Res.* **53**, 1429–1439.
- Goldsworthy, R. L., and Greenberg, J. E. (2004). "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Am.* **116**, 3679–3689.
- Greenberg, J. E., Peterson, P. M., and Zurek, P. (1993). "Intelligibility-weighted measures of speech-to-interference ratio and speech system performance," *J. Acoust. Soc. Am.* **94**, 3009–3010.
- Han, K., and Wang, D. L. (2011). "An SVM based classification approach to speech separation," in *Proceedings of ICASSP*, pp. 5212–5215.
- Hazrati, O., and Loizou, P. C. (2012). "Tackling the combined effects of reverberation and masking noise using ideal channel selection," *J. Speech Lang. Hear. Res.* **55**, 500–510.
- Hidaka, T., Beranek, L. L., and Okano, T. (1995). "Interaural cross-correlation, lateral fraction, and low- and high-frequency sound levels as measures of acoustical quality in concert halls," *J. Acoust. Soc. Am.* **98**, 988–1007.

- Hu, Y., and Loizou, P. (2008). "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Proc.* **16**, 229–238.
- Kim, G., Lu, Y., Hu, Y., and Loizou, P. (2009). "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Am.* **126**, 1486–1494.
- Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T., and Wang, D. L. (2009). "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Am.* **126**, 1415–1426.
- Kokkinakis, K., Hazrati, O., and Loizou, P. C. (2011). "A channel-selection criterion for suppressing reverberation in cochlear implants," *J. Acoust. Soc. Am.* **129**, 3221–3232.
- Kuttruff (2000). *Room Acoustics* (Taylor & Francis, New York), p. 86.
- 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.* **113**, 2237–2248.
- Li, N., and Loizou, P. C. (2008). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Am.* **123**(3), 1673–1682.
- Li, Y., and Wang, D. (2009). "On the optimality of ideal binary time-frequency masks," *Speech Commun.* **51**, 230–239.
- Lochner, J. P. A., and Burger, J. F. (1964). "The influence of reflections on auditorium acoustics," *J. Sound Vib.* **1**, 426–454.
- Loizou, P., and Kim, G. (2011). "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Trans. Audio, Speech, Lang. Proc.* **19**, 47–56.
- Mandel, M. I., Weiss, R. J., and Ellis, D. P. W. (2010). "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Proc.* **18**, 382–394.
- Miyoshi, M., and Kaneda, Y. (1988). "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Proc.* **36**, 145–152.
- Moore, B. C. J. (2003). *An Introduction to the Psychology of Hearing*, 5th ed. (Academic Press, London, U.K.), pp. 1242–1248.
- Nabelek, 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.
- Nilsson, M., Soli, S., and Sullivan, J. (1994). "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," *J. Acoust. Soc. Am.* **95**, 1085–1099.
- Pedersen, M. S., Wang, D. L., Larsen, J., and Kjems, U. (2008). "Two-microphone separation of speech mixtures," *IEEE Trans. Neural Netw.* **19**, 475–492.
- Quackenbush, S., Barnwell, T., and Clements, M. (1988). *Objective Measures of Speech Quality* (Prentice Hall, Englewood Cliffs, NJ), p. 373.
- Roman, N., Srinivasan, S., and Wang, D. L. (2006). "Binaural segregation in multisource reverberant environments," *J. Acoust. Soc. Am.* **120**, 4040–4051.
- Roman, N., Wang, D. L., and Brown, G. (2003). "Speech segregation based on sound localization," *J. Acoust. Soc. Am.* **114**, 2236–2252.
- Roman, N., and Woodruff, J. (2011). "Intelligibility of reverberant noisy speech with ideal binary masking," *J. Acoust. Soc. Am.* **130**, 2153–2161.
- Srinivasan, S., Roman, N., and Wang, D. L. (2006). "Binary and ratio time-frequency masks for robust speech recognition," *Speech Commun.* **48**, 1486–1501.
- Srinivasan, S., and Wang, D. L. (2010). "Robust speech recognition by integrating speech separation and hypothesis testing," *Speech Commun.* **52**, 72–81.
- Studebaker, G. A. (1985). "A 'rationalized' arcsine transform," *J. Speech Hear. Res.* **28**, 455–462.
- Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2011a). "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Proc.* **19**, 2125–2136.
- Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2011b). "An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech," *J. Acoust. Soc. Am.* **130**, 3013–3027.
- Wallach, H., Newman, E. B., and Rosenzweig, M. R. (1949). "The precedence effect in sound localization," *Am. J. Psychol.* **52**, 315–336.
- Wang, D. L. (2005). "On ideal binary masks as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Boston, MA), pp. 181–197.
- Wang, D. L., and Brown, G. J. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley/IEEE Press, Hoboken, NJ).
- Wang, D. L., Kjems, U., Pedersen, M. S., Boldt, J. B., and Lunner, T. (2009). "Speech intelligibility in background noise with ideal binary time-frequency masking," *J. Acoust. Soc. Am.* **125**, 2336–2347.
- Woodruff, J., and Wang, D. L. (2010). "Sequential organization of speech in reverberant environments by integrating monaural grouping and binaural localization," *IEEE Trans. Acoust., Speech, Signal Proc.* **18**, 1856–1866.