Effect of the division between early and late reflections on intelligibility of ideal binary-masked speech

Junfeng Li[a] and Risheng Xia
Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China

Qiang Fang and Aijun Li
Institute of Linguistics, Chinese Academy of Social Sciences, Beijing 100732, China

Jielin Pan and Yonghong Yan
Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China

(Received 20 December 2014; revised 17 March 2015; accepted 24 March 2015)

The ideal binary mask (IBM) that was originally defined in anechoic conditions has been found to yield substantial improvements in speech intelligibility in noise. The IBM has recently been extended to reverberant conditions where the direct sound and early reflections of target speech are regarded as the desired signal. It is of great interest to know how the division between early and late reflections impacts on the intelligibility of the IBM-processed noisy reverberant speech. In this present study, the division between early and late reflections in three rooms was first determined by four typical estimation approaches and then used to compute the IBMs in reverberant conditions. The IBMs were then applied to the noisy reverberant mixture signal for segregating the desired signal, and the segregated signal was further presented to normal-hearing listeners for word recognition. Results showed that the IBMs with different divisions between early and late reflections provided substantial improvements in speech intelligibility over the unprocessed mixture signals in all conditions tested, and there were small, but statistically significant, differences in speech intelligibility between the different IBMs in some conditions tested.
 [http://dx.doi.org/10.1121/1.4919287]

[JFL] Pages: 2801–2810

I. INTRODUCTION

Speech is the most natural means of human−human communication. However, it is often distorted in everyday listening conditions by ambient noise, competing voice, and reverberation. In the past several decades, many studies on speech perception have revealed that human speech understanding remains remarkably robust in adverse listening conditions where various kinds of interferences are present (Assmann and Summerfield, 2004; Meyer *et al.*, 2013; Bradley *et al.*, 1999; Lavandier and Culling, 2008; George *et al.*, 2008; George *et al.*, 2010). The ability of humans to segregate the target signal from an acoustic mixture in adverse conditions is generally thought to involve the process of auditory scene analysis (ASA) (Bregman, 1990). Inspired by the principles of ASA, increased attention has been given to computational auditory scene analysis (CASA) (Wang and Brown, 2006).

Motivated by the auditory masking phenomenon, the research in CASA has suggested that its computational goal for segregating speech from noise is provided by the ideal binary mask (IBM) (Wang, 2005). The key idea of the IBM is to retain the time-frequency units of an acoustic mixture in which target signal is stronger than noise by a certain local criterion (LC), and to discard the remaining units (Wang and Brown, 2006). A series of recent studies have demonstrated that the IBM can dramatically improve the intelligibility of speech masked by different types of noise for normal-hearing and hearing-impaired listeners (Brungart *et al.*, 2006; Anzalone *et al.*, 2006; Li and Loizou, 2008; Brungart *et al.*, 2009; Wang *et al.*, 2009). Brungart *et al.* (2006) showed that the IBM is very effective for normal-hearing listeners to improve speech intelligibility in the presence of competing voice, and found a plateau of nearly perfect intelligibility of an IBM-masked mixture with a range of the LC (from −12 to 0 dB). By using different speech material and filter banks, Li and Loizou (2008) found larger intelligibility benefits for normal-hearing listeners when varying the LC parameter, and a wider range of the LC (from −20 to 5 dB) with almost perfect intelligibility. Anzalone *et al.* (2006) adopted a slightly different approach for computing the IBM by comparing the target signal to a fixed threshold to retain a certain percentage of the total target energy, and reported that more than 9 dB improvement in speech reception threshold (SRT) was obtained for hearing-impaired listeners. Wang *et al.* (2009) demonstrated that the IBM processing yielded an 11 dB improvement in SRT in the cafeteria noise and a 7 dB improvement in speech-shaped noise for normal-hearing listeners, and the SRT improvements were 16 and 9 dB for hearing-impaired listeners in cafeteria noise and speech-shaped noise.

The IBMs used in the aforementioned studies were all defined in anechoic conditions. In addition to noise, reverberation present in many speaking conditions blurs temporal and spectral cues, and flattens formant transitions, which drastically degrades speech intelligibility (Assmann and

[a]Electronic mail: junfeng.li.1979@gmail.com

Summerfield, 2004; Meyer *et al*., 2013; Lavandier and Culling, 2008; George *et al*., 2008; George *et al*., 2010). As suggested by studies in room acoustics (Kuttruff, 2009), reverberation is generally considered to consist of three parts: direct sound, early reflections, and late reverberation. Inspired by the beneficial contribution of early reflections to speech intelligibility (Bradley *et al.,* 2003), Roman and Woodruff (2011) proposed a novel approach for computing the IBM in reverberant conditions by regarding the direct path and early reflections of target signal as the desired signal, and showed that this new IBM processing yielded the 8 and 5.5 dB improvements in SRT over unsegregated signals in speech-shaped noise and simultaneous-talker noise. Hu and Kokkinakis (2014) recently examined the impact of this IBM processing on speech intelligibility in reverberant conditions for cochlear implant listeners. In the computation of this IBM, a fixed threshold of 50 ms was used to distinguish early and late reflections (Roman and Woodruff, 2011, 2013; Hu and Kokkinakis, 2014), which was mainly motivated by the finding in Bradley *et al*. (2003) where the early reflections of up to 50 ms have been found to benefit speech perception for normal-hearing and hearing-impaired listeners. As reverberation varies with the acoustical properties of enclosures, however, the division between early and late reflections should be redefined for each specific acoustic environment (Kuttruff, 2009).

A number of approaches have been reported in the literature to estimate the time boundary that divides early and late reflections, which is generally referred to as *mixing time* in room acoustics. Among these approaches, the fixed values, for instance 50 ms for speech and 80 ms for music, have been suggested to divide early and late reverberation, which are regardless of room properties (Bradley and Soulodre, 1995; Kuttruff, 2009). With the fixed time boundary, Bradley *et al*. (2003) showed the positive impact of early reflections (<50 ms) on speech intelligibility in room environments. Since reverberation varies with the acoustic properties of the room (e.g., surface absorption), several time ranges were suggested for the mixing time, rather than the fixed values, for example, 100−150 ms (Kuttruff, 1993) and 50−200 ms (Hidaka *et al*., 1995). Motivated by the physical generation procedure of early and late reflections, some model-based estimators were reported to correlate the mixing time with the volume and surface area of the room (Polack, 1993; Rubak and Johansen, 1999). The model-based approaches for determining the mixing time of a given room have been widely used in the hybrid reverberation methods in which the early and late reflections were usually generated by different simulation approaches (Jot, 1992; Stewart and Sandler, 2007; Li *et al*., 2012). Based on the Gaussian property of late reverberation, several signal-based algorithms were designed for calculating the mixing time by tracking the statistical characteristics of reverberant signals through certain objective statistical measures (Abel and Huang, 2006; Stewart and Sandler, 2007; Hidaka *et al*., 2007). The signal-based approaches have found their applications in the field of speech dereverberation where the late reverberation was considered harmful for speech perception and automatic speech recognition, and needed to be removed (Gillespie *et al*., 2001; Kokkinakis and Loizou, 2009; Kumatani *et al*., 2011).

From the above-mentioned studies, it is known that the IBM processing yields substantial improvements in speech intelligibility under noise and reverberant conditions, and that the computation of the IBM suggested by Roman and Woodruff (2011) greatly relies on the mixing time that divides early and late reflections. However, very little is so far known about the extent to which the division between early and late reflections does affect the intelligibility of noisy reverberant speech processed by the IBM. Therefore, this present study examines the effect of the mixing time estimated by four state-of-the-art approaches on the intelligibility of the IBM-masked speech in noisy reverberant conditions. Given the configuration (e.g., volume and surface) or the measured impulse responses in three typical rooms, the mixing times were first determined according to the above approaches. Subsequently, the IBMs were created using the early and entire reverberant target and noise signals, and then applied to the noisy reverberant mixture signals that were generated by adding the reverberant noise signal into the reverberant target speech signal at two SNR levels. The IBM-masked signals were then presented to normal-hearing listeners for word identification. Evaluation results indicated that all the IBMs with the different mixing times provided substantial improvements in speech intelligibility over the unprocessed mixture signals in all conditions tested, and there were small, but statistically significant, differences in speech intelligibility between the different IBMs calculated with different divisions between early and late reflections in some conditions tested.

II. IDEAL BINARY MASK IN REVERBERANT CONDITIONS

As the computational goal of CASA for sound separation, the IBM has been extended from anechoic conditions to reverberant conditions by Roman and Woodruff (2011) where the direct sound plus early reflections of target speech are regarded as the desired signal. The computing procedure of this extended IBM in reverberant conditions is briefly summarized below.

Similar to the computation of the IBM in anechoic conditions (Wang and Brown, 2006), the IBM in reverberant conditions is also computed from the cochleagram representation of target and noise signals. Specifically, the cochleagram is calculated from the outputs of a 64-channel gammatone filterbank with the center frequencies from 50 to 8000 Hz equally spaced on the equivalent rectangular bandwidth scale. The output of each filter in the filterbank is divided using 20-ms rectangular frames with 10-ms overlap into a set of time-frequency units, and the cochleagram corresponds to a two-dimensional response energy computed across all the time-frequency units. Suppose that $D(k, l)$ and $R(k, l)$ be the energy (in dB) of the desired and residual signals in the kth frequency channel and the lth time frame, the IBM is then defined as (Roman and Woodruff, 2011)

$$\mathrm{IBM}(k,\ell) = \begin{cases} 1, & D(k,\ell) - R(k,\ell) > \mathrm{LC}; \\ 0, & \text{otherwise}. \end{cases} \quad (1)$$

Finally, the desired signal can be segregated by applying the IBM to the noisy reverberant mixture signal in a synthesis step (Wang and Brown, 2006). It is noted that the desired signal $D(k, \ell)$ consists of both the direct path and early reflections of the target signal, which is derived by convolving the target signal with the direct plus early impulse responses. The residual signal $R(k, \ell)$ is obtained by subtracting the desired signal $D(k, \ell)$ from the noisy reverberant mixture signal. It is clear that the residual signal $R(k, \ell)$ includes both the late reverberant target signal and the reverberant noise signal. The division between early and late reflections could be determined using the four mixing time estimators described in Sec. III.

III. DIVISION BETWEEN EARLY AND LATE REFLECTIONS

As introduced in Sec. I, a number of mixing time estimators have been reported in the literature to divide early and late reflections, which are generally classified into three categories: the fixed estimators (Begault, 1992; Bradley *et al.*, 2003), the model-based estimators (Rubak and Johansen, 1999), and the signal-based estimators (Abel and Huang, 2006; Stewart and Sandler, 2007). Specifically, four representative mixing time estimators were evaluated in this present study, including the fixed mixing time ($t_m = 50$ ms), the model-based estimator designed on the room volume and surface, and the signal-based estimators reported in Abel and Huang (2006) and Stewart and Sandler (2007). The definitions of the four mixing time estimators are detailed below.

A. Fixed mixing time estimator

Based on the knowledge of room acoustics, Bradley *et al.* (2003) suggested a constant value of 50 ms for the mixing time that is independent of the environment, and found that early reflections (<50 ms) greatly contributed to speech intelligibility. This fixed mixing time (50 ms) that was widely used in many previous studies (Roman and Woodruff, 2011, 2013; Hu and Kokkinakis, 2014) was also examined in this present study.

B. Model-based mixing time estimator

According to Sabine's theory in room acoustics, the mean free path length in a diffuse sound field is given by (Kuttruff, 2009)

$$l_m = 4V/S, \quad (2)$$

where S is the total surface area of the room and V its volume. With the assumption of a diffuse field for late reverberation, Rubak and Johansen (1999) suggested to estimate the mixing time from the mean free path length l_m, given by

$$t_{\mathrm{model}} = 4l_m/c \approx 47V/S, \quad (3)$$

where c denotes the sound velocity.

C. Abel's mixing time estimator

Inspired by the important role of echo density in the perceptual quality of reverberation, Abel and Huang (2006) proposed to determine the mixing time based on the normalized echo density profile (NED). Given the room impulse response $h(t)$, the NED is defined as the fraction of impulse response taps lying outside the window standard deviation, described as

$$\eta(n) = \frac{1}{\mathrm{ercf}\left(1/\sqrt{2}\right)} \sum_{l=n-\delta}^{n+\delta} w(l) \mathbf{1}\{|h(l) > \sigma|\}, \quad (4)$$

where $\mathrm{ercf}\,(1/\sqrt{2}) = 0.3173$ is the expected fraction of samples lying outside the standard deviation from the mean for a Gaussian distribution, $\mathbf{1}\{\cdot\}$ is the indicator function which returns 1 when its argument is true and 0 otherwise, $w(l)$ is a weighting function having the unit sum $\sum_l w(l) = 1$, and σ is the standard deviation of the impulse response within the window. The NED starts from zero and approaches one as reverberation evolves from early reflections to late reverberation. In order to account for the natural fluctuation in the NED for real rooms, Abel and Huang (2006) further suggested a new NED threshold $1 - \sigma_{\mathrm{late}}$, where σ_{late} denotes the standard deviation of the NED in late reverberation. The mixing time is finally determined as the instant where the NED becomes $1 - \sigma_{\mathrm{late}}$ for the first time.

D. Stewart's mixing time estimator

Based on the assumption of a diffuse field for late reverberation, Stewart and Sandler (2007) suggested calculating the mixing time using the kurtosis, which is a fourth-order moment of a statistic process, defined as

$$\gamma_4 = \frac{E(x-\mu)^4}{\sigma^4} - 3, \quad (5)$$

where E denotes the expectation operator, μ is the mean, and σ is the standard deviation of the process x. Since this kurtosis measure typically approaches zero toward the end of an impulse response due to its increasing Gaussian nature, the mixing time is determined as the time instant when the kurtosis calculated in the sliding window reaches zero.

IV. EXPERIMENT I: THE EFFECT OF DIVISION BETWEEN EARLY AND LATE REFLECTIONS ON THE IBM-MASKED REVERBERANT SPEECH IN STATIONARY NOISE

A. Method

1. Stimuli

In the speech intelligibility test, the database developed by Ma and Shen (2004) was adopted as the target speech material. This database consists of 10 tables, each of which contains 75 phonetically balanced Chinese words with consonant-vowel structure. In each table, every three words are randomly combined to form one nonsense sentence, producing a total of 25 sentences. The sentences are uttered by

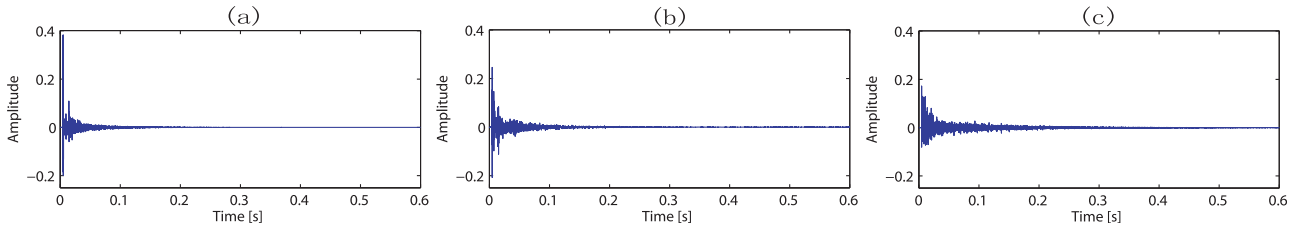

FIG. 1. (Color online) The room impulse responses recorded in the meeting room (a), office room (b), and lecture room (c), which were used in the present study.

one female speaker and recorded in a soundproof booth at a sampling rate of 16 kHz and stored in a 16 bit format. The interfering noise signal used in this listening test was the speech-shaped noise that was obtained by passing white noise through the average spectrum of the speech database.

To generate reverberant conditions, room impulse responses (RIRs) were selected from the database provided by Jeub *et al.* (2009) in which the RIRs were measured in four typical rooms with different dimensions and acoustic properties. To highlight the effect of reverberation on speech intelligibility, only the RIRs corresponding to the farthest distances between the sound source and the recording microphone in three rooms (office, meeting, and lecture) were adopted in the present intelligibility test. All the RIRs were stored with a sampling frequency of 48 kHz. The RIRs used in this study are plotted in Fig. 1 and the configurations and acoustic properties of the three rooms are listed in Table I.

Both target speech and noise signal were first upsampled to 48 kHz and then convolved with the RIR in each room condition, which was followed by being downsampled to 16 kHz. The long-term mean square level of reverberant target speech in the absence of noise was fixed across all sentences in all conditions tested. The reverberant noise signal was scaled accordingly to reach the desired SNR level and then added to the reverberant target signal at the SNRs of 0 and 5 dB, generating the noisy reverberant mixtures.

For the RIR in each room condition, its mixing time was determined using the four estimation approaches described in Sec. III, and the results are listed in Table II. The estimated mixing times were used for separating early reflections from the RIRs, which were further exploited to generate the different IBMs. Specifically, the early reverberant target signal $d(n)$ was generated by convolving the separated early reflections with anechoic target signal, and the remaining signal $r(n)$ was obtained by subtracting $d(n)$ from the noisy reverberant mixture signal. Based on the time-frequency representations of $d(n)$ and $r(n)$, the IBMs can be computed according to Eq. (1). The derived IBMs corresponding to the four mixing time estimators were denoted as $\text{IBM}_{\text{fixed}}$, $\text{IBM}_{\text{model}}$, IBM_{abel}, and $\text{IBM}_{\text{stewart}}$, respectively. The IBMs were finally applied to the noisy reverberant mixture signals in a synthesis step to generate the segregated waveform stimuli (Wang and Brown, 2006).

TABLE I. The configurations and acoustic properties of the three rooms examined in this study. d_{LM} denotes the distance between the sound source and the recording microphone in collecting the room impulse responses.

	Meeting room	Office room	Lecture room
Room dimensions [m]	[8.00 5.00 3.10]	[5.00 6.40 2.90]	[10.80 10.90 3.15]
d_{LM} [m]	2.8	3.0	10.2
Reverberation time [s]	0.25	0.48	0.83

2. Subjects

Ten normal-hearing listeners (five females and five males) participated in this experiment. All subjects were native Chinese-speaking listeners and were paid for their participation. Their ages varied between 23 and 28 with an average of 25.

3. Procedure

In each environment, there were a total of seven testing conditions, including the noiseless anechoic speech signal (CLEAN), the noiseless reverberant speech signal (REVERB), the noisy reverberant mixture signal (MIXTURE), and the processed signals by four IBMs ($\text{IBM}_{\text{fixed}}$, $\text{IBM}_{\text{model}}$, IBM_{abel}, and $\text{IBM}_{\text{stewart}}$). The conditions of CLEAN and REVERB were included for comparison. All these stimuli were presented to each subject at a comfortable listening level through HDA-200 headphones in a soundproof booth.

Prior to the test, each subject went through a training session to become familiar with the testing procedure. In the test, each subject participated in a total of 34 listening conditions [2 SNRs $\times$ 3 rooms $\times$ 5 algorithms (4 IBMs + 1 MIXTURE) + 3 REVERBs + 1 CLEAN)]. One list of sentences (i.e., 25 sentences) was used per condition, and none of the lists was repeated across conditions. Thus, each subject listened to 850 nonsense sentences [25 sentences $\times$ 34 conditions] in the listening test. In the test, the presentation orders of the stimuli and the listening conditions were randomized across each subject. Subjects were asked to write down the words they heard where they were allowed to guess the content. During the test, a break was administered after every six testing conditions and additional breaks were possible as needed.

TABLE II. Estimated mixing time values in the three reverberant environments tested.

	Estimated mixing time [ms]			
Conditions	Fixed	Model	Abel	Stewart
Meeting	50	36.29	40.75	44.31
Office	50	33.52	28.67	50.90
Lecture	50	46.83	26.77	30

B. Results

The mean percentages of words correctly identified in speech-shaped noise across 10 subjects for the seven testing conditions in the three rooms are plotted in Fig. 2. The error bars represent the standard errors of the mean.

As shown in Fig. 2, the word recognition scores for the CLEAN condition was 95.73%, and were 97.09%, 92.06%, and 87.45% for the REVERB condition in the meeting, office and lecture rooms. It is obvious that speech intelligibility degraded as the amount of reverberation increased. For the MIXTURE condition where both speech-shaped noise and reverberation existed, the speech recognition scores averaged across two SNRs further dropped to 73.27% in the meeting room, 66.33% in the office room, 55.47% in the lecture room. In the three rooms tested, the speech intelligibility ratings for the unprocessed MIXTURE condition at SNR of 0 dB were consistently much lower than those at SNR of 5 dB. The speech intelligibility ratings in the MIXTURE condition were 68.53% and 78.00% at 0 and 5 dB in the meeting room, 60.13% and 72.53% in the office room, and 46.53% and 64.40% in the lecture room. In comparison of the word recognition scores in the MIXTURE condition, the IBMs that were computed using the mixing times determined with the four estimators greatly increased the speech recognition scores in all conditions tested. For instance, in the meeting room at SNR of 5 dB, the IBM_{fixed} processing yielded the highest speech intelligibility ratings up to 95.87%, which was about 18% improvement over the MIXTURE condition; in the lecture room at SNR of 0 dB, the IBM_{model} processing improved speech intelligibility from 46.53% to 82%, which corresponded to a speech intelligibility improvement of larger than 35%. Moreover, the slight differences in speech intelligibility were observed across the four IBMs with different mixing times in some conditions tested. For example, the intelligibility ratings of the processed speech in the lecture room at SNR of 0 dB were 76.40% for IBM_{fixed}, 82.00% for IBM_{model}, 71.60% for IBM_{abel}, and 79.73% for $IBM_{stewart}$.

To examine the effects of reverberant rooms (meeting, office, and lecture), SNR levels (0 and 5 dB) and processing conditions (1 CLEAN + 1 REVERB + 1 MIXTURE + 4 IBMs), the word recognition scores were subjected to statistical analysis using the scores as the dependent variable, and the room, SNR and processing condition as the three within-subjects factors. Three-way analysis of variance (ANOVA) with repeated measures indicated the significant effects of room effect [$F(2, 18) = 91.23$, $p < 0.001$], SNR [$F(1, 9) = 74.46$, $p = 0.09$], and processing condition [$F(6, 54) = 99.95$, $p < 0.001$]. There were significant interactions between room and SNR [$F(2, 18) = 24.36$, $p < 0.001$], between room and processing condition [$F(12, 108) = 13.68$, $p < 0.001$], and between SNR and processing condition [$F(6, 54) = 51.97$, $p < 0.001$]. There was also significant interaction among room, SNR and processing condition [$F(12, 108) = 3.93$, $p = 0.0001$].

Following ANOVA, to further investigate the similarities between the IBMs with different mixing times, the *post hoc* tests (multiple paired comparisons according to Ryan (1959) with appropriate correct) were done between the word recognition scores obtained by the different IBMs. A difference between word recognition scores was treated as significant if the significance level $p < 0.05$. The test was applied separately for each mixture condition at different SNRs and in different rooms. The symbol "*" in the figures indicates significant pairwise differences ($p < 0.05$) between two IBMs computed with different mixing times in that condition. The small, but statistically significant, differences in speech intelligibility ratings could be found across the four IBMs defined with the different mixing times in certain conditions, for example, in the lecture room at SNR of 0 dB.

C. Discussion

The results indicated that speech intelligibility degraded in reverberant conditions, which might be caused primarily by self-masking effects that give rise to flattened formant transitions and by overlap-masking effects that tend to fill the gaps in the temporal envelope of speech (Assmann and Summerfield, 2004). More severe degradation of speech intelligibility was found with increased reverberation time, as shown by the speech intelligibility ratings in the meeting

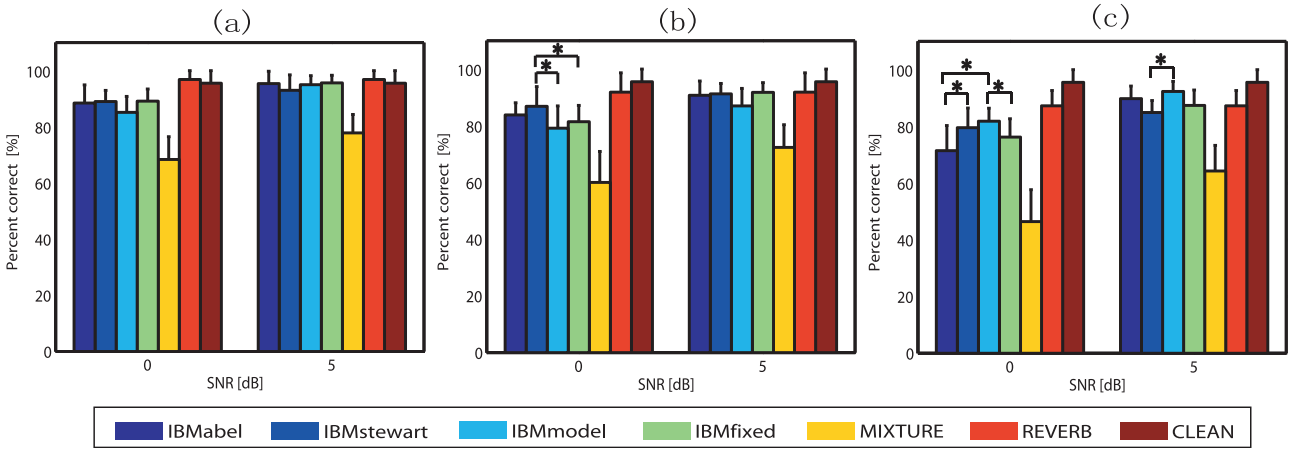

FIG. 2. (Color online) Mean percent correct scores of the IBM-masked speech signals in the *speech-shaped noise* condition. Three reverberant conditions were considered: (a) meeting room, (b) office room, and (c) lecture room. The error bars indicate standard deviations of the mean.

and lecture rooms, which was widely proved in many previous studies (Assmann and Summerfield, 2004; George *et al.*, 2008; George *et al.*, 2010; Hazrati and Loizou, 2012; Kokkinakis and Loizou, 2011; Roman and Woodruff, 2011, 2013). When additionally including the speech-shaped noise, the speech intelligibility ratings were further decreased by up to 28.56%, 31.93%, 40.92% at SNR of 0 dB in the meeting, office, and lecture rooms. The negative impact of noise was likely attributed to its mask on the weak consonants to a greater degree than the higher intensity vowels (Assmann and Summerfield, 2004). In the "worst" MIXTURE condition tested (i.e., at SNR of 0 dB in the meeting room) where both speech-shaped noise and reverberation were present, the degradation of speech intelligibility amounted up to 49.20% and 40.92% compared with those in the CLEAN condition and in the corresponding REVERB condition. This significant degradation of speech intelligibility came from the combined effects of noise and reverberation, which were more detrimental to speech intelligibility than either reverberation or noise-alone effects (Assmann and Summerfield, 2004; Roman and Woodruff, 2011; Hazrati and Loizou, 2012).

In comparison of the unprocessed MIXTURE condition, the IBMs defined with the four mixing time estimators consistently improved the speech intelligibility in the three rooms and at two SNRs, which was in line with the results in Roman and Woodruff (2011, 2013). In the "best" condition tested (i.e., in the meeting room at 5 dB), the IBMs yielded the nearly perfect speech intelligibility ratings, which were almost equivalent to that in the CLEAN condition. The high recognition accuracy indicated that the IBMs enabled recovering all the acoustic cues that are crucial for speech understanding from the noisy reverberant mixture signals in this "best" condition tested, which was also observed in Roman and Woodruff (2013). With the increased intensity of noise and reverberation, the IBMs failed to recover all the perceptually important acoustic cues in the lecture room at SNR of 0 dB where the speech intelligibility ratings of the IBMs-processed speech were lower than that in the CLEAN condition.

In the "more clean" conditions where the noise and reverberation were relatively low, for example, in the meeting room and in the office room at 5 dB, no significant difference in speech intelligibility was found across the IBMs with different estimated mixing times, as shown in Fig. 2. As listed in Table II, it was noted that quite different values were obtained for the mixing times by the four approaches described in Sec. III in both meeting and office rooms. The similar speech intelligibility ratings in these conditions provided by the IBMs that were defined with these quite different mixing times might be attributed to the noise-dominant effects in these low reverberant environments, and to the inclusion of early reflections in the computation of the IBMs. In the meeting room characterized by the low reverberation, since the harmful effect on speech intelligibility was mainly from the noise signal, the IBMs were therefore able to improve speech intelligibility, which was consistent with the results reported in the previous studies (Wang and Brown, 2006; Brungart *et al.*, 2006; Anzalone *et al.*, 2006; Wang *et al.*, 2009). In the office room at 5 dB where the effects of both reverberation and noise were modest, the IBMs with the different mixing times yielded the comparable speech intelligibility improvements. This result indicated that the inclusion of early reflections in the computation of IBMs was very important for speech understanding in reverberation, which was not so sensitive to the exact length of early reflections involved at least in this modest reverberant and relatively low noise condition.

In the "more adverse" conditions tested (in the lecture room and in the office room at 0 dB), significant differences in speech intelligibility were found between the IBMs with different estimated mixing times. In the lecture room, the IBM_{model} yielded the highest speech intelligibility ratings among the four IBMs tested, although the mixing time (46.83 ms) used in the computation of the IBM_{model} was between the minimum value given by the Abel's estimator (26.77 ms) and the maximum value given by the fixed estimator (50 ms). Therefore, it might be because that the IBM_{Abel} treated part of early reflections as harmful late reflections due to the too short mixing time used, and that the IBM_{fixed} regarded part of harmful late reflections as the desired signal due to the too large mixing time. The IBM_{model} adopted the appropriate value for the mixing time in its computation, yielded the highest speech intelligibility among the four IBMs tested, which could be more clearly observed in the lecture room at 0 dB. Therefore, the direct sound plus appropriate amount of early reflections were very important for improving speech understanding in reverberant conditions, which was consistent with the findings of Bradley *et al.* (2003) and Roman and Woodruff (2011, 2013). The similar results were also observed in the office room at the 0-dB condition where the $IBM_{stewart}$ took the appropriate amount of early reflections as the desired signal, yielding the highest speech recognition scores.

V. EXPERIMENT II: THE EFFECT OF DIVISION BETWEEN EARLY AND LATE REFLECTIONS ON THE IBM-MASKED REVERBERANT SPEECH IN NON-STATIONARY INTERFERING NOISE

The evaluation results of experiment I showed that the IBMs improved speech recognition scores over the unprocessed mixture signals in the stationary speech-shaped noise condition. Stationary and non-stationary noises were usually characterized by different properties in the time-frequency domain, for example, the high variation of temporal and spectral envelope for non-stationary interfering noise. Though normal-hearing listeners are able to use the gaps in the temporal envelope of speech under the non-stationary noise condition to preserve the intelligibility of the target speech, reverberation tends to fill these gaps (Assmann and Summerfield, 2004). Experiment II was designed to examine the abilities in speech intelligibility of the IBMs with the different divisions between early and late reflections when applied to the non-stationary noise (e.g., competing talker voice) reverberant condition.

A. Method

The target speech material and the RIRs were the same as those used in experiment I. The interfering noise was a competing talker's voice that was randomly selected for each sentence trail from a pool of 25 sentences uttered by a male speaker. Both the target speech and the interfering noise were first convolved with the corresponding RIR, and then mixed to generate the noisy reverberant mixture at SNRs of 0 and 5 dB. The mixing times that divide early and late reflections were estimated for each RIR according to the approaches described in Sec. III, which were further used for creating the early reverberant target signal and the reverberant interfering signal. The IBMs were computed according to Eq. (1) and then applied to the noisy reverberant mixture signals for segregating the desired signal.

Ten normal-hearing listeners that did not participate in experiment I participated in this experiment. Their ages varied between 23 and 38 with an average of 27 and they were paid for their participation. Prior to the testing, each subject listened to a training session as well to get familiar with the testing procedure. During the test, subjects participated in a total of 34 listening conditions which includes a total of seven different classes of stimuli (CLEAN, REVERB, MIXTURE, $\text{IBM}_{\text{fixed}}$, $\text{IBM}_{\text{model}}$, IBM_{abel}, and $\text{IBM}_{\text{stewart}}$) in the three rooms at two SNRs. Subjects were asked to write down the words they heard and a break was instructed during the test. The testing procedure was the same as that in experiment I.

B. Results

Figure 3 showed that the speech recognition ratings in the REVERB condition were lower than those in the CLEAN condition, especially when the reverberation time was relatively large (e.g., in the lecture room). After adding the non-stationary competing-talker noise, the average speech recognition ratings for the MIXTURE condition dropped from 97.75% to 81.07% in the meeting room, from 92.82% to 74.27% in the office room, and from 80.65% to 50.53% in the lecture room. Compared with the unprocessed MIXTURE condition, the IBMs computed with different mixing times consistently improved the speech recognition ratings in all rooms and SNRs tested. This benefit in speech intelligibility was more obvious in the "worst" condition tested (i.e., in the lecture room at 0 dB) where the speech intelligibility was improved from 40.13% to 74.47% on average by the four IBMs. The differences in speech intelligibility were also observed across the four IBMs with different mixing times in some other conditions tested. For example, the speech recognition ratings in the lecture room at 5 dB were 84.13% for $\text{IBM}_{\text{fixed}}$, 81.73% for $\text{IBM}_{\text{model}}$, 87.87% for IBM_{abel}, and 74.13% for $\text{IBM}_{\text{stewart}}$.

To examine the effects of reverberant rooms (meeting, office, and lecture), SNR levels (0 dB and 5 dB) and processing conditions (1 CLEAN + 1 REVERB + 1 MIXTURE + 4 IBMs), the word recognition scores were subjected to statistical analysis using the scores as the dependent variable, and the room, SNR and processing condition as the three within-subjects factors. Three-way ANOVA with repeated measures indicated significant effects of room effect [$F(2, 18) = 225.56$, $p < 0.001$] and processing condition [$F(6, 54) = 244.57$, $p < 0.001$], and no significant effect of SNR [$F(1, 9) = 3.56$, $p = 0.09$]. There were significant interaction between room and SNR [$F(2, 18) = 95.29$, $p < 0.001$], between room and processing condition [$F(12, 108) = 36.36$, $p < 0.001$], and between SNR and processing condition [$F(6, 54) = 5.57$, $p = 0.0001$]. There was also significant interaction among room, SNR and processing condition [$F(12, 108) = 25.28$, $p < 0.001$]. To further investigate the similarities between the IBMs with different mixing times, the *post hoc* tests with Ryan's method were done between the word recognition scores obtained by the different IBMs. The test was applied separately for each mixture condition at different SNR and in different room. The symbol "*" in Fig. 3 indicates significant pairwise differences ($p < 0.05$) between two IBMs computed with the different mixing times.

C. Discussion

Compared with the CLEAN and REVERB conditions, the MIXTURE condition yielded the consistent degradation of speech intelligibility especially in the high reverberant

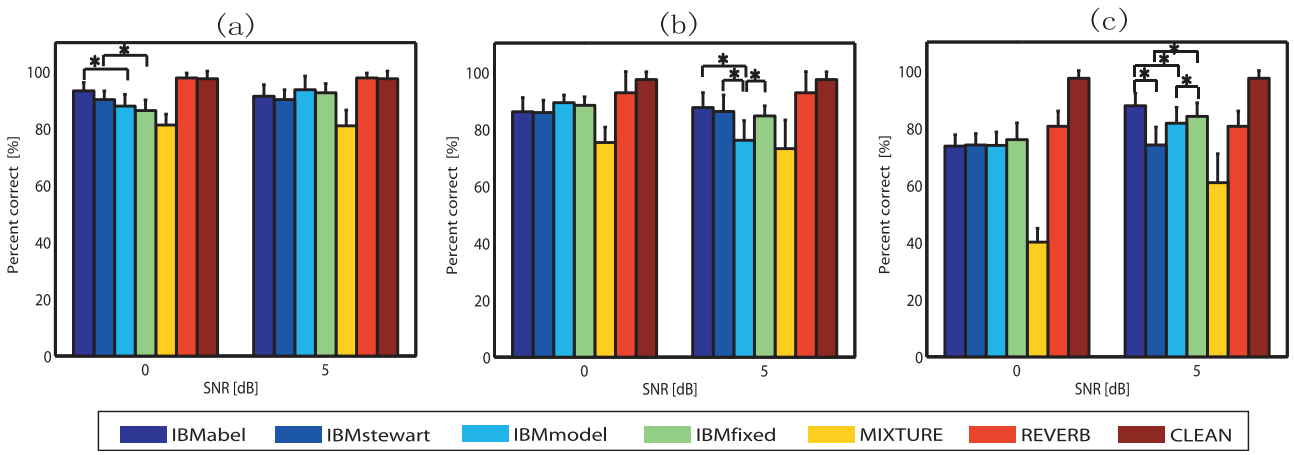

FIG. 3. (Color online) Mean percent correct scores of the IBM-masked speech signals in the *competing-talker interference* condition. Three reverberant conditions were considered: (a) meeting room, (b) office room, and (c) lecture room. The error bars indicate standard deviations of the mean.

lecture room, which was due to the combined effects of noise and reverberation and was in line with the results of experiment I. It was noted that in the low and modest reverberant environments (i.e., the meeting and office rooms), the degradation caused by the competing-talker voice was quite small relative to that by the speech-shaped noise. This result might come from the positive function of the gaps in the temporal envelope of speech in speech understanding under the non-stationary noise condition which helps to preserve the intelligibility of the target speech (Assmann and Summerfield, 2004; Roman and Woodruff, 2011, 2013). Speech intelligibility was largely degraded in the lecture room, which might be due to the blurred temporal cues and flattened formant transitions of the target and competing-talker voice (Assmann and Summerfield, 2004; Kokkinakis and Loizou, 2011; Hazrati and Loizou, 2012).

The IBMs defined with the four mixing time estimators improved speech intelligibility in all conditions tested, which was consistent with the results of experiment I and the results in Roman and Woodruff (2011, 2013). Non-stationary competing-voice interference was relatively sparse in the time-frequency domain even under reverberant conditions. Based on the sparsity of the target speech and non-stationary interfering signals, the IBMs might be able to eliminate most time-frequency units dominated by the interfering signal and to alleviate the negative effect of reverberation to a certain degree on the target speech, which contributed to speech understanding (Wang and Brown, 2006; George *et al.,* 2008; Brungart *et al.,* 2009; Roman and Woodruff, 2011).

Similar to the results obtained in the speech-shaped noise, it was found significant differences in speech intelligibility across the IBMs with different mixing times under some conditions tested. For instance, in the office room at 5 dB, the IBM_{model} provided the lowest speech recognition ratings among the four IBMs. This might be attributed to the relative small value (36.29 ms) for the mixing time, which led to that the IBM_{model} regarded part of early reflections of target signal as the undesired signal to be removed (Roman and Woodruff, 2011). No significant differences were found among the other three IBMs (IBM_{fixed}, IBM_{abel}, $IBM_{stewart}$), though different values for the mixing time were used in their computation. This indicated that the appropriate value for the mixing time to be used in the computation of IBM might be in a range, rather than a specific value, to improve speech intelligibility in noisy reverberant conditions.

VI. CONCLUDING REMARKS

It is well known that the IBM has been suggested as the computational goal of computational auditory scene analysis. The IBM was originally designed in anechoic conditions and was found capable to improve speech intelligibility for both normal-hearing and hearing-impaired listeners across a variety of anechoic conditions in the presence of background noise (Brungart *et al.,* 2006; Anzalone *et al.,* 2006; Li and Loizou, 2008; Wang *et al.,* 2009). Inspired by the observation that early reflections positively contribute to speech understanding (Bradley *et al.,* 2003), Roman and Woodruff (2011) proposed to extend the definition of IBM processing from anechoic to reverberant conditions where the direct path and early reflections of target speech are regarded as the desired signal. The effect of the division between early and late reflections (i.e., the mixing time) on speech intelligibility of the IBM-processed noisy reverberant signal has not been extensively studied. The experiments presented in this paper were designed to examine the effect of the division between early and late reflections on the intelligibility of the IBM-processed speech in three reverberant rooms under the stationary (speech-shaped) and non-stationary (competing-talker voice) noise conditions. Specifically, four typical approaches were considered for estimating the division between early and late reflections, including the fixed value (50 ms) (Bradley *et al.,* 2003; Roman and Woodruff, 2011), the model-based estimator (Rubak and Johansen, 1999) and two signal-based estimators (Abel and Huang, 2006; Stewart and Sandler, 2007). The results of the current study have important implications for the implementation of the IBM processing strategy in practical noisy reverberant conditions.

The results obtained from experiments I and II showed that the IBMs that regard the direct sound plus early reflections of target speech as the desired signal yielded substantial improvements in speech intelligibility for normal-hearing listeners under noisy reverberant conditions, which was consistent with the results in Roman and Woodruff (2011, 2013). Originated from the IBM definition in anechoic conditions, the IBMs examined in this study were able to recover the time-frequency units dominated by the desired signal that were important for speech understanding (Wang and Brown, 2006; Brungart *et al.,* 2006; Li and Loizou, 2008). Furthermore, the IBMs were able to improve speech intelligibility in reverberant conditions by regarding the early reflections of the target signal as the desired signal, which was attributed to the positive impact of early reflections to speech intelligibility (Bradley *et al.,* 2003; Roman and Woodruff, 2011).

In both speech-shaped noise and competing-talker interference conditions, speech intelligibility was gradually degraded as the amount of reverberation increased, such as from the degradation of the speech recognition ratings in the meeting room to that in the lecture room. It is usually believed that the degradation of speech intelligibility in reverberant conditions is mainly from the negative impact of late reflections on speech understanding (Assmann and Summerfield, 2004; George *et al.,* 2008; Kokkinakis and Loizou, 2011). In each reverberant room, it was found that the speech recognition ratings of the unprocessed mixture speech in the speech-shaped noise condition were much lower than those in the competing-talker interference condition. It might be because normal-hearing listeners could preserve speech intelligibility by using the gaps in the temporal envelope of speech in reverberation in the presence of non-stationary competing-talker interference (Assmann and Summerfield, 2004; Roman and Woodruff, 2011; Hazrati and Loizou, 2012).

In each room condition tested, the estimated values for the mixing time varied greatly across the four estimation approaches examined in the present study. The fixed estimator (50 ms) was empirically determined according to the

results of psychoacoustic experiments (Bradley and Soulodre, 1995; Bradley *et al.*, 2003). The mixing time estimated by the model-based approach was determined by the surface area and volume of the room (Rubak and Johansen, 1999). The signal-based estimator for the mixing time was mainly dependent on the statistical properties of the impulse response or the reverberant signal (Abel and Huang, 2006; Stewart and Sandler, 2007). The difference in the principle on which the three classes of mixing time estimators were designed might account for the variation of the estimated mixing times in each room condition. The difference between the two signal-based estimators might be mainly attributed to the different objective measures [shown in Eqs. (4) and (5)] in describing the Gaussian characteristics of the reverberant signal, and to the impulse responses that were recorded in real rooms, rather than the artificial ones created through a certain reverberation simulation method.

It is interesting that no significant difference in speech intelligibility was found across the processed signals by the IBMs defined with the different values of mixing time in some conditions tested. This implied that there might be a range of value rather than a specific value for the mixing time. The IBMs computed using the mixing time value that lies in the appropriate range provided the improved speech intelligibility to a comparable degree in these conditions. The similar speech intelligibilities obtained by the IBMs in the low reverberant conditions might be partially attributed to the dominant impact of background noise than reverberation. In the other conditions tested, the small but statistically significant difference in speech intelligibility was observed across the IBMs-processed signals. The IBM defined with the value in the appropriate range for the mixing time yielded the higher speech intelligibility, and the IBMs with the mixing time value out of the range yielded the relatively lower speech intelligibility in these conditions. This result might come from that the IBMs with the too low value for the mixing time regard part of early reflections as harmful and that the IBMs with the too large value for the mixing time regard the late reflections as beneficial. It implied that the IBM with the appropriate value for the mixing time integrated the beneficial part of early reflections in speech understanding, which greatly improved speech intelligibility in noisy reverberant conditions.

The four estimators examined in this present study were originally developed for estimating the mixing time for a given reverberant environment. Although the physical mixing time provided an important implication to the speech intelligibility of the noisy reverberant signals processed by the IBMs, one further work might be to develop a perceptual mixing time that could more directly quantify the impact of the division between early and late reflections on speech intelligibility. Moreover, the results of this present study would contribute to design a speech intelligibility enhancement system in noisy reverberant conditions.

ACKNOWLEDGMENTS

This work was partially supported by the National 973 Program (2013CB329302), the National Natural Science Foundation of China (No. 11461141004), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Nos. XDA06030100 and XDA06030500), and the National 863 Program (No. 2012AA012503).

Abel, J., and Huang, P. (**2006**). "A simple, robust measure of reverberation echo density," in *Proceedings of the 121st Convention of the Audio Engineering Society*, preprint 6985, pp. 1–10.

Anzalone, M., Calandruccio, L., Doherty, K., and Carney, L. (**2006**). "Determination of the potential benefit of time-frequency gain manipulation," Ear Hear. **27**, 480–492.

Assmann, P., and Summerfield, Q. (**2004**). "Perception of speech under adverse conditions," in *Speech Processing in the Auditory System*, edited by W. A. S. Greenbergy and A. Popper (Springer-Verlag, New York), pp. 231–308.

Begault, D. (**1992**). "Perceptual effects of synthetic reverberation on three-dimensional audio systems," J. Audio Eng. Soc. **40**, 895–904.

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

Bradley, J., and Soulodre, G. (**1995**). "The influence of late arriving energy on spatial impression," J. Acoust. Soc. Am. **97**, 2263–2272.

Bregman, A. (**1990**). *Auditory Scene Analysis* (MIT, Cambridge, MA), pp. 47–184, 411–453.

Brungart, D., Chang, P., Simpson, B., and Wang, D. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.

Brungart, D., Chang, P., Simpson, B., and Wang, D. (**2009**). "Multitalker speech perception with ideal time-frequency segregation: Effects of voice characteristics and number of talkers," J. Acoust. Soc. Am. **125**, 4006–4022.

George, E., Festen, J., and Houtgast, T. (**2008**). "The combined effects of reverberation and nonstationary noise on sentence intelligibility," J. Acoust. Soc. Am. **124**, 1269–1277.

George, E., Goverts, S., Festen, J., 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.

Gillespie, B., Malvar, H., and Florencio, D. (**2001**). "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proceedings of the IEEE Conference Acoustics, Speech, and Signal Processing*, pp. 3701–3704.

Hazrati, O., and Loizou, P. (**2012**). "Tackling the combined effects of reverberation and masking noise using ideal channel selection," J. Speech, Lang., Hear. Res. **55**, 500–510.

Hidaka, T., Okano, T., and Beranek, L. (**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.

Hidaka, T., Yamada, Y., and Nakagawa, T. (**2007**). "A new definition of boundary point between early reflection and late reverberation in reverberant environments," J. Acoust. Soc. Am. **122**, 326–332.

Hu, Y., and Kokkinakis, K. (**2014**). "Effects of early and late reflections on intelligibility of reverberated speech by cochlear implant listeners," J. Acoust. Soc. Am. **135**, EL22–EL28.

Jeub, M., Schafer, M., and Vary, P. (**2009**). "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proceedings of the International Conference on Digital Signal Processing*, pp. 1–5.

Jot, J. (**1992**). "An analysis/synthesis approach to real-time artificial reverberation," in *Proceedings of the IEEE Conference Acoustics, Speech, and Signal Processing*, pp. 221–224.

Kokkinakis, K., and Loizou, P. (**2009**). "Selective-tap blind dereverberation for two-microphone enhancement of reverberant speech," IEEE Signal Process. Lett. **16**, 961–964.

Kokkinakis, K., and Loizou, P. (**2011**). "The impact of reverberant self-masking and overlap-masking effects on speech intelligibility by cochlear implant listeners," J. Acoust. Soc. Am. **130**, 1099–1102.

Kumatani, K., McDonough, J., and Raj, B. (**2011**). "Maximum kurtosis beamforming with a subspace filter for distance speech recognition," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 179–184.

Kuttruff, H. (**1993**). "Auralization of impulse response modeled on the basis of ray-tracing results," J. Audio Eng. Soc. **41**, 876–880.
Kuttruff, H. (**2009**). *Room Acoustics*, 5th ed. (Spon Press, New York), Chaps. 4–6.
Lavandier, M., and Culling, J. (**2008**). "Speech segregation in rooms: Monaural, binaural and interacting effects of reverberation on target and interfere," J. Acoust. Soc. Am. **213**, 2237–2248.
Li, J., Xia, R., and Yan, Y. (**2012**). "A hybrid approach for simulation of room reverberation," J. Acoust. Soc. Am. **131**, 3217–3217.
Li, N., and Loizou, P. (**2008**). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," J. Acoust. Soc. Am. **123**, 1673–1682.
Ma, D., and Shen, H. (**2004**). *Acoustical Manual* (Chinese Science Publisher, Beijing), Chap. 19.
Meyer, J., Dentel, L., and Meunier, F. (**2013**). "Speech recognition in natural background noise," PLoS One **1**, 1–14.
Polack, J. (**1993**). "Playing billiards in the concert hall: The mathematical foundations of geometrical room acoustics," Appl. Acoust. **38**, 235–244.
Roman, N., and Woodruff, J. (**2011**). "Intelligibility of reverberant noisy speech with ideal binary masking," J. Acoust. Soc. Am. **130**, 2153–2161.
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.
Rubak, P., and Johansen, L. (**1999**). "Artificial reverberation based on a pseudo-random impulse response," in *Proceedings of the 106th Convention of the Audio Engineering Society*, preprint 4900, pp. 1–15.
Ryan, T. (**1959**). "Multiple comparison in psychological research," Psychol. Bull. **56**, 26–47.
Stewart, R., and Sandler, M. (**2007**). "Statistical measures of early reflections of room impulse responses," in *Proceedings of the 10th International Conference on Digital Audio Effects*, preprint DAFx-07, pp. 1–4.
Wang, D. (**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., and Brown, G. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley/IEEE Press, Hoboken, NJ), Chap. 3.
Wang, D., Kjems, U., Pedersen, M., Boldt, J., and Lunner, T. (**2009**). "Speech intelligibility in background noise with ideal binary time-frequency masking," J. Acoust. Soc. Am. **125**, 2336–2347.