Improvement of intelligibility of ideal binary-masked noisy speech by adding background noise

Shuyang Cao, Liang Li, and Xihong Wu[a)]
Department of Machine Intelligence, Peking University, Beijing 100871, China

(Received 14 July 2010; revised 25 November 2010; accepted 4 February 2011)

When a target-speech/masker mixture is processed with the signal-separation technique, ideal binary mask (IBM), intelligibility of target speech is remarkably improved in both normal-hearing listeners and hearing-impaired listeners. Intelligibility of speech can also be improved by filling in speech gaps with un-modulated broadband noise. This study investigated whether intelligibility of target speech in the IBM-treated target-speech/masker mixture can be further improved by adding a broadband-noise background. The results of this study show that following the IBM manipulation, which remarkably released target speech from speech-spectrum noise, foreign-speech, or native-speech masking (experiment 1), adding a broadband-noise background with the signal-to-noise ratio no less than 4 dB significantly improved intelligibility of target speech when the masker was either noise (experiment 2) or speech (experiment 3). The results suggest that since adding the noise background shallows the areas of silence in the time-frequency domain of the IBM-treated target-speech/masker mixture, the abruption of transient changes in the mixture is smoothed and the perceived continuity of target-speech components becomes enhanced, leading to improved target-speech intelligibility. The findings are useful for advancing computational auditory scene analysis, hearing-aid/cochlear-implant designs, and understanding of speech perception under "cocktail-party" conditions.
 [DOI: 10.1121/1.3559707]

PACS number(s): 43.71.Gv [MAH] Pages: 2227–2236

I. INTRODUCTION

The ideal binary mask (IBM) has been proposed as a signal-processing algorithm for computational auditory scene analyses (Wang, 2005), containing the binary values of 1 and 0 in a time-frequency (T-F) matrix. The spectrum of the signal/masker mixture is first decomposed in the frequency domain using a bank of gammatone filters (Patterson *et al*., 1988) and then the energy is assigned along the time domain (Wang and Brown, 2006). The IBM is defined as the comparison in signal-to-noise ratio (SNR) in each T-F unit [the element of the two-dimensional (2-D) T-F representation of the target/masker mixture] against a local SNR criterion (LC) [i.e., the threshold in decibel (dB)]: When the SNR within a T-F unit exceeds the LC, the unit is assigned the value of 1; when the local SNR is below the LC, the unit is assigned the value of 0. Thus, an IBM-treated signal/masker mixture can be synthesized by applying the T-F matrix with the binary values to the original signal/masker mixture (see panels A–D in Fig. 1). In some studies (e.g., Li and Loizou, 2008), short-time Fourier transform is also used to decrease frequency resolution at low frequencies and increase frequency resolution at high frequencies.

Obviously, varying the LC affects the effects of the IBM manipulation. Brungart *et al*. (2006) examined the intelligibility of a mixture processed by the IBM with different LCs and different maskers, and reported that there is a plateau in performance at LC values from −12 to 0 dB. They suggested that the choice of the LC of −6 dB, which is located around the center of the performance plateau, is better than the commonly used 0-dB LC for improving the speech intelligibility (also see Wang *et al*., 2009). Li and Loizou (2008) varied the LC and got a similar performance plateau ranging from −20 to 0 dB.

It has been well demonstrated that the IBM signal manipulation improves the speech intelligibility (e.g., Anzalone *et al*., 2006; Brungart *et al*., 2006, 2009; Li and Loizou, 2007, 2008; Wang *et al*., 2008, 2009; Kjems *et al*., 2009). For example, Li and Loizou (2008) found that the intelligibility benefit brought by the IBM manipulation is 7 dB under speech-shaped-noise masking, 10 dB under modulated-speech-shaped-noise masking, and 15 dB under two-talker-speech masking. Wang *et al*. (2009) found that in normal-hearing listeners, the IBM signal manipulation improves the speech intelligibility by 11 dB under cafeteria-noise masking and 7 dB under speech-shaped-noise masking. Interestingly, the improvement is even larger in hearing-impaired listeners: 16 dB under cafeteria-noise masking and 9 dB under speech-shaped-noise masking.

Nevertheless, since signals of the T-F units with SNRs below the LC are removed by the IBM manipulation, the 2-D T-F representation of the target-speech/masker matrix contains numerous IBM-induced "honeycomb-like" areas of sudden, audible silences (with the binary value of 0), leading to that the target-speech sound is interrupted by the temporal and spectral gaps (Fig. 1). It is known that when an interrupted target sound is filled with another louder sound, the target sound is perceived as a continuous stream through interruption. This phenomenon has been called as "auditory induction," "continuity illusion," "perceptual restoration," or "phonemic restoration," which is probably due to an

[a)]Author to whom correspondence should be addressed. Electronic mail: wxh@cis.pku.edu.cn

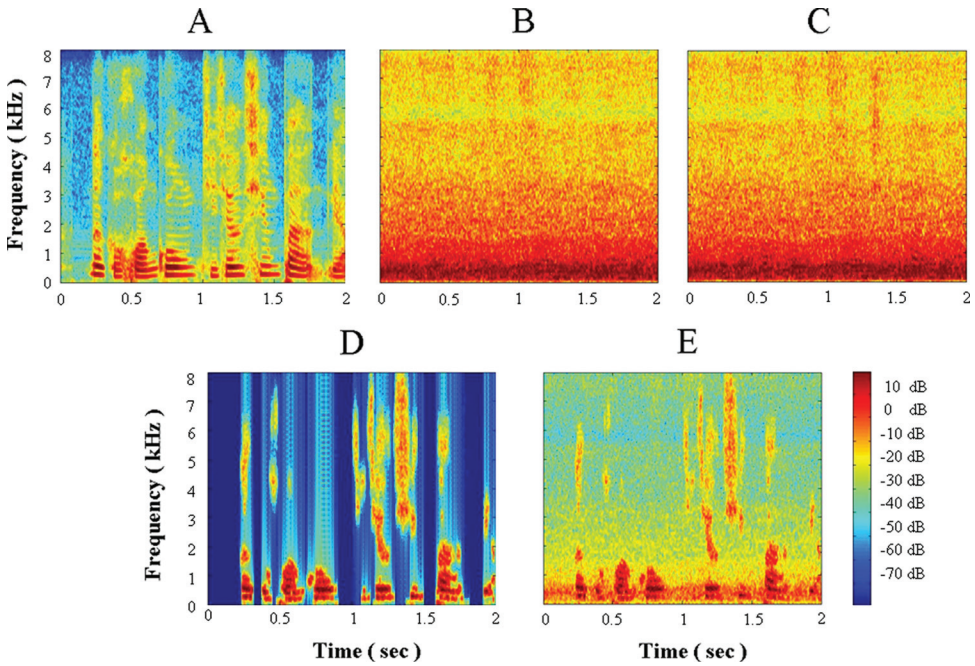

FIG. 1. The spectrograms of the following stimuli: A target-speech stimulus (panel A), the masking steady-state speech-spectrum noise (panel B), the target-speech/noise-masker mixture with the SNR of −10 dB when the IBM manipulation was not introduced (panel C), the IBM-treated target/masker mixture (panel D), and the IBM-treated target/masker mixture after the background noise with the SNR of 12 dB [relative to the pre-IBM-treated (the original unprocessed) target speech] was added (panel E).

enhancement of perceived continuity of target spectrotemporal energy (e.g., Bregman, 1990; Woods *et al*., 1996; Srinivasan and Wang, 2005; Shinn-Cunningham and Wang, 2008; Baskent *et al*., 2009). In addition to the perceived continuity, intelligibility of speech can also be improved by filling in speech gaps with un-modulated, steady-state noise (Warren, 1970; Powers and Wilcox, 1977; Bashford *et al*., 1992). Thus, it is important to investigate whether adding an un-modulated broadband-noise background with an appropriate SNR can improve intelligibility of target speech in the target/masker mixture that is treated by the IBM manipulation.

In this study, the effects of the IBM manipulation on intelligibility of Chinese target speech under masking conditions were first examined when the masker was steady-state speech-spectrum noise, two-talker Chinese speech, or two-talker English speech (experiment 1). And then, the effects of adding a noise background on intelligibility of target speech in the IBM-treated target/masker mixture were examined when the masker in the target-speech/masker mixture was speech-spectrum noise (experiment 2). Finally, in experiment 3, the effects of adding the noise background on intelligibility of target speech in the IBM-treated target/masker mixture were examined when the masker was either one-talker speech or two-talker speech.

In the present study, each nonsense sentence was presented only once. Since in total 36 and 72 testing conditions were used in experiment 2 and experiment 3, respectively, and 18 sentences were used for each testing condition, a large number of speech sentences were required for each of these two experiments. To satisfy the requirement of the large sentence numbers, and particularly to guarantee the quality of the speech stimuli, target sentences in both experiments 2 and 3 were recited by a synthesized voice. In fact, the artificially voiced sentences sounded very naturally.

II. EXPERIMENT 1: EFFECTS OF THE IBM MANIPULATION ON SPEECH INTELLIGIBILITY

A. Methods

1. Participants

Twelve young university students (19–25 yr old, eight females and four males) participated in experiment 1 but not in other experiments. In this and the following two experiments (experiments 2 and 3), all the participants, who had spoken Mandarin Chinese as the native language and had learned English for 7–10 yr, had normal [pure-tone threshold no more than 25 dB hearing level (HL)] and bilaterally symmetrical (no more than 15 dB difference between the two ears) hearing at frequencies from 125 to 8000 Hz, confirmed by the audiometry. They gave their written informed consent to participate in the experiments and were paid a modest stipend for their participation.

2. Stimuli

Speech stimuli were Chinese "nonsense" sentences, which are syntactically correct but not semantically meaningful. Direct English translations of the sentences are similar but not identical to the English nonsense sentences that were developed by Helfer (1997) and also used in studies by Freyman *et al*. (1999) and Li *et al*. (2004). Each of the Chinese sentences has 12 characters (also 12 syllables) including three key components: subject, predicate, and object, which are also the three keywords, with two characters (also two syllables) for each (one syllable for each character). For example, the English translation of one Chinese nonsense sentence is "One appreciation could retire his ocean" (the keywords are underlined). Note that the sentence structure cannot provide any contextual support for recognizing the keywords. The development of the Chinese nonsense sentences is described by Yang *et al*. (2007).

Target speech was spoken by a young-female talker (talker A) with a median and constant rate. There were three types of maskers used in experiment 1: Steady-state speech-spectrum noise, two-Chinese-talker speech, and two-English-talker speech. All stimuli were recorded digitally onto computer disks, sampled at 16 kHz and saved as 16-bit PCM wave files. Acoustic signals were generated using the 24-bit Creative Sound Blaster PCI128 with a built-in anti-aliasing filter (Creative Technology, Ltd., Singapore) processed by a computer and presented diotically to the participant through headphones (Sennheiser HD 600, Dublin, Ireland).

The noise masker was a stream of steady-state speech-spectrum noise (Yang *et al.*, 2007). The Chinese-speech masker was a 47-s loop of digitally-combined continuous recordings for Chinese nonsense sentences (whose keywords did not appear in target sentences) spoken by two Chinese young-female talkers (talkers B and C). Each of the two masking talkers spoke different sentences and the sound-pressure levels (SPLs) were the same across their speech sounds within a testing session. The English-speech masker was a 47-s loop of digitally-combined continuous recordings for English nonsense sentences spoken by two native North American young-female talkers (talker D and E).

Calibration of the sound levels of the headphone was carried out with the Larson Davis Audiometer Calibration and Electroacoustic Testing System (AUDit and System 824, Larson Davis, USA) with "A" weighting. The sound-pressure level for target speech was fixed at 60 dB SPL.

3. Ideal binary masking segregation

The procedures of the IBM manipulation used in this study were the same as used by Wang *et al.* (2009). Each of the stimuli (target-speech, maskers, and target/masker mixture) was processed through a bank of 64-channel gammatone filters and with center frequencies ranging from 55 to 7743 Hz on an approximately logarithmic scale. The filter response was then windowed into 20-ms frames with a 10-ms overlap to produce a matrix of T-F units. The LC was set at −6 dB. Figure 1 illustrates the spectrograms of the following stimuli related to this experiment: A target-speech stimulus (panel A), the masking noise (panel B), the target/masker mixture without the IBM manipulation (panel C), and the IBM-treated target/masker mixture (panel D).

4. Design and procedures

There were 24 (three masker types: noise, Chinese speech, English speech; two IBM conditions: without IBM, with IBM; and four SNRs at each IBM condition) testing conditions for each participant, and 18 target sentences were used in each testing condition. The presentation order for the six masker/IBM combinations was partially counterbalanced across 12 participants using a Latin square order, and the presentation order of the four SNRs at each of the combinations was arranged randomly.

For the target/masker mixture without the IBM manipulation, the level of each of the maskers was set at −12, −8, −4, or 0 dB. For the mixture processed by the IBM, the four SNRs were −16, −8, −4, and 0 dB when the masker was noise, and −24, −20, −16, and −12 dB when the masker was either Chinese speech or English speech. The selection of these SNRs was based on our pilot experiments.

In a sound attenuated chamber (EMI Shielded Audiometric Examination Acoustic Suite), the participant initiated a trial by pressing a key on the computer keyboard. In the target/masker mixture, the masker was first presented and about 1 s later target speech was presented. Both target speech and the masker were terminated at the same time. The participant was instructed to loudly repeat the whole target sentence immediately after all the stimuli ended. The participants' performance was scored on the numbers of correctly identified keywords in target sentences by the experimenters who sat outside the chamber using headphones.

To ensure that all the participants fully understood and correctly followed the experimental instructions, there was one training session before formal testing. Sentences used in training were different from those used in formal testing.

B. Results

A logistic psychometric function

$$y = 1/[1 + e^{-\sigma(x-\mu)}]$$

was fit to each individual participant's data, using the Levenberg–Marquardt method (Wolfram, 1991), where y is the probability of correct identification of keywords, x is the SNR corresponding to y, μ is the SNR corresponding to 50% correct identification (the threshold), and σ determines the slope of the psychometric function.

Figure 2 presents the group-mean percent-correct keyword intelligibility as a function of the SNR along with the group-mean best-fitting psychometric functions (curves), when the masker was noise (top panel), Chinese speech (middle panel), or English speech (bottom panel). Obviously, with the increase of the SNR, participants' keyword intelligibility monotonically increased. More importantly, the IBM manipulation shifted the psychometric functions to the left, indicating a reduction of the performance threshold and an improvement of target-speech intelligibility.

The psychometric functions in Fig. 2 were used to determine the group-mean thresholds, which are shown in Fig. 3 for each of the three masking conditions. Clearly, the threshold decreased (intelligibility of target speech was improved) following the IBM manipulation.

A 3 (masker type: noise, Chinese speech, and English speech) by 2 (IBM conditions: without IBM, with IBM) two-way within-subject analysis of variance (ANOVA) shows a significant interaction between masker type and IBM condition ($F_{2,22} = 299.546$, $p < 0.001$). Multiple t-tests confirm that the IBM manipulation caused a significant improvement in target-speech intelligibility (making the threshold μ significantly lower) when the masker was noise ($T_{11} = 6.895$, $p < 0.001$), Chinese speech ($T_{11} = 14.879$, $p < 0.001$), or English speech ($T_{11} = 9.320$, $p < 0.001$). The IBM-induced threshold shift was 7.6, 14.9, and 9.3 dB, when the masker was noise, Chinese speech, and English speech, respectively.

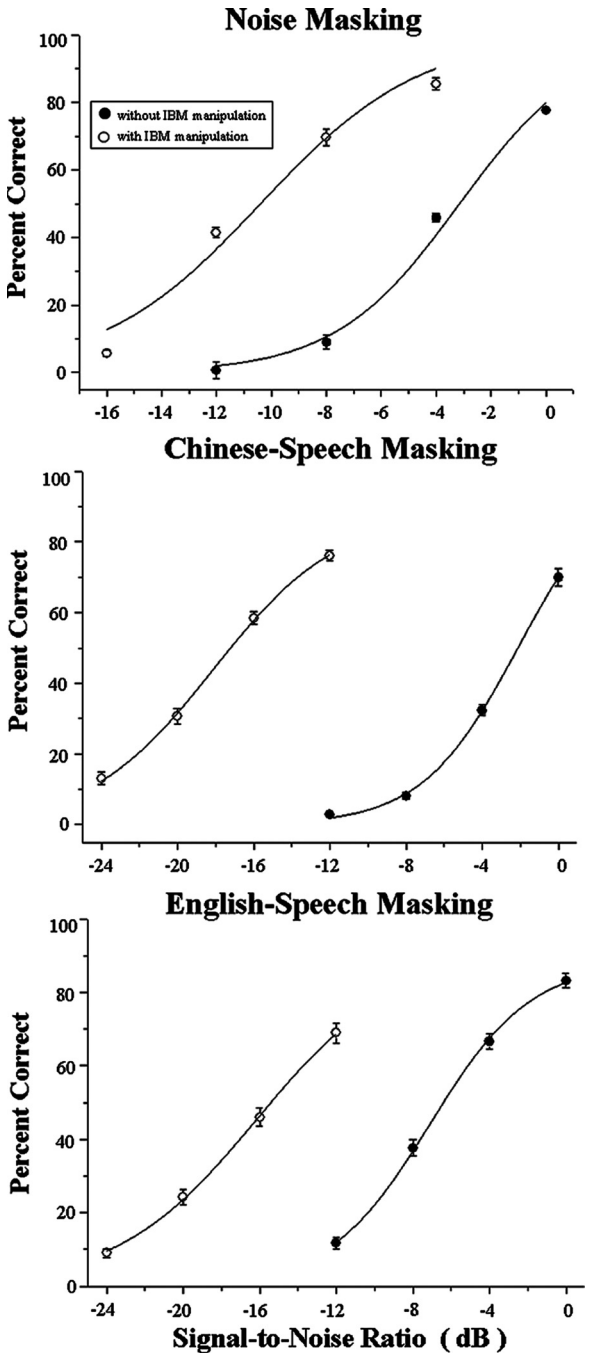

FIG. 2. Group-mean percent-correct keyword recognition in experiment 1 as a function of the SNR along with the group-mean best-fitting psychometric functions (curves), when the masker was steady-state speech-spectrum noise (top panel), two-Chinese-talker speech (middle panel), or two-English-talker speech (bottom panel). Solid circles represent the conditions without the IBM manipulation; open circles represent the conditions under which the target/masker mixture was processed by the IBM manipulation. In this and following figures, the error bars represent the standard errors of the mean.

III. EXPERIMENT 2: EFFECTS OF ADDING THE NOISE BACKGROUND WHEN THE MASKER WAS NOISE

A. Methods

1. Participants

Twelve young university students (19–25 yr old, seven females and five males) participated in experiment 2 but not in other experiments of this study.

2. Stimuli

Speech stimuli were also Chinese "nonsense" sentences as used in experiment 1, but target speech was recited by an artificially synthesized young-female voice (voice O, see below for details).

The speech-synthesis method based on the hidden Markov model (HMM) has been successfully used for achieving the text-to-speech (TTS) transformation (i.e., converting written text into audible speech) (Masuko *et al.*, 1996; Yoshimura *et al.*, 1999). Free software HTS is also available (Zen *et al.*, 2007a). In this study, acoustic signals of target speech were generated by the HMM-based speech-synthesis system. At first, a Chinese corpus including 6000 sentences with a news-broadcast style, which was both phonetically and prosodically rich, was downloaded from the website (see King and Karaiskos, 2009) and sampled for model training with a sampling rate of 16 kHz. Using the method developed by Zen *et al.* (2007b), some critical parameters of speech features (including the mel-cepstrum, log F_0, and band aperiodicity measures) were extracted and the five-state left-to-right HMM structure (with no skip) was adopted. Then a five-dimensional multivariate Gaussian distribution was incorporated to model the distribution of the state duration probability. The context-dependent HMMs for each stream were constructed using the decision-tree-based context-clustering method with the minimum-description length (MDL) criterion developed by Shinoda and Watanabe (1997). At the synthesis stage, the speech-parameter sequence for each sentence stimulus was generated from the corresponding HMMs under the dynamic feature constraints. Then using the method developed by Fukada *et al.* (1992), a speech waveform was synthesized by the well-known algorithm of the Mel Log Spectrum Approximation Filter with the generated parameters. Finally, an acoustic model was established by a training procedure using the speech corpus

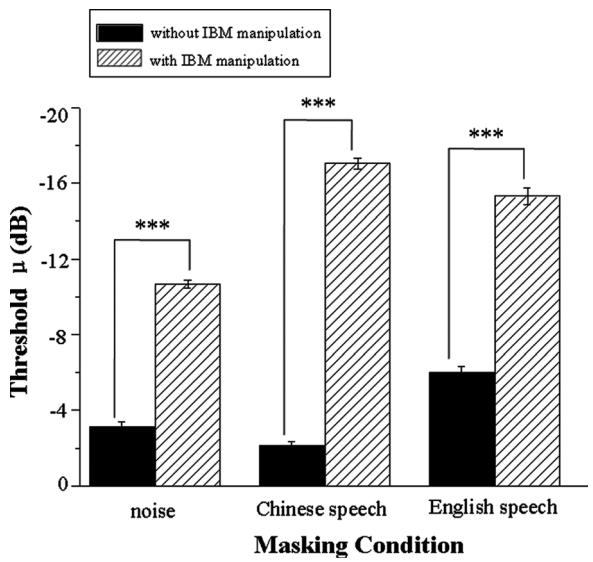

FIG. 3. Group-mean thresholds for recognizing target-speech keywords in experiment 1. Black histograms indicate the thresholds under conditions without the IBM manipulation, striated histograms indicate the thresholds under conditions with the IBM manipulation.

with the voice of a selected female talker (talker O). In total 1700 Chinese nonsense sentences with the target voice of talker O were prepared for this study.

All acoustic signals were generated using the 24-bit Creative Sound Blaster PCI128 (which had a built-in anti-aliasing filter) processed by a computer (Pentium IV processor, Intel Corporation, Santa Clara, CA) and presented diotically to the participant through headphones (Sennheiser HD 600). Calibration of the SPL was carried out with the Larson Davis Audiometer Calibration and Electroacoustic Testing System with "A" weighting. The sound level of target speech was fixed at 60 dB SPL.

Both the noise masker and the noise background were streams of steady-state speech-spectrum noise (Yang *et al.*, 2007). Relative to target-speech stimuli, the SPL of the masking noise was adjusted to produce four SNRs: −6, −10, −14, and −18 dB. Relative to the pre-IBM-treated (the original unprocessed) target speech, the SPL of the background noise, which was added to all the T-F units, was adjusted to produce eight SNR conditions: −4, 0, 4, 8, 12, 16, 20, and 24 dB. There was also one condition without the background noise.

3. Ideal binary masking segregation

The procedures of the IBM manipulation used in experiments 2 and 3 were the same as used in experiment 1. Figure 1 compares the spectrogram of the IBM-treated target/masker mixture without adding the background noise (panel D) and the spectrogram of the IBM-treated target/masker mixture after the background noise was added (panel E).

4. Design and procedures

There were 36 (four SNRs for the masking noise, nine SNR conditions for the background noise) testing conditions for a participant, and 18 target sentences were used in each condition. The presentation order for the nine SNR conditions for the background noise was partially counterbalanced across 12 participants, and the presentation order for the four SNRs for the masking noise was arranged randomly at each SNR condition for the background noise.

In the sound attenuated chamber as used in experiment 1, the participant initiated a testing trial by pressing a key on the computer keyboard. A mixture of target speech and masking noise, which were processed by the IBM, was presented to the participant with headphones when a background noise was either introduced or not.

The performance scoring and pre-testing training in experiments 2 and 3 were the same as used in experiment 1.

B. Results

Figure 4 presents the group-mean percent-correct keyword intelligibility as a function of the SNR for the masking noise, along with the group-mean best-fitting psychometric functions (curves), at each of the nine SNR conditions for the background noise. Obviously, with the increase of the SNR for the masking noise from −18 to −6 dB, participants' keyword intelligibility monotonically increased.

The psychometric functions in Fig. 4 were also used to determine the group-mean thresholds. Group-mean thresholds for the nine SNR conditions for the background noise are shown in Fig. 5. The dash line represents the threshold for the condition when the background noise was not provided. A one-way within-subject ANOVA shows that the thresholds were significantly different across the SNR conditions for the background noise ($F_{8,88} = 9.795$, $p < 0.001$). Multiple t-tests show that relative to the condition without the background noise, adding the background noise significantly made the threshold μ lower when the SNR for the background noise was 8 dB ($t_{11} = 5.615$, $p < 0.001$), 12 dB ($t_{11} = 6.077$, $p < 0.001$), or 16 dB ($t_{11} = 6.948$, $p < 0.001$) (the α was adjusted to 0.006).

The results of experiment 2 indicate that after the mixture of target speech and noise masker was manipulated by the IBM, adding a background noise with the SNR from 8 to 16 dB improved intelligibility of target speech in the target/masker mixture. For example, when the SNR for the background noise was 12 dB, the improvement of target intelligibility was 1.5 dB.

IV. EXPERIMENT 3: EFFECTS OF ADDING THE BACKGROUD NOISE WHEN THE MASKER WAS SPEECH

A. Methods

1. Paticipants

Twelve young university students (19–27 yr old, eight females, and four males) participated in experiment 3.

2. Stimuli

Target speech sentences used in experiment 3 were also Chinese nonsense sentences as used in experiments 1 and 2, and recited by the artificially synthesized young-female voice (voice O) as used in experiment 2. However, masking stimuli used in experiment 3 were two types of speech maskers. The first one was a 47-s loop of recordings for Chinese nonsense sentences (whose keywords did not appear in target sentences) recited by one young-female talker (talker B). The second type was a 47-s loop of digitally-combined continuous recordings for Chinese nonsense sentences (whose keywords also did not appear in target sentences) spoken by two different young-female talkers (talkers B and C) (Yang *et al.*, 2007).

The IBM processing was conducted as experiments 1 and 2. The SNR for the one-talker masker was set as −18, −22, −26, or −30 dB; the SNR for the two-talker was set as −14, −18, −22, or −26 dB. Relative to the pre-IBM-treated (the original unprocessed) target speech, the SPL of the background noise, which was added to all the T-F units, was adjusted to produce eight SNR conditions: −4, 0, 4, 8, 12, 16, 20, and 24 dB. There was also one condition without the background noise.

3. Design and procedures

The design and procedures used in experiment 3 were the same as used in experiment 2, except that there were 72 (two masker types, four SNRs for the maskers, nine SNR

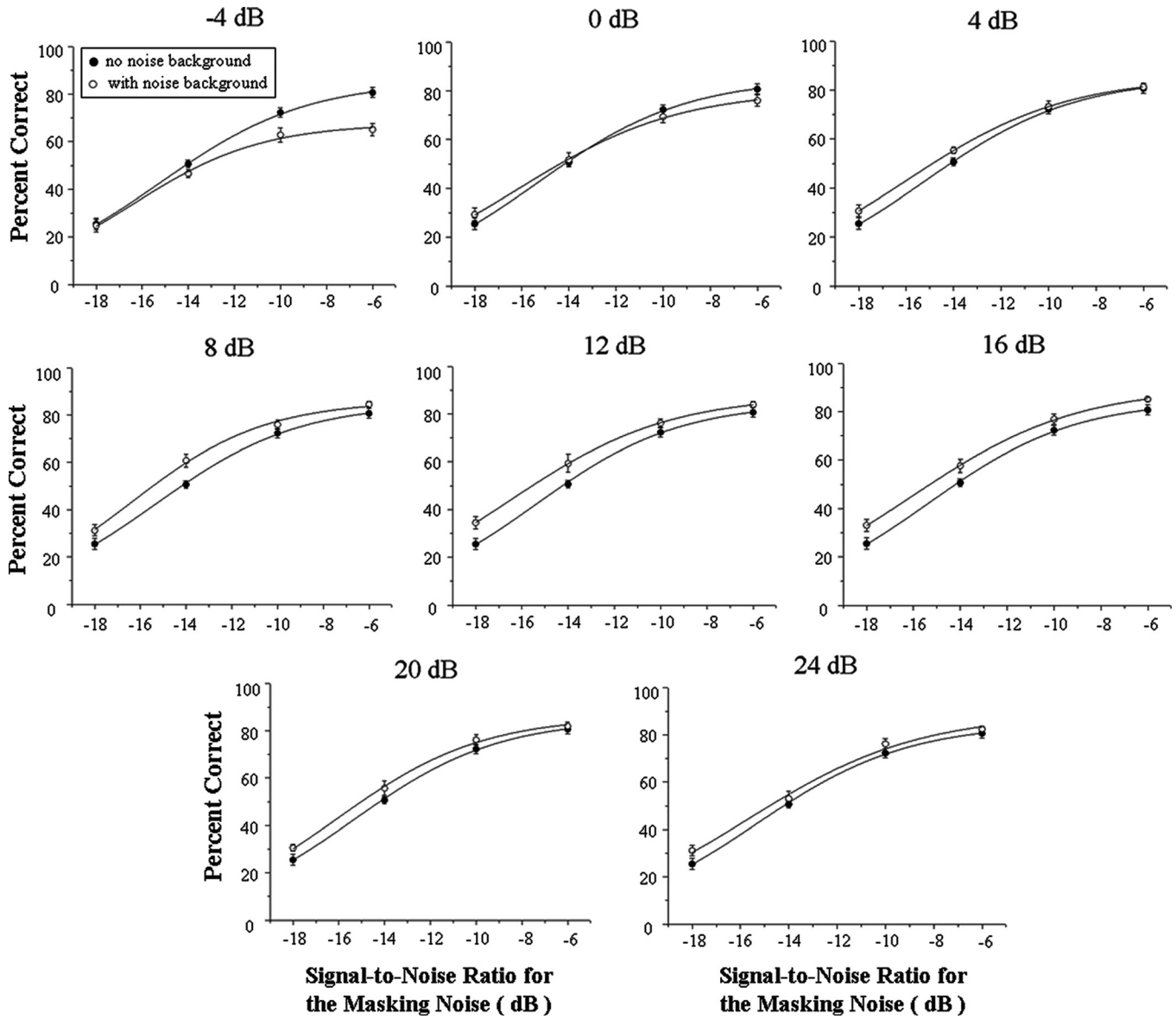

FIG. 4. Group-mean percent-correct keyword recognition as a function of the SNR for the steady-state masking noise in experiment 2, along with the group-mean best-fitting psychometric functions (curves) at each of the SNR conditions for the background noise. The performance under the condition without adding the background noise serves as the baseline in each of the panels. Solid circles are associated with the baseline condition without adding the background noise; open circles are associated with the conditions with the background noise being added.

conditions for the background noise) testing conditions for each participant.

B. Results

Figures 6 and 7 present the group-mean percent-correct keyword intelligibility as a function of the SNR for the masking noise along with the group-mean best-fitting psychometric functions (curves) when the masker was one-talker speech and two-talker speech, respectively, for each of the background-noise conditions. In each panel, the baseline was the performance when the background noise was not provided. With the increase of the SNR for the masking noise, participants' keyword intelligibility monotonically increased.

Figure 8 shows group-mean thresholds for the nine SNR conditions for the background noise when the masker in the target/masker mixture was one-talker speech (upper panel) or two-talker speech (lower panel). The dash lines represent the thresholds for the conditions when the background noise was not provided. A 2 (masker types) by 9 (SNR conditions for the background noise) two-way ANOVA shows that the interaction between the two factors on the threshold was significant ($F_{8,88} = 122.464$, $p < 0.001$).

When the masker was one-talker speech, multiple t-tests show that relative to the condition without the background noise, adding the background noise significantly made the threshold μ lower when the SNR for the background noise was 4 dB ($t_{11} = 4.585$, $p = 0.001$), 8 dB ($t_{11} = 4.438$, $p = 0.001$), 12 dB ($t_{11} = 6.436$, $p < 0.001$), 16 dB ($t_{11} = 5.639$, $p < 0.001$), 20 dB ($t_{11} = 4.268$, $p = 0.001$), 24 dB ($t_{11} = 6.257$, $p < 0.001$), but not 0 dB ($t_{11} = 2.936$, $p = 0.014$) (the α was adjusted to 0.006). Also, adding the background noise significantly increased the threshold when the SNR was -4 dB ($t_{11} = 7.919$, $p < 0.001$).

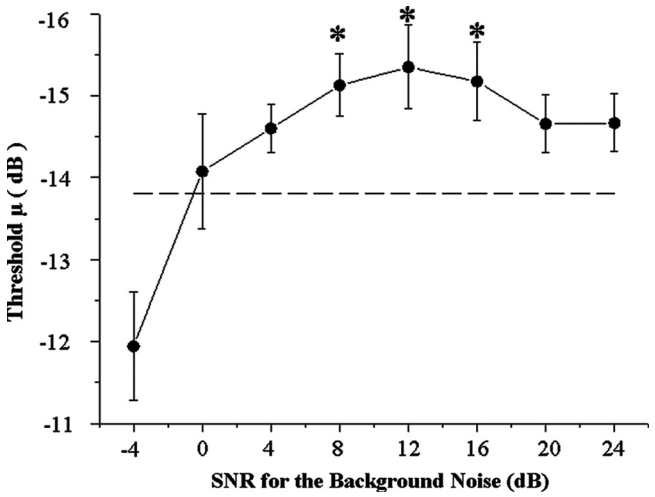

FIG. 5. Group-mean thresholds for recognizing target-speech keywords in experiment 2 for the eight different levels of the background noise. The dash line shows the group-mean thresholds for recognition of target speech in the IBM-treated target/masker mixture without adding the background noise.

When the masker was two-talker speech, multiple *t*-tests show that relative to the condition without the background noise, adding the background noise significantly made the threshold μ lower when the SNR for the background noise was 8 dB ($t_{11}=5.881$, $p<0.001$), 12 dB ($t_{11}=7.039$, $p<0.001$), 16 dB ($t_{11}=4.541$, $p=0.001$), or 20 dB ($t_{11}=7.212$, $p<0.001$). Also, adding the background noise significantly increased the threshold when the SNR was –4 dB ($t_{11}=11.119$, $p<0.001$).

The results of experiment 3 indicate that after the mixture of target speech and one-talker-speech masker was manipulated by the IBM, adding a background noise with the SNR from 4 to 24 dB improved intelligibility of target speech in the target/masker mixture. For example, when the SNR for the background noise was 16 dB, the improvement of target intelligibility was 3.3 dB. However, adding a background noise with the SNR of −4 dB had a significant masking effect on intelligibility of target speech.

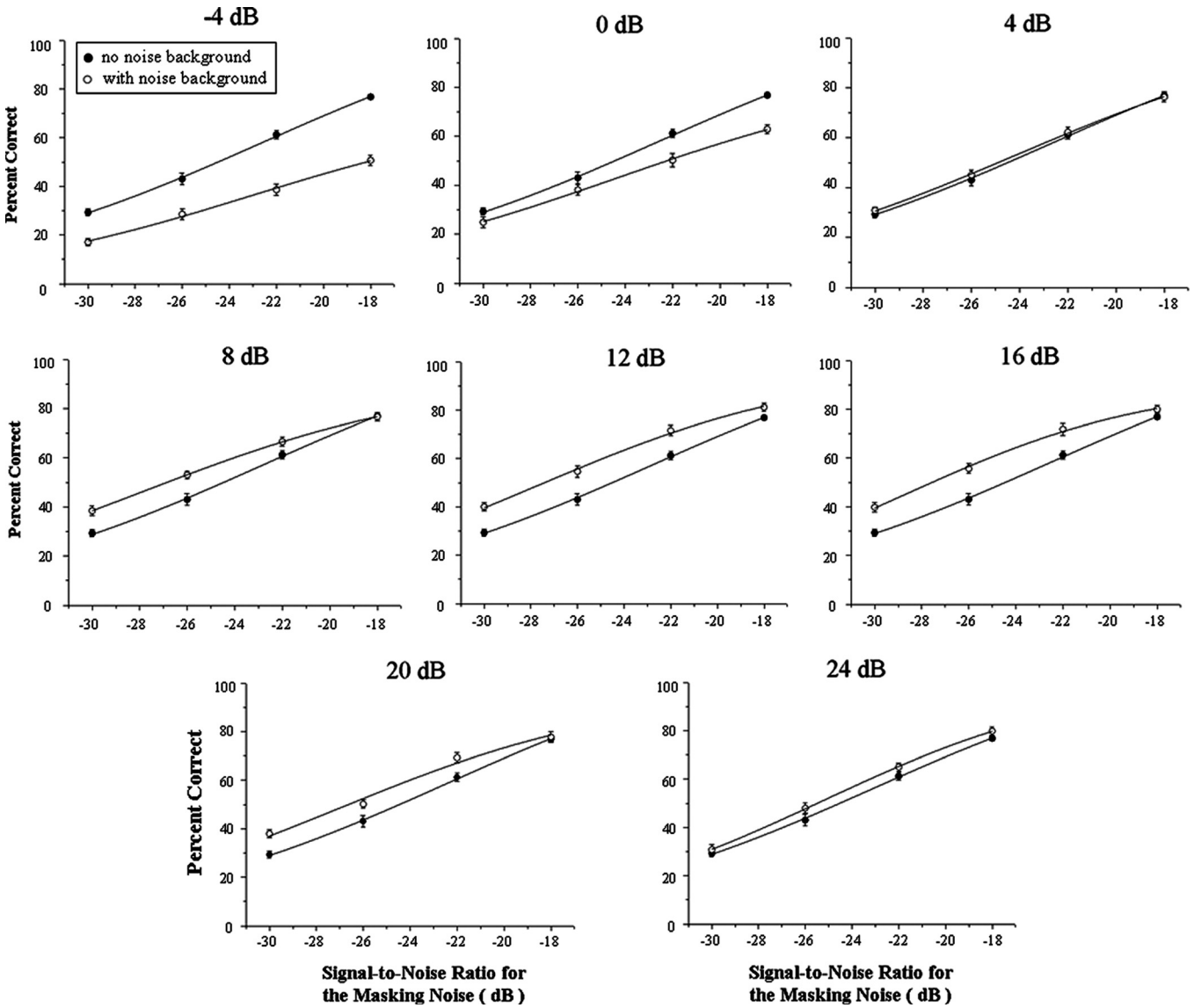

FIG. 6. Group-mean percent-correct keyword recognition as a function of the SNR for the one-talker-speech masker in experiment 3, along with the group-mean best-fitting psychometric functions (curves) at each of the nine SNR conditions for the background noise. The performance under the condition without adding the background noise serves as the baseline in each of the panels. Solid circles are associated with the baseline condition without adding the background noise; open circles are associated with the conditions with the background noise being added.

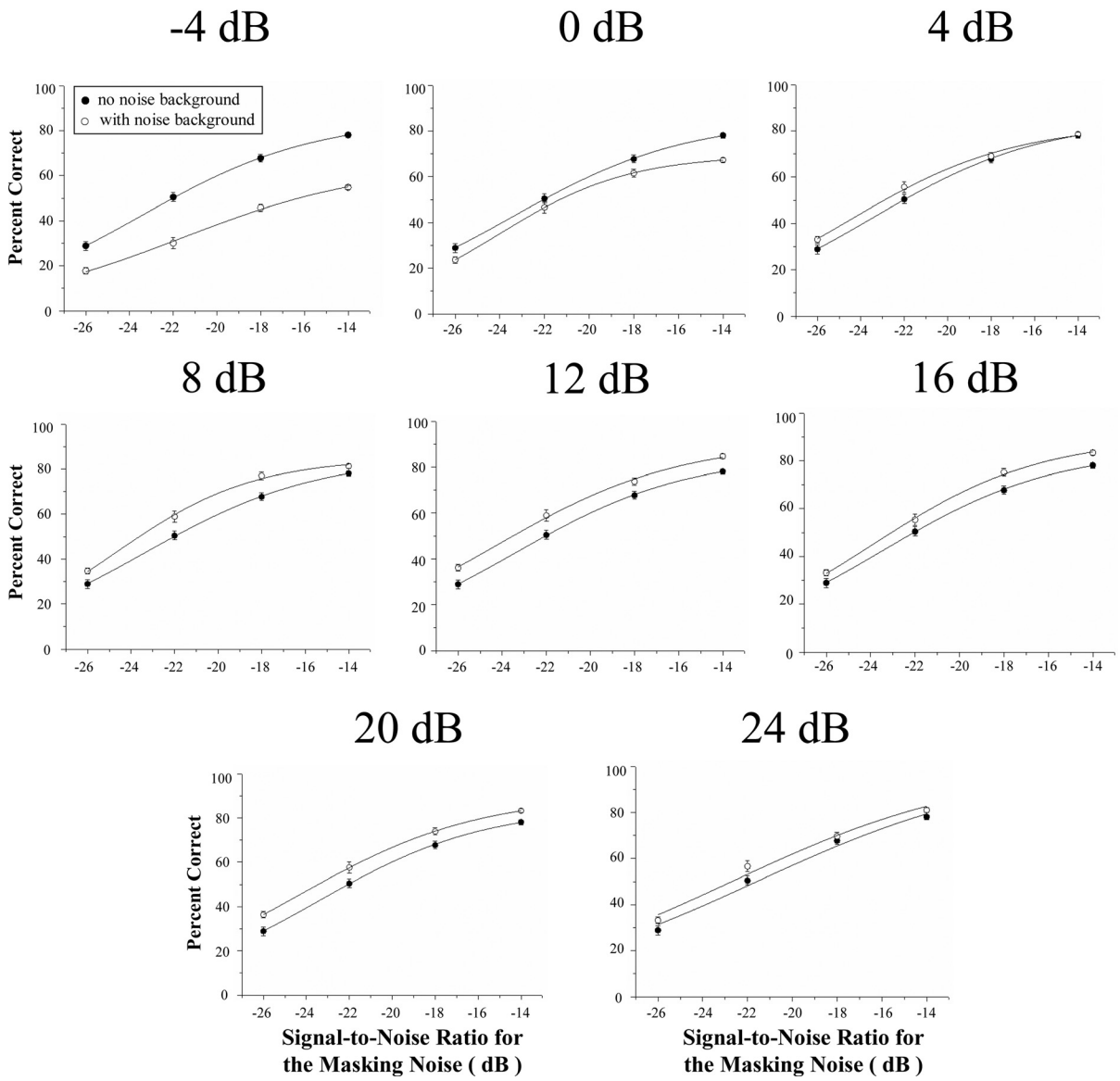

FIG. 7. Group-mean percent-correct keyword recognition as a function of the SNR for the two-talker-speech masker in experiment 3, along with the group-mean best-fitting psychometric functions (curves) at each of the nine SNR conditions for the background noise. The performance under the condition without adding the background noise serves as the baseline in each of the panels. Solid circles are associated with the baseline condition without adding the background noise; open circles are associated with the conditions with the background noise being added.

When the masker was two-talker speech, adding a background noise with the SNR from 4 to 24 dB improved intelligibility of target speech in the target/masker mixture. For example, when the SNR for the background noise was 12 dB, the improvement of target intelligibility was 1.9 dB. Also, adding a background noise with the SNR of −4 dB had a significant masking effect on intelligibility of target speech.

V. GENERAL DISCUSSION

The results of experiment 1 of this study show that after introducing the IBM manipulation of the target/masker mixture, intelligibility of target speech in the mixture was remarkably improved when the masker was noise, Chinese (native) speech, or English (foreign) speech. More specifically, the IBM manipulation led to the largest reduction of speech-intelligibility threshold (14.9 dB) when the masker was native (Chinese) speech, followed by that of 9.3 dB when the masker was foreign (English) speech. The smallest improvement (7.6 dB) occurred when the masker was speech-spectrum noise. The results are in agreement with previous studies showing that the IBM manipulation improves speech intelligibility (Anzalone *et al.*, 2006; Brungart *et al.*, 2006, 2009; Li and Loizou, 2007, 2008, Wang *et al.*, 2008, 2009; Kjems *et al.*, 2009). Moreover, since noise masking and speech masking are different in mechanisms (for the concepts of energetic masking and informational masking, see Arbogast *et al.*, 2002; Brungart, 2001; Durlach *et al.*, 2003; Freyman *et al.*, 1999; Kidd *et al.*, 1994; Li *et al.*, 2004; Shinn-Cunningham *et al.*, 2005; Summers and Molis, 2004; Wu *et al.*, 2005), the results of experiment 1 of this study support the notion that the IBM manipulation mainly reduces informational masking of target speech (Brungart *et al.*, 2006).

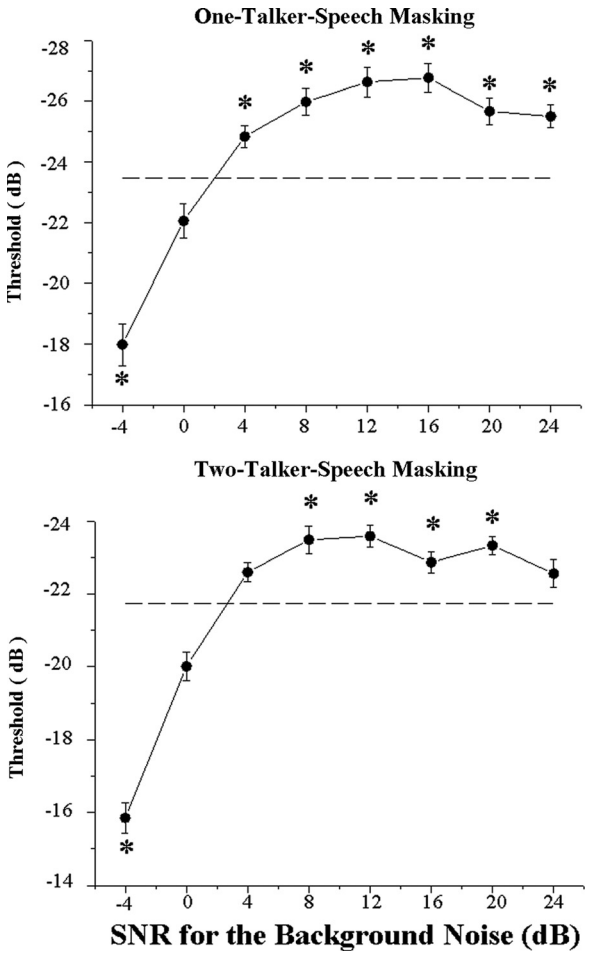

FIG. 8. Group-mean thresholds for recognizing target-speech keywords in experiment 3 for the eight different levels of the background noise when the masker was one-talker speech (top panel) or two-talker speech (bottom panel). The dash line in each panel shows the group-mean thresholds for recognition of target speech in the IBM-treated target/masker mixture without adding the background noise.

As mentioned in Sec. I, since units in the 2-D T-F representation of the target/masker mixture are removed by the IBM manipulation when their SNRs are below the LC, numerous temporal and spectral gaps (with the binary value of 0) are introduced into the target/masker mixture. The results of experiments 2 and 3 show that adding a background of steady-state speech-spectrum noise further improved intelligibility of target speech in the IBM-treated target/masker mixture, regardless of whether the masker was noise, one-talker speech, or two-talker speech. More specifically, the reduction of the speech-intelligibility threshold was 1.5, 3.3, and 1.9 dB when the masker was noise, one-talker speech, and two-talker speech, respectively. The results are consistent to previous reports that intelligibility of speech is improved by filling in speech gaps with un-modulated, steady-state noise (Warren, 1970; Powers and Wilcox, 1977; Bashford *et al.*, 1992).

Providing a noise background with an appropriate SNR (no less than 4 dB) does not cause a significant masking effect on target speech, but weakens the abruption of transient changes that occur in the IBM-treated target/masker mixture, leading to both the enhancement of perceived continuity of the target-speech stream (Bregman, 1990; Woods *et al.*, 1996; Srinivasan and Wang, 2005) and the facilitation of the formation of the target-speech object. Consequently, intelligibility of target speech is improved.

It has been known that a 1-dB reduction in the speech-reception threshold is equal to a 7%–19% increase in the percent correct measurement (Moore, 2007). Since adding a steady-state noise background is feasible, the speech-enhancing method established by this study may be useful in hearing-improving applications, such as hearing aids and cochlear implants.

VI. CONCLUSIONS

The present study confirms the enhancing effect of the IBM manipulation on intelligibility of target speech at various masking conditions, supporting the view that the speech-intelligibility improvement is largely due to a reduction of informational masking. More importantly, the present study for the first time shows that providing a steady-state speech-spectrum noise background to the IBM-treated target-speech mixture significantly improves intelligibility of target speech, probably due to an enhancement of perceived continuity of the target-speech stream. Thus, a critical issue in future studies is whether the combination of the IBM manipulation and background-noise addition is useful for improving hearing-aid and cochlear-implant devices designed for listeners with impaired hearing, particularly under noisy, multiple-voicing conditions.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (90920302; 60811140086), the "973" National Basic Research Program of China (2009CB825404), the Chinese Ministry of Education (20090001110050), the Speech and Hearing Research Center at Peking University, and the Key Laboratory on Machine Perception (Ministry of Education).

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

Arbogast, T. L., Mason, C. R., and Kidd, G. (**2002**). "The effect of spatial separation on informational and energetic masking of speech," J. Acoust. Soc. Am. **112**, 2086–2098.

Bashford, J. A., Jr., Riener, K. R., and Warren, R. M. (**1992**). "Increasing the intelligibility of speech through multiple phonemic restorations," Percept. Psychophys. **51**, 211–217.

Baskent, D., Eiler, C., and Edwards, B. (**2009**). "Effects of envelope discontinuities on perceptual restoration of amplitude-compressed speech," J. Acoust. Soc. Am. **125**, 3995–4005.

Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT, Cambridge, MA), pp. 345–349.

Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.

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

Brungart, D. S., 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.

Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., and Kidd, G. (**2003**). "Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity," J. Acoust. Soc. Am. **114**, 368–379.

Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (**1999**). "The role of perceived spatial separation in the unmasking of speech," J. Acoust. Soc. Am. **106**, 3578–3588.

Fukada, T., Tokuda, K., Kobayashi, T., and Imai, S. (**1992**). "An adaptive algorithm for mel-cepstral analysis of speech," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP-92,* San Francisco, CA, pp. 137–140.

Helfer, K. S. (**1997**). "Auditory and auditory–visual perception of clear and conversational speech," J. Speech Lang. Hear. Res. **40**, 432–443.

Kidd, G., Mason, C. R., Deliwala, P. S., Woods, W. S., and Colburn, H. S. (**1994**). "Reducing informational masking by sound segregation," J. Acoust. Soc. Am. **95**, 3475–3480.

King, S., and Karaiskos, V. (**2009**). "The Blizzard Challenge 2009," in *Proceedings of the Blizzard Challenge Workshop*, Edinburgh, UK.

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.

Li, L., Daneman, M., Qi, J. G., and Schneider, B. A. (**2004**). "Does the information content of an irrelevant source differentially affect speech recognition in younger and older adults?" J. Exp. Psychol. Hum. Percept. Perform. **30**, 1077–1091.

Li, N., and Loizou, P. C. (**2007**)**.** "Factors influencing glimpsing of speech in noise," J. Acoust. Soc. Am. **122**, 1165–1172.

Li, N., and Loizou, P. C. (**2008**). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," J. Acoust. Soc. Am. **123**, 1673–1682.

Masuko, T., Tokuda, K., Kobayashi, T., and Imai, S. (**1996**). "Speech synthesis from HMMs using dynamic features," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP-96*, Atlanta, GA, pp.389–392.

Moore, B. C. J. (**2007**). *Cochlear Hearing Loss*, 2nd ed. (Wiley, Chichester, UK), pp. 201–232.

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (**1988**). "An Efficient Auditory Filterbank Based on The Gammatone Function," Report No. 2341 (Applied Psychology Unit, Cambridge, UK), pp. 1–33.

Powers, G. L., and Wilcox, J. C. (**1977**). "Intelligibility of temporally interrupted speech with and without intervening noise," J. Acoust. Soc. Am. **61**, 195–199.

Shinn-Cunningham, B. G., Ihlefeld, A., and Satyavarta, L. E. (**2005**). "Bottom-up and top-down influences on spatial unmasking," Acust. Acta Acust. **91**, 967–979.

Shinn-Cunningham, B. G., and Wang, D. (**2008**). "Influences of auditory object formation on phonemic restoration," J. Acoust. Soc. Am. **123**, 295–301.

Shinoda, K., and Watanabe, T. (**1997**). "Acoustic modeling based on the MDL criterion for speech recognition," in *Proceedings of EUROSPEECH*, Rhodes, Greece, pp. 99–102.

Srinivasan, S., and Wang, D. (**2005**). "A schema-based model for phonemic restoration," Speech Commun. **45**, 63–87.

Summers, V., and Molis, M. R. (**2004**). "Speech recognition in fluctuating and continuous maskers: Effects of hearing loss and presentation level," J. Speech Lang. Hear. Res. **47**, 245–256.

Wang, D. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Norwell, MA), pp. 181–197.

Wang, D. L., and Brown, G. J. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley, NY/IEEE, Hoboken, NJ), pp. 14–25.

Wang, D. L., Kjems, U., Pedersen, M. S., Boldt, J. B., and Lunner, T. (**2008**). "Speech perception of noise with binary gains," J. Acoust. Soc. Am. **124**, 2303–2307.

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.

Warren, R. (**1970**). "Perceptual restoration of missing speech sounds," Science **167**, 392–393.

Wolfram, S. (**1991**). *Mathematica: A System for Doing Mathematics by Computer* (Addison-Welsey, New York), pp. 672–682.

Woods, W. S., Hansen, M., Wittkop, T., and Kollmeier, B. (**1996**). "A scene analyzer for speech processing," in *Proceedings of the ESCA Workshop on the Auditory Basis of Speech Perception, ABSP-1996* (International Speech Communication Association, Keele, UK), pp. 232–235.

Wu, X.-H., Wang, C., Chen, J., Qu, H.-W., Li, W.-R., Wu, Y.-H., Schneider, B. A., and Li, L. (**2005**). "The effect of perceived spatial separation on informational masking of Chinese speech," Hear. Res. **199**, 1–10.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (**1999**). "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proceedings of EUROSPEECH 1999*, pp. 2347–2350.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., and Tokuda, K. (**2007a**). "The HMM-based speech synthesis system (HTS) version 2.0," in *Proceedings of the 6th ISCA Workshop Speech Synthesis (SSW-6)*, August, Bonn, Germany.

Zen, H., Toda, T., Nakamura, M., and Tokuda, K. (**2007b**). "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," IEICE Trans. Inf. Syst. E90-D (1), 325–333.