Informational masking of speech produced by speech-like sounds without linguistic content

Jing Chen, Huahui Li, Liang Li, and Xihong Wu[a)]
Department of Machine Intelligence, Speech and Hearing Research Center, and Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing 100871, People's Republic of China

Brian C. J. Moore
Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, England

(Received 31 December 2010; revised 29 January 2012; accepted 1 February 2012)

This study investigated whether speech-like maskers without linguistic content produce informational masking of speech. The target stimuli were nonsense Chinese Mandarin sentences. In experiment I, the masker contained harmonics the fundamental frequency (F0) of which was sinusoidally modulated and the mean F0 of which was varied. The magnitude of informational masking was evaluated by measuring the change in intelligibility (releasing effect) produced by inducing a perceived spatial separation of the target speech and masker via the precedence effect. The releasing effect was small and was only clear when the target and masker had the same mean F0, suggesting that informational masking was small. Performance with the harmonic maskers was better than with a steady speech-shaped noise (SSN) masker. In experiments II and III, the maskers were speech-like synthesized signals, alternating between segments with harmonic structure and segments composed of SSN. Performance was much worse than for experiment I, and worse than when an SSN masker was used, suggesting that substantial informational masking occurred. The similarity of the F0 contours of the target and masker had little effect. The informational masking effect was not influenced by whether or not the noise-like segments of the masker were synchronous with the unvoiced segments of the target speech. © *2012 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.3688510]

PACS number(s): 43.66.Dc, 43.71.Bp, 43.71.An [MAA] Pages: 2914–2926

I. INTRODUCTION

Listeners often find it difficult to understand speech when it is presented with background sounds, such as noise or interfering talkers. Two main factors are thought to contribute to this difficulty: (1) energetic masking, which occurs when peripheral neural activity elicited by a signal is overwhelmed by that elicited by the masker, leading to a degraded or noisy neural representation of the signal, or (2) informational masking, which is also called "non-energetic masking," and is conceptualized as anything that reduces intelligibility once energetic masking has been accounted for, including effects such as difficulty in determining how to assign acoustic elements in the mixture to the target and masker (Watson, 1987; Freyman *et al*., 1999; Freyman *et al*., 2001, 2004; Brungart *et al*., 2001; Li *et al*., 2004; Wu *et al*., 2005; Mattys *et al*., 2009). The effect of energetic masking on speech intelligibility has been well documented and can be evaluated using models such as the Articulation Index (French and Steinberg, 1947; Fletcher and Galt, 1950) and the Speech Intelligibility Index (ANSI, 1997). The effects of informational masking on speech intelligibility are more complicated, involving multiple levels of processing, and are rarely described by current computational models (Houtgast and Steeneken, 1985; ANSI, 1997; Elhilali *et al*., 2003; Rhebergen *et al*., 2006).

Several researchers have studied the effects of informational masking on speech perception by manipulating the stimulus characteristics. Brungart *et al*. (2001) found that the recognition of speech in multitalker environments generally worsened when the target and masking talkers had similar voice characteristics: The target was more intelligible when the masker and the target were spoken by different-gender talkers than when they were spoken by same-gender talkers or the same talker. The number of masking talkers was also manipulated in several studies (Freyman *et al*., 2004; Simpson and Cooke, 2005; Wu *et al*., 2007). The results showed that speech recognition was a non-monotonic function of the number of masking talkers. The effects of informational masking can be reduced by introduction of a difference in perceived location of the target and masker via the precedence effect (Freyman *et al*., 2001; Li *et al*., 2004; Wu *et al*., 2005; Huang *et al*., 2008) (see following text for more details of this method). This effect is called here the "releasing effect." When sentences were used as test materials, the releasing effect was largest with two competing talkers for both English and Chinese, indicating that two-talker speech produced the most informational masking (Freyman *et al*., 2004; Rakerd *et al*., 2006; Wu *et al*., 2007). Also, a native-language speech masker produced more informational masking than a non-native speech masker (Freyman *et al*., 2001; Wu *et al*., 2011). Similarly, time-reversed speech produced

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

 0001-4966/2012/131(4)/2914/13/$30.00

less informational masking than normal speech, but performance with a time-reversed native speech masker was poorer than for a non-native speech masker, perhaps due to increased forward masking for the former (Rhebergen *et al*., 2005).

It is generally assumed that two kinds of processes play a role in speech perception: signal-driven processes and knowledge-driven processes (Bregman, 1990). The relative importance of signal-driven and knowledge-driven processes in producing informational masking and release from informational masking remains unclear. At the acoustic level, the main ways in which speech differs from steady speech-spectrum noise (SSN), which is often regarded as a purely energetic masker (see, however, Stone *et al*., 2011) are: (1) speech is highly amplitude modulated (AM), and the AM is partially correlated in different frequency regions; (2) speech includes periodic or quasi-period segments the fundamental frequency (F0) of which varies over time; and (3) speech tends to alternate between periodic segments with a harmonic structure and non-periodic segments with a noise-like structure. Freyman *et al*. (2001) studied the effects of the characteristics of AM using a masker that was SSN modulated by the single- or multi-channel envelope extracted from two-talker speech. The releasing effect of perceived spatial separation was not greater when the masker was AM noise than when it was steady SSN, indicating that the AM itself did not induce informational masking. However, to our knowledge, it has not been investigated whether a periodic sound with F0 modulation (F0M) leads to informational masking. It is known that F0 differences play a role in the perceptual separation of a target talker from a background talker (Brokx and Nooteboom, 1982; Bird and Darwin, 1998; Binns and Culling, 2007). The identification of two concurrent vowels improves with increasing F0 difference between them (Culling and Summerfield, 1995). Also, if one vowel in a mixture of vowels is modulated in F0, it becomes more prominent than the other unmodulated vowels (McAdams, 1989). Reducing the F0 variation of sentences increases the speech recognition threshold in background sounds, especially in competing speech (Binns and Culling, 2007). These results are consistent with the possibility that the characteristics of F0M can influence informational masking. In experiment I, we explored this issue by using maskers sounds with a harmonic structure and with an F0 that was either constant or changed over time in ways with a varying degree of similarity to the target speech. The maskers were never perceived as having any meaning. We assumed that under these conditions, any informational masking produced by these maskers would be caused by signal-driven processes.

As mentioned earlier, the effects of informational masking can be evaluated by introducing a perceived spatial separation between the target speech and the masker via the precedence effect. For example, when the target and masker are both presented via a loudspeaker to the listener's right and a loudspeaker to the listener's left, and the sound from the right loudspeaker leads that from the left loudspeaker by 3 ms, both the target and masker are perceived as coming from the right loudspeaker (Wallach *et al*., 1949; Zurek, 1980; Litovsky *et al*., 1999). In other words, the target and masker are perceived as being co-located. However, if the delay between the two loudspeakers is reversed for the masker only, the target is still perceived as coming from the right loudspeaker, but the masker is perceived as coming from the left loudspeaker. Thus the relative perceived locations of the target and masker can be manipulated without substantially changing sound levels or spectra at the two ears (Freyman *et al*., 1999; Li *et al*., 2004). It has been confirmed for both Chinese and English speech materials that when the masker is speech, a perceived spatial separation between the target speech and masker can lead to a 3-8 dB release from masking, but when the masker is SSN, the release from masking is only about 1 dB (Freyman *et al*., 1999; Li *et al*., 2004; Wu *et al*., 2005).

The large effect of perceived spatial separation for the speech masker but not the noise masker is thought to occur because informational masking is large for the former but not for the latter. Phonemes, syllables, and words from the masking speech may be confused with those from the target speech. Potentially, this source of informational masking can be reduced by using a speech-like synthesized masker, such as that described in the following text, which has no linguistic content. The releasing effect of perceived spatial separation would be expected to be less than when the masker is speech but more than when it is steady SSN. Any releasing effect for the speech-like but non-linguistic masker can reasonably be interpreted as reflecting informational masking produced by signal-driven processes.

In the present study, synthesized harmonic tones with no formant structure and no linguistic content were used as maskers to investigate the effect of F0 modulation on the intelligibility of the target speech. In experiment I, the paradigm of perceived spatial separation was used to assess whether the mean value of F0 and the pattern of F0M (steady or sinusoidal F0M) can influence informational masking. In experiment II, to make the masker more similar to speech, the masking harmonics were synthesized with the original or modified pitch contour of the target sentence, and bursts of SSN were inserted in the masker at times corresponding to unvoiced portions of the target sentence. The effect of similarity of the F0 contours of the target and masker was evaluated. In experiment III, the timing of the noise bursts in the masker relative to the unvoiced portions of the target speech was manipulated to assess the importance of synchrony of acoustic features in the target and masker. In all experiments, performance was compared with that obtained using an SSN masker.

II. EXPERIMENT I: SINUSOIDAL MODULATION OF F0 IN HARMONIC COMPLEXES

A. Method

1. Listeners

Sixteen university students participated, 12 female and 4 male, with a mean age of 21 yr (range: 19-24 yr). In this and all subsequent experiments reported in this paper, all of the listeners had audiometric thresholds better than 20 dB HL at all audiometric frequencies from 0.25 to 8 kHz and all

had less than a 15-dB difference in threshold between the two ears at any frequency. Their first language was Mandarin Chinese.

2. Apparatus

Listeners were seated in a chair at the center of an anechoic chamber (Beijing CA Acoustics), which was 560 cm in length, 400 cm in width, and 193 cm in height. All signals were generated at a 22050 Hz sampling rate by a 24-bit Creative Sound Blaster PCI128 (which had a built-in anti-aliasing filter) using the audio editing software COOLEDIT PRO 2.0. The analog outputs were delivered from two loudspeakers (Dynaudio Acoustics, BM6 A, each with a built-in amplifier), which were in the frontal azimuthal plane at $\pm 45^{\circ}$ azimuth. The loudspeaker height was 140 cm, which was approximately ear level for a seated listener with average body height. The distance between the loudspeaker and the center of the listener's head was 200 cm.

3. Stimuli

The target stimuli were Chinese "nonsense" sentences (Yang *et al.*, 2007). Each of the sentences has a subject, verb, and object, which are also the three key words, with two characters for each (one syllable for each character). The meaning of the sentences did not provide any contextual information to aid recognition of the key words, e.g. "Yi1zhi1 Ma3yi3 Zheng4zai4 Xuan1nao4 Zhe4ge4 Shu1bao1 (An ant is roaring this bag)," where the key words are underlined, and the digits indicate the tonal pattern. The target sentences were spoken by a young female, who was asked to keep to a medium speech rate during recording. The sentences were scaled in amplitude so that each had the same root-mean-square (RMS) value. There were 54 lists of target sentences, with 15 sentences per list.

The following equation was used to define the fundamental frequency of the masker:

$$F0(t) = F0_{mean} + \beta \times F0_{mean} sin2\pi f_m t, \quad (1)$$

where $F0(t)$ is the sinusoidally modulated F0, $F0_{mean}$ is the mean F0, f_m is the modulation frequency, and β is the modulation depth. The values of $F0_{mean}$ and β were determined by analyzing the F0 contours of the target speech. The F0 contour of each sentence was extracted using "The Snack Sound Toolkit" (Sjolander, 2006). The mean value of F0 was 252 Hz, and the modulation depth, defined as the ratio of the standard deviation to the mean of the F0 contour for a given sentence, was about 0.2. In experiment I, β was fixed at 0.2, and $F0_{mean}$ was manipulated around 252 Hz. Because there were about five syllables per second in the target speech, f_m was set to 5 Hz. In a comparison condition, called "flat," β was set to 0, giving a steady masker. The calculated F0 from Eq. (1) was used to modulate the F0 of a complex periodic sound with all harmonics of equal amplitude. To ensure that the bandwidth of the masker was identical to that of the target, the frequency of the highest harmonic was limited to 11 025 Hz. The F0M harmonic tone was then filtered by a speech-spectrum filter, which was constructed based on the amplitude spectrum of the steady SSN used by Yang *et al.* (2007). Figure 1 shows time waveforms and spectrograms of the synthesized harmonic tones for conditions flat (upper) and F0M (lower). Note that the amplitude envelopes are flat for both conditions.

The signal-to-masker ratio (SMR) was calculated based on RMS values and was fixed at -8 dB. This value was

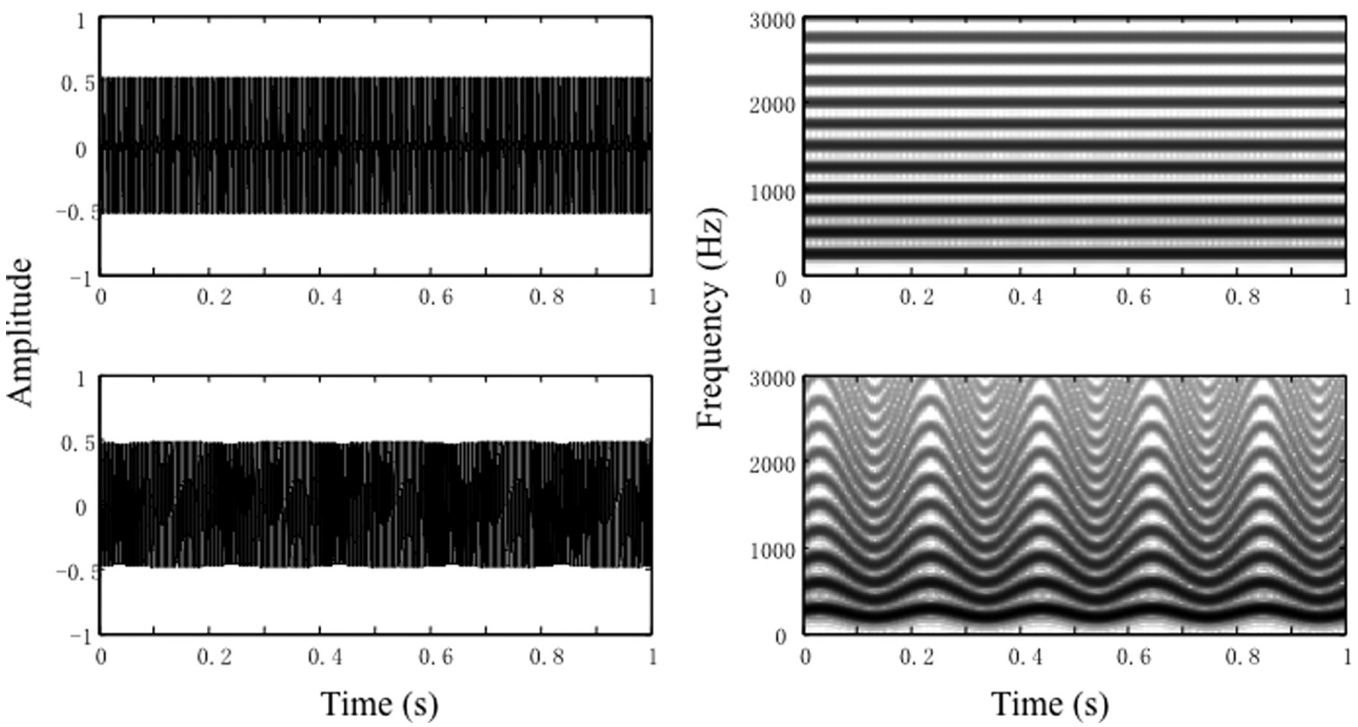

FIG. 1. Time-domain waveforms (left panels) and narrowband spectrograms (right panels) of the synthesized harmonics used in experiment I. The upper panels represent the harmonics without any F0 modulation, and the lower ones represent the sinusoidally modulated harmonics with a modulation depth of 0.2.

selected based on pilot experiments to ensure that speech intelligibility varied over a reasonable range. The target speech was presented at 62 dBA as measured using a Brüel and Kjær sound level meter (Type 2230) at the position corresponding to the center of the listener's head.

4. Design and procedure

Three factors were manipulated: $F0_{mean}$, $F0$ modulation depth, and the perceived location of the masker. Seven values of $F0_{mean}$ were used, 150, 178, 212, 252, 300, 356, and 424 Hz, which correspond to −9, −6, −3, 0, 3, 6, and 9 semitones, respectively, relative to 252 Hz. For the target, the right loudspeaker always led the left loudspeaker by 3 ms; for the masker, the right loudspeaker either led the left loudspeaker by 3 ms or lagged the left loudspeaker by 3 ms. Thus the target and the masker were perceived as being either co-located on the right side or spatially separated (target on the right and masker on the left). In total, there were 28 ($7 \times 2 \times 2$) conditions, and 15 target sentences were used for each condition. These 28 conditions were organized into four blocks: flat and co-located, flat and separated, F0M and co-located, and F0M and separated. For every group of four listeners, two of them were tested with the two blocks of flat first and then the two blocks of F0M, and the other two were tested in the opposite order. For each group of two listeners, one was tested co-located first and then separated, and the other was tested in the opposite order. In each block, the seven values of $F0_{mean}$ were presented in random order for each listener.

The listener pressed a button to start each trial. The masker and the target began and ended simultaneously. Listeners were instructed to verbally repeat the whole target sentence as well as they could immediately after the trial was completed. The experimenter, who sat outside the anechoic chamber, scored whether the key words had been identified correctly. A key word was scored as correct only if both syllables of the key word were repeated correctly.

To ensure that all the listeners fully understood and correctly followed the instructions, there was a training session, including 15 sentences, before the formal test. The sentences used for training were different from those used for formal testing.

B. Results and discussion

Figure 2 shows mean percent-correct word identification as a function of $F0_{mean}$. The squares and circles represent conditions flat and F0M, respectively. The solid and dashed curves represent conditions co-located and separated, respectively. Speech intelligibility was clearly higher for the flat condition than for the F0M conditions, especially when $F0_{mean}$ was above 212 Hz. For the F0M conditions, when the target and the masker were perceived as co-located, identification improved when $F0_{mean}$ was made either higher or lower than 252 Hz, and the greatest releasing effect of perceived spatial separation occurred for $F0_{mean} = 252$ Hz. The effect of $F0_{mean}$ was small for the flat condition. Scores obtained when the masker was SSN at −8 dB SMR are shown in Fig. 2 by two parallel lines, solid for the co-located

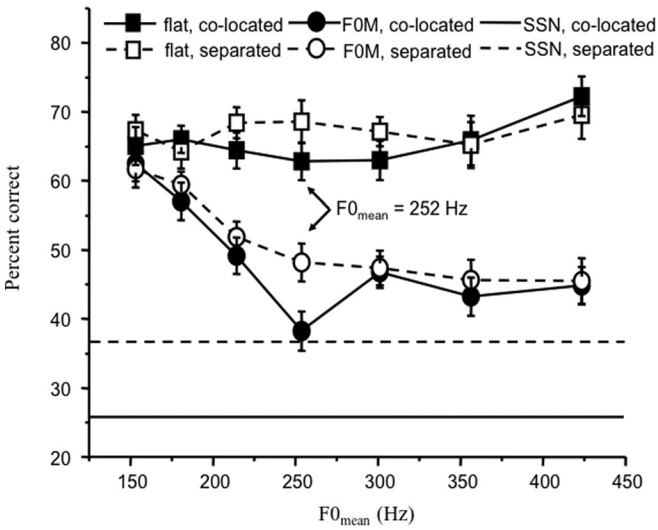

FIG. 2. Mean percent-correct identification of key words across 16 listeners as a function of $F0_{mean}$ for the four masking conditions of experiment I: (1) "flat" harmonics co-located with the target (filled rectangles and solid line); (2) "flat" harmonics spatially separated from the target (open rectangles and dashed line); (3) F0M harmonics co-located with the target (filled circles and solid line); (4) F0M harmonics spatially separated from the target (open circles and dashed line). The horizontal lines show scores obtained using a steady SSN masker co-located (solid line) or spatially separated (dashed line) from the target, drawn from another study (Chen *et al.*, 2008).

condition and dashed for the separated condition. These data were taken from another experiment (Chen *et al.*, 2008) that used the same test materials and the same apparatus and also used young normal-hearing subjects. Speech recognition was higher when the masker was composed of synthesized harmonic tones than when it was SSN, presumably because the spectral gaps in the former allowed more glimpses of the target speech, and/or because the random fluctuations in amplitude of the "steady" noise had a deleterious effect (Drullman, 1995; Stone *et al.*, 2011).

A three-factor within-subject ANOVA confirmed that there was a significant effect of $F0_{mean}$ [$F(6, 90) = 14.2$, $P < 0.001$], and of presence or absence of F0 modulation [$F(1, 15) = 188.1$, $P < 0.001$], but no effect of perceived location [$F(1, 15) = 3.2$, $P > 0.05$]. However, there were significant interactions between $F0_{mean}$ and perceived location [$F(6, 90) = 2.7$, $P = 0.018$], and between $F0_{mean}$ and presence or absence of F0 modulation [$F(6, 90) = 13.8$, $P < 0.001$]. Separate two (presence or absence of F0 modulation) by two (perceived location) within-subject ANOVAs showed that for each $F0_{mean}$ except 150 Hz, there was a significant difference between scores for the flat and F0M conditions [$F(1, 15) \geq 12.6$, $P \leq 0.003$]. These two-way ANOVAs also revealed significant effect of perceived location for $F0_{mean} = 212$ Hz [$F(1, 15) = 7.3$, $P = 0.017$] and $F0_{mean} = 252$ Hz [$F(1, 15) = 9.1$, $P = 0.009$], indicating that perceived spatial separation only led to release from masking when $F0_{mean}$ was equal to or near the mean target F0. Pairwise t-tests showed significant effects of perceived spatial separation for conditions F0M [$t(15) = -2.63$, $P = 0.019$] and flat [$t(15) = -2.73$, $P = 0.015$] only when $F0_{mean} = 252$ Hz. The lack of a significant effect of perceived spatial separation for condition flat when $F0_{mean} = 212$ Hz might

have been due to limited statistical power as the number of subjects was relatively small. Two (perceived location) by seven ($F0_{mean}$) within-subject ANOVAs were conducted for the F0M and the flat conditions, respectively, showing a significant effect of $F0_{mean}$ only for the F0M conditions [$F(6, 90) = 28.0$, $P < 0.001$]. One-way ANOVA and pairwise *t*-tests (Bonferroni corrected) confirmed that for the F0M conditions, when the target and the masker were perceived as co-located, identification for the two lowest $F0_{mean}$ values (150 Hz and 178 Hz) was significantly better than for the other $F0_{mean}$ values [$t(15) \geq 4.26$, $P \leq 0.014$]. Similar effects were observed when the target and the masker were perceived as separated [$t(15) \geq 4.39$, $P \leq 0.011$], except that the difference between scores for $F0_{mean} = 178$ Hz and $F0_{mean} = 212$ Hz did not reach significance.

For condition F0M, performance improved when $F0_{mean}$ was decreased below 252 Hz. This may indicate a role for informational masking the effects of which would decrease when the F0s of the target and masker were made more different. The asymmetrical pattern, whereby the masker was less effective for F0s below than above that of the target, is consistent with the result of Summers *et al*. (2010). They reported that an extraneous competitor formant has less impact on the intelligibility of a dichotically presented sentence when its F0 differs from that of the target formants. Furthermore, competitor formants with F0s above that of the target were more effective than those with F0s below. A similar trend can be seen in the results of Darwin (1981) using the /ru/-/li/ paradigm. Summers *et al*. (2010) offered two possible explanations for this asymmetry: (1) a progressive change in the excitation pattern toward fewer, more intense, and better-resolved harmonics as the F0 of the masker was increased, which could induce a greater masking effect, and (2) pitch perception may be dominated by the higher F0 when two harmonic complex tones with different F0s are mixed in the same frequency region (Deeks and Carlyon, 2004).

Performance for values of $F0_{mean}$ above 178 Hz was markedly poorer for condition F0M than for condition flat. This might have occurred because the F0 modulation was translated to AM in the auditory system, and the AM induced by the masker interfered with the processing of the AM of the target; AM processing is important for speech intelligibility (Shannon *et al*., 1995). Another possibility is that the F0 modulation did introduce some informational masking, but that the manipulation of perceived spatial separation was not effective in reducing that informational masking. However, this seems unlikely given the success of perceived spatial separation in reducing informational masking in other studies as reviewed in the introduction.

In summary, the results showed that: F0M harmonic maskers led to poorer performance than steady harmonic maskers; maskers the mean F0 of which was the same as that of the target speech reduced intelligibility more than those the mean F0s of which differed from that of the target when the target and the maskers were perceived as co-located; and the releasing effect of perceived spatial separation was significant only for maskers the mean F0 of which was the same as that of the target speech. Although the releasing effect was significant for F0mean = 252 Hz for the F0M masker, it was small (about 10%) and comparable with the releasing effect for the SSN (about 10%). The results suggest no effect of informational masking for the steady harmonic maskers and weak effects of informational masking for the F0M harmonic maskers with mean F0 close to that of the target.

The informational masking produced by the F0M masker may have been weak because the target and the masker were very dissimilar, and the masker had a predictable structure with no abrupt changes. The role that F0M plays in informational masking for a speech interferer may have been underestimated by the use of sinusoidally F0M harmonic tones as maskers because similarity and uncertainty, which are key factors underlying informational masking (Durlach *et al*., 2003), were not simulated by the F0M signals. For a speech masker, the F0 is modulated in a much more complex way, and unvoiced segments occur that resemble noise bursts, without any harmonic structure. To test the role of F0M in informational masking for speech in a more appropriate way, the maskers used in experiment II were synthesized signals with F0 contours resembling those in speech and with noise bursts representing unvoiced parts. To control the similarity between the target and masker, the F0 contour used for synthesizing the maskers was based on the F0 contour of the target.

Because the releasing effect of perceived spatial location was relatively small for the harmonic tone maskers, a different approach was taken in experiment II. The effects of energetic masking were taken into account using a method based on the speech intelligibility index (SII) (ANSI, 1997). Any effects of masking above those predicted from the SII were taken as indicating informational masking.

III. EXPERIMENT II: SPEECH-LIKE MASKERS

A. Method

1. Listeners

Ten inexperienced university students (19-24 yr old, mean age = 22 yr, 5 females) participated.

2. Apparatus

All apparatus was the same as for experiment I except that the analog outputs were delivered from only one loudspeaker, which was in the frontal azimuthal plane at 0° azimuth and 200 cm away from the listener.

3. Stimuli

The target stimuli were the same nonsense Chinese sentences as used in experiment I. The masking stimuli were synthesized signals with four types of F0 contours and SSN. The intention with the former was to synthesize signals with similar acoustic characteristics to speech, including a harmonic structure with fluctuating F0 contour during voiced parts, and noise-like structure during unvoiced parts except that there was no formant structure. Formants supply essential cues for phoneme identification, so the synthesized

signals were completely unintelligible and so should not activate knowledge-driven forms of informational masking. However, they might be expected to lead to signal-driven informational masking. Maskers with this property were synthesized in the following way. F0 was extracted frame by frame for each sentence of the target speech. The value was set to 0 if the frame contained silence or an unvoiced signal. An F0 function of time, $F0(t)$, was created by piecewise linear interpolation. For example, if the F0 values for two adjacent voiced frames were $F01$ and $F02$, the frame duration was d ms, the initial time of *frame 1* was $t0$, and the sampling rate was fs kHz, then the F0 between the center of the first frame and that of the next frame was computed using the formula:

$$F0(t) = F01 + \frac{(F02 - F01) \times t}{d \times fs}, \quad (t0 + 0.5d \le t \le t0 + 1.5d). \tag{2}$$

The instantaneous phase of sampling point t, $\phi(t)$, was computed using the following formula:

$$\phi(t) = \phi(t-1) + 2\pi F0(t)/fs. \tag{3}$$

The waveform for the time-varying fundamental component was constructed as

$$A(t) = \sin(\phi(t)). \tag{4}$$

The higher harmonics were synthesized in a similar way, where the value of $F0(t)$ was multiplied by a series of integers and all harmonics had equal amplitude. The frame duration was 10 ms, the sampling rate was 22.05 kHz, and the initial phase of each harmonic was 0. In frames the F0 value of which was 0, the waveform was constructed with Gaussian noise. To avoid abrupt spectral changes in the transitions from harmonic to noise-like segments and vice versa, a raised-cosine window function with duration of 5 ms was applied to each end of every signal segment. The connected waveform was filtered through a speech-spectrum filter to make its long-term spectrum similar to that of the SSN. The amplitude of the noise bursts was adjusted so that the mean level during the bursts was the same as that during the harmonic segments of the masker.

The similarity of the target speech and masking synthesized signals was manipulated by using maskers with different F0 contours. Based on the original F0 contour of the target speech, these F0 contours were modified using the following formula:

$$F'0(t) = \overline{F0} \times \exp(m \times \ln(F0(t)/\overline{F0})), \tag{5}$$

where $F'0(t)$ represents the modified F0 contour, $F0(t)$ represents the original F0 contour, and $\overline{F0}$ represents the mean F0 of the sentence. This formula is similar to that used by Binns and Culling (2007). They manipulated F0 contours by setting m to 1, 0.5, 0.25, 0, and −1. Values of $m = 1$, 0, and −1 lead to original, monotonized and inverted F0 contours, respectively. Setting m to 0.5 or 0.25 results in F0 contours with the same shape as the original contour but with a reduced amount of F0 fluctuation. Based on the assumption that increased F0 fluctuations should be more effective in producing informational masking than reduced F0 fluctuations, we used $m = 2$ instead of $m = 0.5$ and 0.25. In summary, four values of m were used, 1, 0, −1, and 2, corresponding to conditions original, flat, inverted, and amplified, respectively, as illustrated in Fig. 3.

Figure 4 shows time waveforms and spectrograms of a sample target sentence and the corresponding five types of masker. For the time waveforms, all four synthesized maskers have lower envelope fluctuations than for the target speech. As can be seen from the spectrograms, the periodic and non-periodic parts of the synthesized maskers are aligned with those of the target speech. None of the synthesized maskers led to any phoneme perception, presumably because they contained no formant information.

The SSN was constructed by adding together 57 sentences spoken by each of 25 female speakers and another 56 sentences spoken by 25 different female speakers as described by Yang *et al.* (2007). The very large number of sentences meant that the SSN sounded like noise rather than babble. Note that the spectrum of the SSN was not exactly the same as the mean spectrum of the target speech as it was based on a different speech corpus. The SSN employed is used as a standard speech masker in the Key Laboratory of Machine Perception of Peking University.

4. Design and procedure

Psychometric functions for recognition of the target speech were measured. Two factors were manipulated: (1) type of masker (original, flat, amplified, inverted, and SSN)

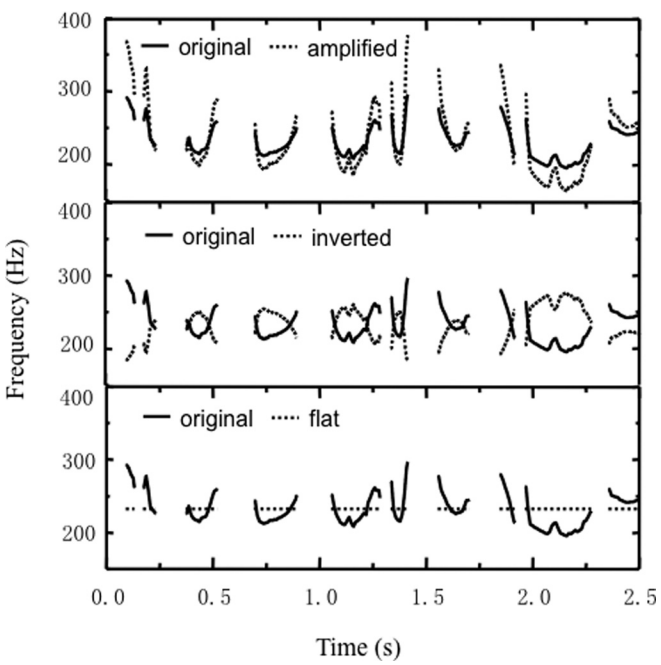

FIG. 3. Examples of manipulated F0 contours used in experiment II. The manipulation $m = 1$ corresponds to condition "original," represented by the solid line in each panel; the manipulations $m = 0$, −1, and 2 correspond to the three conditions: flat (bottom panel), inverted (middle panel), and amplified (top panel), represented by the dotted lines.

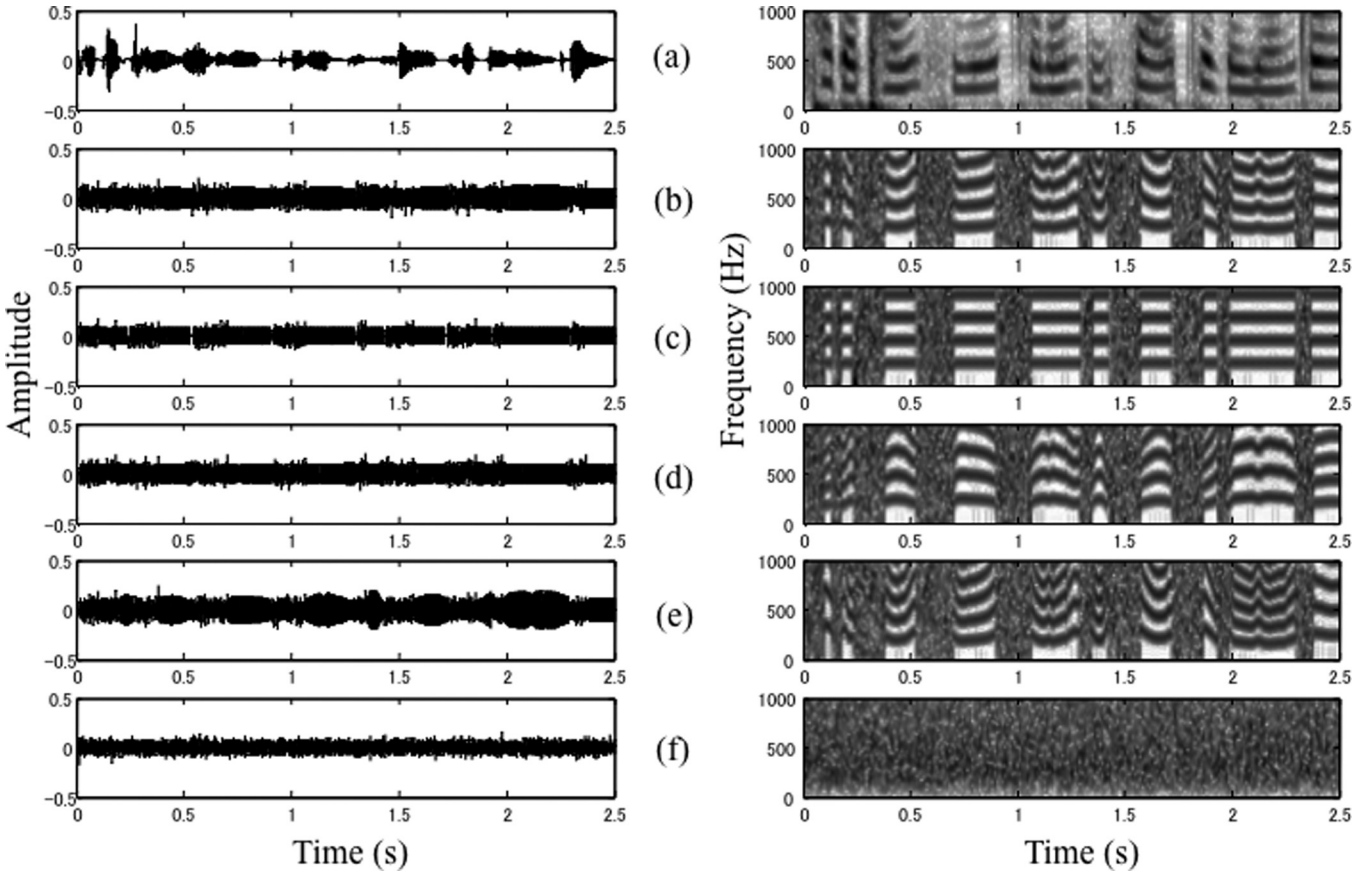

FIG. 4. Waveforms (left) and spectrograms (right) of the maskers used in experiment II. Row (a) represents the target speech. Rows (b) to (e) represent maskers with four manipulations of the F0 contour: original, flat, inverted and amplified, respectively. Row (f) represents the SSN.

and (2) SMR (−9, −5, −1, and 3 dB). There were 20 (5 × 4) conditions (15 sentences per condition) for each listener, and they were organized into five blocks according to the type of masker. In each block, 60 sentences (15 at each SMR) were presented with SMRs in a random order, and the order of the five blocks was determined using a Latin-square design. For each listener, 20 test lists were assigned to the 20 conditions randomly. The test procedure and scoring method were similar to those for experiment I. Note that the masking "sentence" was always based on the target sentence.

B. Results and discussion

Figure 5 shows average percent correct word identification as a function of SMR for the five maskers. The smooth curves are logistic function fits to the data of the form,

$$p(y) = \frac{1}{1 + e^{-\sigma(x-\mu)}} \tag{6}$$

where $p(y)$ is the probability of correctly identifying the key words at SMR x, μ is the SMR corresponding to 50% correct, and σ is the slope of the psychometric function. The parameters μ and σ were fitted using the Levenberg–Marquardt method (Wolfram, 1991). The results indicate that maskers with F0M harmonics led to poorer intelligibility than the SSN, and the synthesized masker the F0 contour of which was the same as that of the target produced the lowest scores.

Similar psychometric functions were fitted to the data for individual listeners. Figure 6 shows the mean threshold values (μ) and slope values (σ) for each masker type. The threshold was lowest for the SSN. A one-way ANOVA indicated that the effect of masker type was significant [$F(4, 36) = 20.2$, $P < 0.001$)]. Pairwise t-tests (Bonferroni corrected) indicated that the threshold for condition "original" was significantly higher than for condition "amplified" [$t(9) = 5.27$, $P = 0.005$] but was not significantly different from that for conditions "flat" [$t(9) = 1.79$, $P > 0.05$] or "inverted"

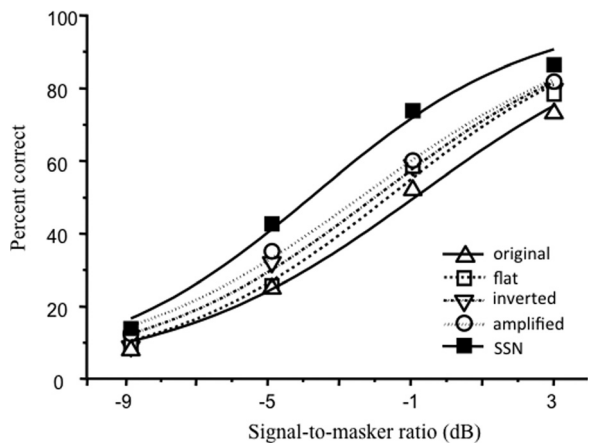

FIG. 5. Symbols show the mean percent correct identification of key words across 10 listeners as a function of SMR for the five masking conditions of experiment II (see key). The curves are fitted psychometric functions.

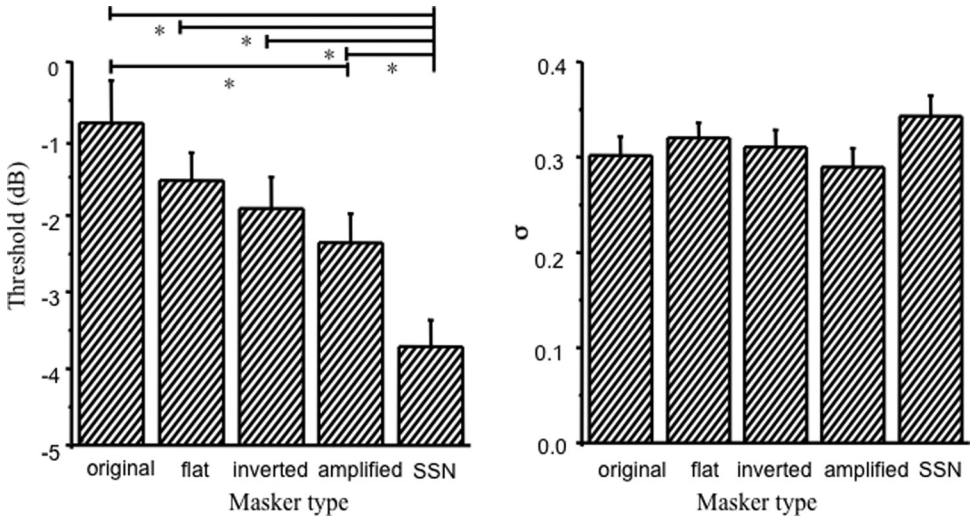

FIG. 6. Average threshold values, μ (left), and slope values, σ (right), for each type of masker. Error bars indicate ± 1 standard error of the mean. Significant differences between conditions are indicated by "*".

$[t(9) = 2.49, \quad P > 0.05]$. Thresholds for all synthesized maskers were significantly higher than that for SSN $[t(9) \geq 4.55, P \leq 0.014]$, even for the flat masker, which was not F0 modulated. This suggests that the main feature of the synthesized maskers that led to greater masking than the SSN was the alternation between periodic and non-periodic parts. It may also have been the case that the synchrony of harmonic and noise-like parts of the target and masker was important. This possibility was addressed in experiment III.

Figure 6 (right) shows the slope parameter σ for the five maskers. Slope values were similar across the five maskers. A one-way ANOVA of the σ values indicated that the effect of masker type was not significant $[F(4, \ 36) = 1.78, P > 0.05]$. Previous work has shown that the slope is steeper for an SSN masker than for a speech masker (Baer and Moore, 1994; Wu *et al.*, 2005). However, it should be noted that the non-SSN maskers used here did not have the large amplitude fluctuations that occur in speech and that contribute to the shallow psychometric function when a speech masker is used.

The averaged one-third octave spectra for 60 sentences (4 lists) of the target and for each type of masker are presented in the left panel of Fig. 7. As noted earlier, the spectrum of the SSN differed somewhat from that of the target sentences. The spectrum of the SSN was similar to the spectra of the synthesized maskers except for small differences for frequencies between 2 and 8 kHz.

SII values for the 20 conditions (5 types of masker $\times$ 4 SMRs) used in experiment II were calculated, using the method described in ANSI (1997). Sixty sentences (4 lists) from the test corpus were used as the target samples for each condition. For each sample, the level of the corresponding masker was set according to the SMR, and then the one-third octave spectra of the target and masker were used as the input to the SII calculation procedure. The mean SII value for each condition was calculated by averaging the values for the 60 samples. The mean values are shown in the right panel of Fig. 7.

The pattern of the SII values was quite different from that for the data. For the given SMRs, the SII values for the masker "amplified" were higher than for all other maskers; the SII values for the maskers "original" and "inverted" were almost the same and both were close to the SII values for the SSN masker; and the SII values for the masker "flat" were the lowest. This pattern contrasts with the data for which performance was poorer for the masker with the original than with the inverted F0 contours, both giving poorer performance than with the SSN masker. SII scores were higher for the masker with amplified F0 contours than for the SSN masker, whereas for the data the reverse was true.

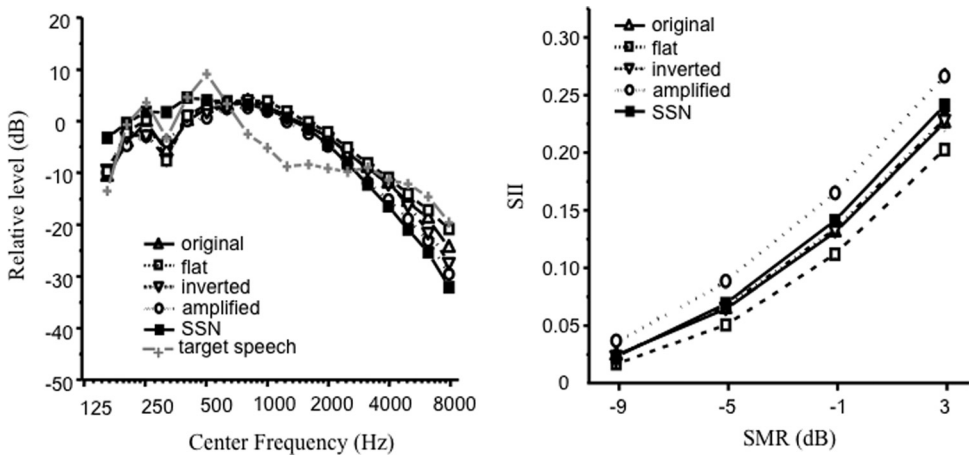

FIG. 7. Left: averaged one-third-octave band spectra for the target (open circles) and five types of maskers used in experiment II. Right: SII values calculated for the stimuli used in experiment II.

Evidently, the SII model cannot account for the effects of masker type on performance, suggesting that something other than energetic masking had a strong influence.

In summary, the speech-like but unintelligible maskers led to lower intelligibility than the SSN, and performance was poorest when the F0 contour of the masker matched that of the target sentence. The ordering of the results across conditions was not the same as predicted by the SII, suggesting that the results cannot be accounted for entirely in terms of energetic masking.

It is instructive to compare the results of experiments I and II for conditions that were similar across the two experiments. Recall that experiment I was conducted using an SMR of −8 dB. For the co-located SSN of experiment I (solid horizontal line in Fig. 2), the mean score was approximately 26%. Based on the psychometric function fitted to the mean results, the corresponding score for experiment II for an SMR of −8 dB was about 21%, which is in reasonable agreement. For the flat masker of experiment I for the condition where the mean F0 of the masker equaled the mean F0 of the target, the mean score was about 63%. For the flat masker of experiment II, the mean score for an SMR of −8 dB was only about 12%. This very large difference across the two experiments was probably caused mainly by the fact that the masker was a continuous harmonic sound in experiment I but alternated between harmonic and noise-like portions in experiment II. It is also possible that the low score in experiment II was partly caused by the noise-like portions of the masker being synchronized to the unvoiced portions or silences in the target speech. Experiment III was conducted to assess the importance of this second factor.

IV. EXPERIMENT III: EFECT OF THE TIMING OF THE NOISE-LIKE BURSTS

A. Method

1. Listeners and apparatus

Six young university students participated (19-28 yr old, mean age = 22 yr, 2 females). They had no previous experience listening to the sentences used in this experiment. All apparatus was the same as for experiment II.

2. Stimuli

Six types of masker were used. Example waveforms (left) and spectrograms (right) of the six maskers are shown in Fig. 8. The maskers were SSN, flat as used in experiment I (called here flatI), and flat as used in experiment II (called here flatII). The masker flatI was a steady harmonic complex tone while the masker flatII alternated between harmonic and noise-like segments, and the noise-like segments were synchronized to the unvoiced segments of the target. Note that $F0_{mean}$ was fixed at 252 Hz for the masker flatI.

The fourth masker was produced by replacing the harmonic tone in masker flatI with a 30-ms noise burst periodically every 200 ms. This masker is called flatI+P (P for

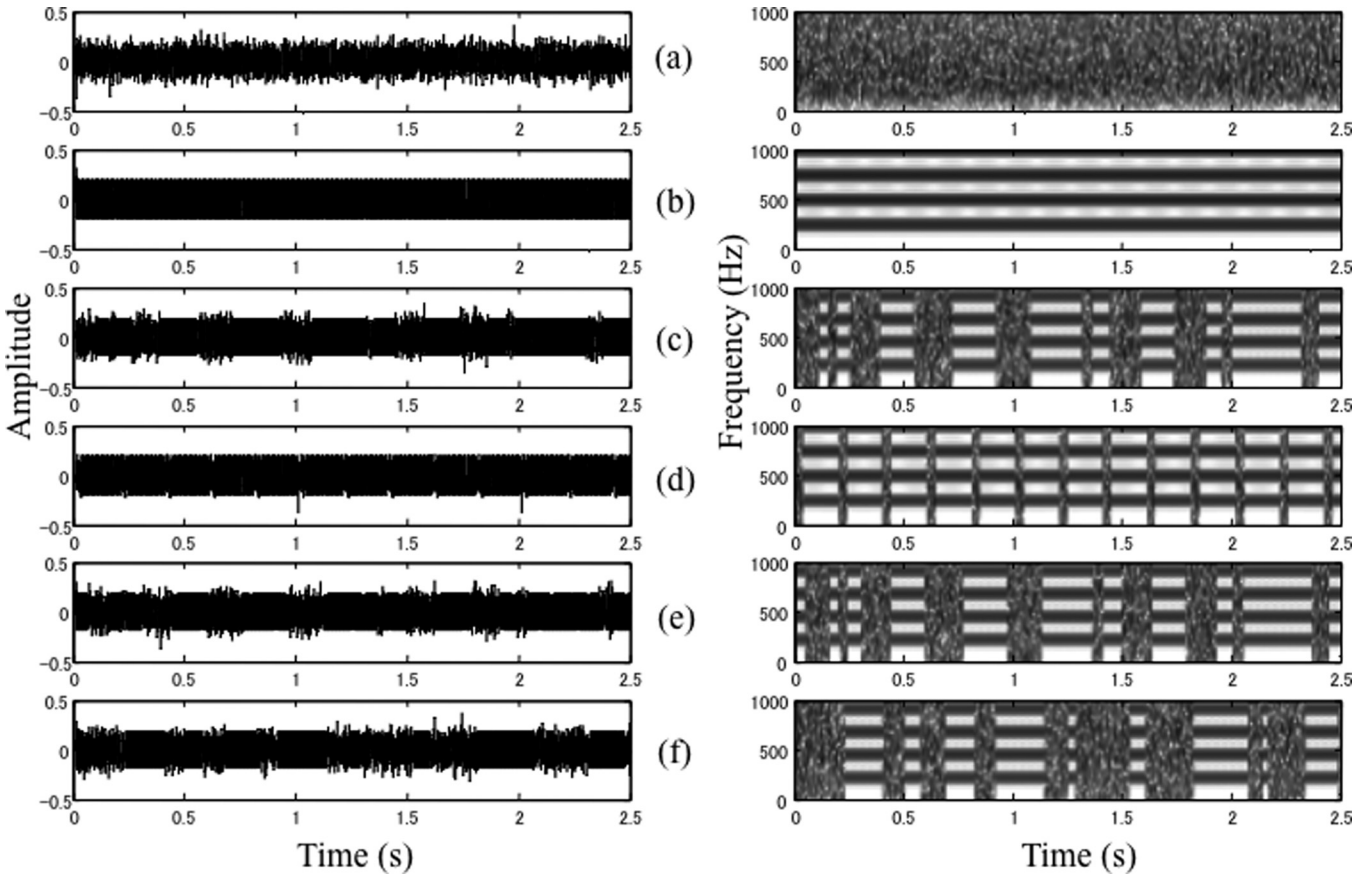

FIG. 8. Waveforms (left) and spectrograms (right) of the maskers used in experiment III. Rows (a) to (f) represent conditions SSN, flatI, flatII, flatI + NB, flatII_SHI, and flatII_IND, respectively.

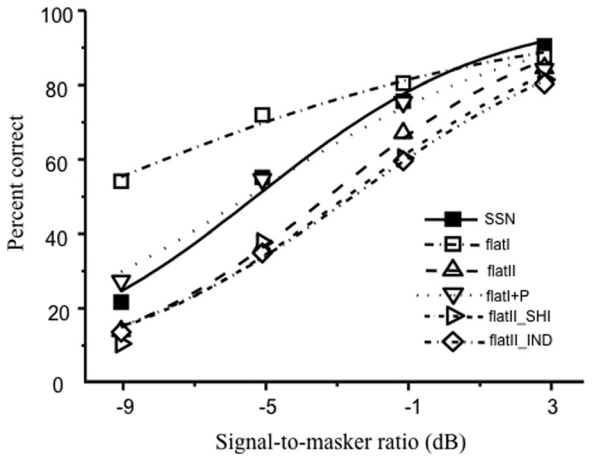

FIG. 9. Symbols show the mean percent correct identification of key words across six listeners as a function of SMR for the six masking conditions of experiment III. The curves are fitted psychometric functions.

periodic). The fifth masker was modified based on flatII; each noise burst was delayed by 50 ms relative to its original position. This masker is called flatII_SHI (SHI stands for shifted). The sixth masker was also based on flatII, but for each trial, the timing and the duration of the noise bursts were based on the timing and duration of the unvoiced segments in an independent sentence. A different independent sentence was used for each trial. This masker is called flatII_IND. The independent sentences were spoken by the same talker as for the target, and the sentences were similar to the target sentences, but the content was different from that for the target on each trial. For the last two maskers, portions of the signal "vacated" by the shifted noise burst were replaced with the harmonic signal.

3. Design and procedure

The design and procedure were the same as for experiment II. Two factors were manipulated: type of masker and SMR (−9, −5, −1, and 3 dB). There were 24 (6×4) conditions for each listener, and they were organized into six blocks according to the type of masker. The test order of the six blocks was determined using a Latin-square design.

B. Results and discussion

Figure 9 shows average percent correct word identification as a function of SMR for the six maskers. The smooth curves are logistic function fits to the data. Performance with masker flatI (open squares) was better than with the SSN masker (filled squares), especially for the SMRs of −9 and −5 dB, consistent with the results of experiment I. Performance with masker flatII (up-pointing triangles) was worse than with the SSN masker, consistent with the results of experiment II. When noise bursts alternated regularly with the harmonic tone (condition flatI+P, down-pointing triangles), performance was close to that for the SSN masker. Performance was similar for the maskers where the noise bursts were delayed relative to those in condition flatII (condition flatII_SHI, right-pointing triangles) or were temporally positioned based on an independent sentence (condition flatII_IND, diamonds), and both led to performance close to that for condition flatII.

Logistic psychometric functions were fitted to the data of individual subjects. Figure 10 shows the mean threshold values (μ) and slope values (σ) for each masker type. A one-way ANOVA on the threshold values indicated that the effect of masker type was significant [$F(5, 25) = 53.1$, $P < 0.001$]. Pairwise t-tests (Bonferroni corrected) indicated that the threshold for condition flatI was significantly lower than for all other conditions [$t(5) \leq -6.66$, $P \leq 0.017$], and the threshold for condition SSN was also significantly lower than for all other conditions [$t(5) \leq -6.16$, $P \leq 0.025$] except flatI+P.

These results suggest that the largest informational masking occurs when the noise segments in the masker alternate with the harmonic segments in an irregular, speech-like manner (conditions flatII, flatII_SHI, and flatII_IND). When the alternation is regular (condition flatI+P), somewhat less informational masking occurs. The synchrony of the noise bursts in the masker with the unvoiced portions of the target does not appear to be important as performance was similar (and did not differ significantly) for condition flatII (where such synchrony did occur) and for conditions flatII_SHI and flatII_IND, for which synchrony did not occur. However, the

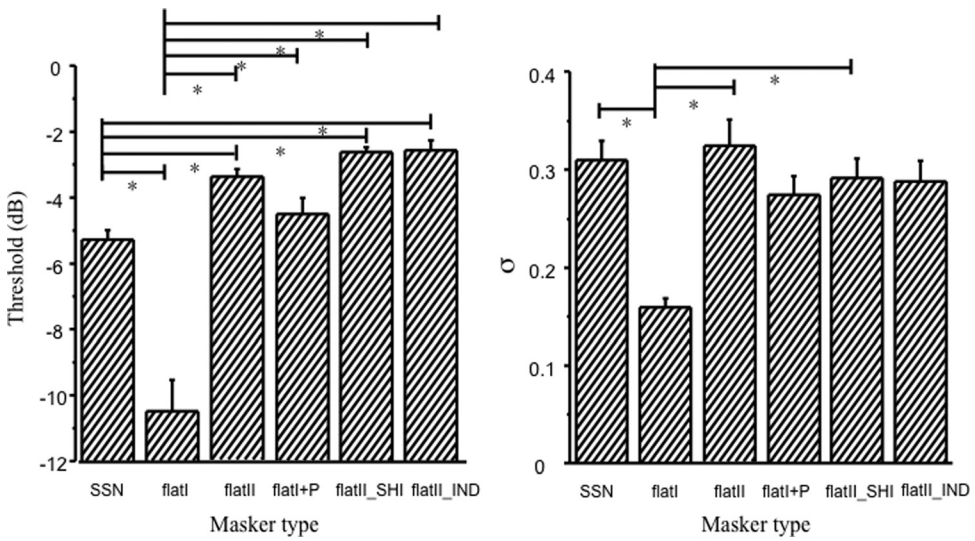

FIG. 10. Average threshold values, μ (left), and slope values, σ (right), for each type of masker in experiment III. Error bars indicate ± 1 standard error of the mean. Significant differences between conditions are indicated by "*".

similarity of performance across conditions flatII, flatII_SHI, and flatII_IND may have been produced by the interaction of competing effects; this is discussed in more detail in the following text.

Figure 10 (right) shows the slope parameter σ for the six maskers. A one-way ANOVA indicated that the effect of masker type on the σ values was significant [$F(5, 25) = 7.9$, $P < 0.001$]. Pairwise t-tests (Bonferroni corrected) indicated that the slope for condition flatI was significantly lower than for the conditions SSN, flatII, and flatII_SHI [$t(5) \leq -6.16$, $P \leq 0.025$]. Excluding condition flatI, the slopes were similar and not significantly different across conditions. The shallow slope for condition flatI was a consequence of the relatively good performance for that condition at low SMRs. This good performance probably reflects the lack of informational masking produced by the flatI masker.

IV. GENERAL DISCUSSION

A. Effects of mean F0

The results reviewed in the introduction showed that the identification of speech in a speech masker improves with increasing F0 difference between the target and masker. In contrast, experiment I showed that when the masker was composed of unmodulated harmonics (condition flat), there was little effect of F0 and no consistent effect of perceived spatial separation of the target and masker. This may have happened because the flat masker was easily perceptually segregated from the target even when the target and masker were co-located, and therefore the masker produced little informational masking. When the maskers were composed of F0M harmonics, a spatial release from masking was found only when the mean F0 of the masker matched that of the target (252 Hz). Also, performance for the co-located condition was poorest when the mean F0 of the masker matched that of the target. This is consistent with previous studies showing effects of F0 differences of the target and masker and is consistent with the idea that the F0M masker produced a small amount of informational masking. However, it is not clear why performance did not improve progressively when the mean F0 of the masker was increased from 300 to 424 Hz, which led to an increasing difference between the mean F0s of the target and masker.

Overall, the results suggest that the flat masker used in experiment I produced negligible informational masking, but the F0M masker may have produced a small amount of informational masking, especially when its mean F0 equaled that of the target.

B. Effects of F0 contour

While it is clearly established that differences in mean F0 between a target talker and competing talker(s) can facilitate tracking of the target talker (Brokx and Nooteboom, 1982; Assmann and Summerfield, 1989; Bird and Darwin, 1998; Darwin and Hukin, 2000; Darwin *et al.*, 2003), it is less clear whether differences in F0 contour between the target and background have a beneficial effect. The role of F0 contour in speech-on-speech masking has been assessed by Binns and Culling (2007). They reported that speech reception thresholds (SRTs) for speech in SSN increased slightly when the F0 contour was flattened or inverted (by 0.4 and 1.3 dB, respectively). The increase was greater when a single-talker masker was used, but no effect was found when the F0 contour of the masker was manipulated. In their work, the effect of the relationship between the F0 of the target and of the masker was not evaluated. The present study focused on the similarity of the F0 contours of the target and the masker; the F0 contour of the masker was manipulated based on the F0 contour of the target.

In experiment II, the F0M of the original masker was almost identical to that of the target. As a result, the only cue that could be used to segregate the target speech from the original masker was the short-term spectral envelope and changes in spectral envelope over time, presumably supplemented by knowledge-driven processes. Consistent with this, performance was poorer for condition "original" than for the other conditions. However, the effect was small, and the difference between condition "original" and the other conditions was only significant for condition "amplified." Furthermore, performance may have been worse for condition "original" than for the other conditions due to energetic masking because for that condition, the harmonics of the masker always coincided exactly in frequency with the harmonics of the target.

It is noteworthy that performance for condition "flat" was only very slightly (non-significantly) better than for condition "original," despite the fact that the F0 contours of the target and masker were almost identical for the latter, but very different for the former. Also the results of experiment I suggest that a harmonic masker with a flat F0 produces very little informational masking. Overall, the results suggest that the similarity of the F0 contour of the target and masker has very little influence on intelligibility or on informational masking. This is consistent with earlier results suggesting that human listeners have poor sensitivity to the "coherence" of F0M across sounds. For example, listeners have difficulty determining whether two tones are modulated in phase or out of phase (Carlyon, 1991). Also, listeners do not seem to be able to use differences in the pattern of F0M across sounds to segregate those sounds (Culling and Summerfield, 1995; Lyzenga and Moore, 2005).

C. The effect of synchrony of features of the target and masker

Synchronous fluctuations in amplitude across different frequency components in a complex sound tend to promote perceptual grouping of those components (Darwin, 1984; Bregman, 1990). Synchrony of onsets appears to be especially important. Based on this, one might have thought that the synchrony of the unvoiced segments of the target and the noise segments of the masker (and corresponding synchrony of the voiced segments of the target and the harmonic segments of the masker) would promote perceptual fusion of the target and masker and lead to especially strong informational masking. However, the results of experiment III showed that masker flatII, for which such synchrony was

present, did not lead to poorer performance than maskers flatII_SHI and flatII_IND, for which no such synchrony was present.

It is possible that the similarity of performance obtained with maskers flat II, flatII_SHI, and flatII_IND was a result of two competing factors. When the noise bursts in the masker were not synchronous with those of the target, this may have decreased informational masking. However, it might also have somewhat increased energetic masking because noise segments in the masker would have partially overlapped with harmonic segments of the target, and noise masks a harmonic complex sound more effectively than a harmonic complex sound masks noise or than a complex tone masks another complex tone (Gockel *et al.*, 2002; Micheyl *et al.*, 2006).

However, it is likely that any increase in energetic masking produced by asynchrony of the noise bursts (conditions flatII_SHI and flatII_IND) was small because the SSN masker led to significantly lower thresholds than the flatII, flatII_SHI, and flatII_IND maskers. It can reasonably be concluded that informational masking was produced by all three of the maskers with noise segments that alternated with harmonic segments in an irregular, speech-like manner.

V. CONCLUSIONS

The effect of synthetic speech-like maskers on speech intelligibility was explored. To avoid strong effects of knowledge-driven processes, the maskers were synthesized harmonic signals, or harmonic-plus-noise signals, without any linguistic information or formant structure. The following factors were varied: The mean F0 of the masker relative to that of the target, the similarity of the F0 contour of the target and masker, and the presence and relative timing of noise-like portions in the masker. The main findings and conclusions are:

(1) Experiment I showed that a sinusoidally F0 modulated masker the mean F0 of which was the same as that of the target speech reduced intelligibility more than F0M maskers with mean F0s different from that of the target when the target and maskers were perceived as co-located. The releasing effect of perceived spatial separation was significant only for the former. F0M maskers led to poorer speech identification than steady harmonic maskers; for the latter, variation of the F0 of the masker had hardly any effect. Both types of harmonic masker led to better performance than obtained with SSN. The results suggest no effect of informational masking for the steady harmonic maskers and weak effects of informational masking for the F0M harmonic maskers with mean F0 close to that of the target.

(2) Experiment II showed that when the maskers were synthetic non-speech sounds with periodic segments synchronized to voiced segments of the target speech and noise-like segments synchronized to unvoiced segments of the target speech, performance was much poorer than obtained with the continuous harmonic maskers in experiment I. Also performance with the synthetic maskers in experiment II was poorer than obtained with SSN. The similarity of the F0 contour of the target and masker had only a small effect on intelligibility consistent with previous work showing that listeners are relatively insensitive to the coherence of F0M across sounds.

(3) Experiment III showed that the important feature leading to informational masking was irregular alternation of harmonic and noise segments in the masker. Synchrony between noise segments of the masker and unvoiced segments of the target was not important. Regular alternation of harmonic and noise segments in the masker led to some informational masking, but not as much as for the maskers with irregular alternation.

(4) The results show that speech-like maskers without any linguistic content can produce signal-driven informational masking.

ACKNOWLEDGMENTS

The work was supported in part by the National Natural Science Foundation of China (Grant Nos. 60535030, 90920302, and 30711120563), and a HGJ Grant of China (Grant No. 2011ZX01042-001-001). J. Chen was supported by a Newton International Fellowship from the Royal Society, UK. B. Moore was supported by the MRC (UK). We thank Michael Akeroyd, Brian Roberts, and two anonymous reviewers for very helpful comments on an earlier version of this paper.

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

Assmann, P. F., and Summerfield, Q. (**1989**). "Modeling the perception of concurrent vowels: vowels with the same fundamental frequency," J. Acoust. Soc. Am. **85**, 327–338.

Baer, T., and Moore, B. C. J. (**1994**). "Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech," J. Acoust. Soc. Am. **95**, 2277–2280.

Binns, C., and Culling, J. F. (**2007**). "The role of fundamental frequency contours in the perception of speech against interfering speech," J. Acoust. Soc. Am. **122**, 1765–1776.

Bird, J., and Darwin, C. J. (**1998**). "Effects of a difference in fundamental frequency in separating two sentences," in *Psychophysical and Physiological Advances in Hearing* (Whurr, London), pp. 263–269.

Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (Bradford Books, MIT Press, Cambridge, MA), pp. 38–42.

Brokx, J. P. L., and Nooteboom, S. G. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phon **10**, 23–36.

Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (**2001**). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **110**, 2527–2538.

Carlyon, R. P. (**1991**). "Discriminating between coherent and incoherent frequency modulation of complex tones," J. Acoust. Soc. Am. **89**, 329–340.

Chen, J., Wu, X. H., Zou, X. F., Zhang, Z. P., Xu, L. J., Wang, M. Y., Li, L., and Chi, H. S. (**2008**). "Effect of speech rate on speech-on-speech masking (A)," J. Acoust. Soc. Am. **123**, 3713.

Culling, J. F., and Summerfield, Q. (**1995**). "The role of frequency modulation in the perceptual segregation of concurrent vowels," J. Acoust. Soc. Am. **98**, 837–846.

Darwin, C. J. (**1981**). "Perceptual grouping of speech components differing in fundamental frequency and onset time," Q. J. Exp. Psychol. **33A**, 185–207.

Darwin, C. J. (**1984**). "Perceiving vowels in the presence of another sound: constraints on formant perception," J. Acoust. Soc. Am. **76**, 1636–1647.

Darwin, C. J., Brungart, D. S., and Simpson, B. D. (**2003**). "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," J. Acoust. Soc. Am. **114**, 2913–2922.

Darwin, C. J., and Hukin, R. W. (**2000**). "Effectiveness of spatial cues, prosody, and talker characteristics in selective attention," J. Acoust. Soc. Am. **107**, 970–977.

Deeks, J. M., and Carlyon, R. P. (**2004**). "Simulations of cochlear implant hearing using filtered harmonic complexes: Implications for concurrent sound segregation," J. Acoust. Soc. Am. **115**, 1736–1746.
Drullman, R. (**1995**). "Temporal envelope and fine structure cues for speech intelligibility," J. Acoust. Soc. Am. **97**, 585–592.
Durlach, N. I., Mason, C. R., Kidd, G., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (**2003**). "Note on informational masking," J. Acoust. Soc. Am. **113**, 2984–2987.
Elhilali, M., Chi, T., and Shamma, S. A. (**2003**). "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," Speech Commun. **41**, 331–348.
Fletcher, H., and Galt, R. H. (**1950**). "The perception of speech and its relation to telephony," J. Acoust. Soc. Am. **22**, 89–151.
French, N. R., and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90–119.
Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2001**). "Spatial release from informational masking in speech recognition," J. Acoust. Soc. Am. **109**, 2112–2122.
Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2004**). "Effect of number of masking talkers and auditory priming on informational masking in speech recognition," J. Acoust. Soc. Am. **115**, 2246–2256.
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.
Gockel, H., Moore, B. C. J., and Patterson, R. D. (**2002**). "Asymmetry of masking between complex tones and noise: The role of temporal structure and peripheral compression," J. Acoust. Soc. Am. **111**, 2759–2770.
Houtgast, T., and Steeneken, H. J. (**1985**). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Acoust. Soc. Am. **77**, 1069–1077.
Huang, Y., Huang, Q., Chen, X., Qu, T. S., Wu, X. H., and Li, L. (**2008**). "Perceptual integration between target speech and target-speech reflection reduces masking for target-speech recognition in younger adults and older adults," Hear. Res. **244**, 51–65.
Li, L., Daneman, M., Qi, J. G., and Schneider, B. A. (**2004**). "Does the information content of an irrelevant source differentially affect spoken word recognition in younger and older adults?" J. Exp. Psych.: Hum. Per. Perf. **30**, 1077–1091.
Litovsky, R. Y., Colburn, H. S., Yost, W. A., and Guzman, S. J. (**1999**). "The precedence effect," J. Acoust. Soc. Am. **106**, 1633–1654.
Lyzenga, J., and Moore, B. C. J. (**2005**). "Effect of FM coherence for inharmonic stimuli: FM-phase discrimination and identification of artificial double vowels," J. Acoust. Soc. Am. **117**, 1314–1325.
Mattys, S. L., Brooks, J., and Cooke, M. (**2009**). "Recognizing speech under a processing load: dissociating energetic from informational factors," Cogn. Psychol. **59**, 203–243.
McAdams, S. (**1989**). "Segregation of concurrent sounds. I.: Effects of frequency modulation coherence," J. Acoust. Soc. Am. **86**, 2148–2159.
Micheyl, C., Bernstein, J. G. W., and Oxenham, A. J. (**2006**). "Detection and F0 discrimination of harmonic complex tones in the presence of competing tones or noise," J. Acoust. Soc. Am. **120**, 1493–1505.
Rakerd, B., Aaronson, N. L., and Hartmann, W. M. (**2006**). "Release from speech-on-speech masking by adding a delayed masker at a different location," J. Acoust. Soc. Am. **119**, 1597–1605.
Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (**2005**). "Release from informational masking by time reversal of native and non-native interfering speech," J. Acoust. Soc. Am. **118**, 1274–1277.
Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (**2006**). "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," J. Acoust. Soc. Am. **120**, 3988–3997.
Shannon, R. V., Zeng, F. -G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Simpson, S. A., and Cooke, M. (**2005**). "Consonant identification in N-talker babble is a nonmonotonic function of N," J. Acoust. Soc. Am. **118**, 2775–2778.
Sjolander, K. (**2006**). "The Snack Sound Toolkit," http://www.speech.kth.se/snack/index.html (Last viewed November 21, 2010).
Stone, M. A., Fullgrabe, C., Mackinnon, R. C., and Moore, B. C. J. (**2011**). "The importance for speech intelligibility of random fluctuations in"steady"background noise," J. Acoust. Soc. Am. **130**, 2874–2881.
Summers, R. J., Bailey, P. J., and Roberts, B. (**2010**). "Effects of differences in fundamental frequency on across-formant grouping in speech perception," J. Acoust. Soc. Am. **128**, 3667–3677.
Wallach, H., Newman, E. B., and Rosenzweig, M. R. (**1949**). "The precedence effect in sound localization," Am. J. Psychol. **62**, 315–336.
Watson, C. S. (**1987**). "Uncertainty, informational masking, and the capacity of immediate auditory memory," in *Auditory Processing of Complex Sounds*, edited by W. A. Yost and C. S. Watson (Lawrence Erlbaum Associates, N.J.), pp. 267–277.
Wolfram, S. (**1991**). *Mathematica: A System for Doing Mathematics by Computer* (Addison-Welsey, New York), pp. 53–76.
Wu, X. H., Chen, J., Yang, Z. G., Huang, Q., Wang, M. Y., and Li, L. (**2007**). "Effect of number of masking talkers on speech-on-speech masking in Chinese," in *Interspeech2007*, Antwerp, Belgium, pp. 390–393.
Wu, X. H., Wang, C., Chen, J., Qu, H. W., Li, W. R., Wu, Y. H., Schneider, B. A., Li, L. (**2005**). "The effect of perceived spatial separation on informational masking of Chinese speech," Hear. Res. **199**, 1–10.
Wu, X. H., Yang, Z. G., Huang, Y., Chen, J., Li, L., Daneman, M., Schneider, B. A. (**2011**). "Cross-language differences in informational masking of speech by speech: English versus Mandarin Chinese," J Speech Lang Hear Res. **54**, 1506–1524
Yang, Z. G., Chen, J., Huang, Q., Wu, X. H., Wu, Y. H., Schneider, B. A., and Li, L. (**2007**). "The effect of voice cuing on releasing Chinese speech from informational masking," Speech Commun. **49**, 892–904.
Zurek, P. M. (**1980**). "The precedence effect and its possible role in the avoidance of interaural ambiguities," J. Acoust. Soc. Am. **67**, 952–964.