Journal of Experimental Psychology:
Human Perception and Performance
2009, Vol. 35, No. 5, 1618–1628

© 2009 American Psychological Association
0096-1523/09/$12.00 DOI: 10.1037/a0015791

Transient Auditory Storage of Acoustic Details Is Associated With Release of Speech From Informational Masking in Reverberant Conditions

Ying Huang, Qiang Huang, Xun Chen, Xihong Wu, and Liang Li
Peking University

Perceptual integration of the sound directly emanating from the source with reflections needs both temporal storage and correlation computation of acoustic details. We examined whether the temporal storage is frequency dependent and associated with speech unmasking. In Experiment 1, a break in correlation (BIC) between interaurally correlated wideband or narrowband noises was detectable even when an interaural interval (IAI) was introduced. The longest IAI, which varied markedly across participants, could be up to about 20 ms for wideband noise and decreased as the center frequency was increased for narrowband noises. In Experiment 2, when the interval between target speech and its single-reflection simulation (intertarget interval [ITI]) was reduced from 64 to 0 ms, intelligibility of target speech was markedly improved under speech-masking but not noise-masking conditions. The longest effective ITI correlated with the longest IAI for detecting the BIC only in the low-frequency (≤400 Hz) narrowband noise. Thus the ability to temporally store fine details contributes to perceptual integration of correlated leading and lagging sounds, which in turn, contributes to releasing speech from informational masking in noisy, reverberant environments.

auditory memory, temporal integration, informational masking, energetic masking, reverberation

Both transient storage of acoustic features and temporal integration of relevant signals are important for detecting, recognizing, and localizing sounds in everyday environments (Bregman, 1990; Näätänen & Winkler, 1999). At the early stage of auditory perception, fine-structure details of sound waves must be faithfully maintained for periods of time, otherwise auditory processing at later stages would be impossible. Although listeners are not able to be aware of most acoustic details of a wideband sound, the human's auditory system has the dramatic ability to process acoustic details. For example, listeners with normal hearing are very sensitive to small differences between a wideband noise delivered at one ear and its copy delivered at the other ear (Gabriel & Colburn, 1981; Goupell & Hartmann, 2006; Pollack & Trittipoe, 1959). A change in the interaural correlation of wideband noise can cause a change in the perception of the noise, that is, the compactness, number, and placement of wideband noise images depend on the degree of interaural correlation (Blauert & Lindemann, 1986). Thus the human's auditory system can calculate the sound correlation between the two ears and represent the consequence of the calculation at the perceptual level.

Previous studies have shown that human listeners are also able to detect a very transient drop in correlation between the two ears (termed [BIC]; i.e., a transient drop of interaural correlation from 1 to 0 and then returning to 1), showing the marked ability to temporally resolve fast changes in interaural configurations (Akeroyd & Summerfield, 1999; Boehnke, Hall, & Marquardt, 2002). Note that introducing a change in interaural correlation for wideband noises does not change the energy and spectrum in the signals, but it can change the loudness of the signals (Culling, 2007). However, these studies did not investigate whether this binaural ability can be maintained when an interaural interval (IAI) is introduced. Because the preservation of the sensitivity to the BIC even when an IAI is introduced indicates whether fine-structure information of a noise is maintained for the time of the IAI (Huang, Kong, Fan, Wu, & Li, 2008), measuring the IAI when the BIC is detectable can provide a way of investigating the temporal storage of acoustic details.

The sensitivity to the interaural correlation appears to be frequency dependent (Akeroyd & Summerfield, 1999; Culling, Colburn, & Spurchise, 2001; Mason, Brookes, & Rumsey, 2005). For example, when the center frequency of the narrowband noise (with the absolute bandwidth of 100 Hz) is 250 Hz, human listeners can detect the occurrence of the BIC with the duration of 6.5 ms. When the center frequency becomes 1000 Hz and the absolute bandwidth is not changed, the duration threshold increases to 35 ms (Akeroyd & Summerfield, 1999). However, it should be noted that because the bandwidth of the auditory filter, such as the equivalent rectangular bandwidth (ERB), varies remarkably in the linear scale but generally remains constant in the logarithmic scale with the change

Ying Huang, Qiang Huang, Xun Chen, Xihong Wu, and Liang Li, Department of Psychology, and Department of Machine Intelligence, Speech and Hearing Research Center, Key Laboratory on Machine Perception (Ministry of Education), Peking University.

This work was supported by the National Natural Science Foundation of China (Grants 30711120563, 30670704; 60605016; 60535030; 60435010), the National High Technology Research and Development Program of China (Grants 2006AA01Z196, 2006AA010103), the "973" National Basic Research Program of China (2009CB320901), the Trans-Century Training Program Foundation for the Talents by the State Education Commission, and "985" grants from Peking University.

Correspondence concerning this article should be addressed to Liang Li, Department of Psychology, Peking University, Beijing, China, 100871. E-mail: liangli@pku.edu.cn

of the center frequency (Glasberg & Moore, 1990), using a logarithmically constant value of bandwidth is more appropriate for studying the center-frequency effect on interaural integration of fine-structure information of narrowband noises. Thus the frequency effect needs further investigation using a constant bandwidth in the logarithm scale.

One of the most intriguing questions in auditory perception is how listeners are able to detect, identify, and locate individual sound sources in noisy, reverberant environments (Bregman, 1990). To perceptually separate a target signal from a disruptive background in reverberant situations, the auditory system has to be able to differentiate sound waves of the reflections of the signal source from sound waves of other sources (which will not be as highly correlated with the direct wave from the signal). In other words, the auditory system needs to integrate the direct target sound wave with its own correlated reflections. Acoustic details of the waves directly coming from the source must be maintained for a period of time to achieve the source/reflection integration, otherwise the auditory scene will be more cluttered and confused. In fact, when the delay between the direct wave coming from the source and a reflected wave is sufficiently short, all nonspatial attributes of the reflection are perceptually captured by the direct wave (Li, Qi, He, Alain, & Schneider, 2005), leading to a single fused sound image which point of origin is perceived to be around the location of the sound source. This phenomenon is called the "precedence effect" (Blauert, 1997; Litovsky, Colburn, Yost, & Guzman, 1999; Wallach, Newman, & Rosenzweig, 1949), which plays an important role in facilitating the recognition and localization of the source in reverberant environments.

When the target source is speech and the noisy background contains competing speech, target speech is masked by two different types of masking components: (a) energetic masking, and (b) informational masking (Arbogast, Mason, & Kidd, 2002; Best, Ozmeral, Gallun, Sen, & Shinn-Cunningham, 2005; Brungart, 2001; Durlach et al., 2003; Freyman, Helfer, McCall, & Clifton, 1999; Kidd, Mason, Deliwala, Woods, & Colburn, 1994; Li, Daneman, Qi, & Schneider, 2004; Lutfi, 1990; Oxenham, Fligor, Mason, & Kidd, 2003; Shinn-Cunningham, Ihlefeld, Satyavarta, & Larson, 2005; Summers & Molis, 2004; Wu et al., 2005; Yang et al., 2007). Energetic masking occurs when peripheral neural activity elicited by target speech is overwhelmed by that elicited by maskers, leading to a degraded or noisy neural representation of the target. Due to energetic masking, a listener cannot detect some acoustic components of the target speech. Whereas, the masking effects, which cannot be explained by energetic masking (even when target and maskers have negligible spectral overlap), are generally referred to as informational masking, which makes it difficult to attend to and recognize target speech. Higher level perceptual/cognitive processes are involved in analyzing signals in the presence of an informational masker. Typically, the auditory system is unable to segregate audible components of the target speech from those of masking speech.

Masking (particularly informational masking) of target speech can be reduced if the listener can use certain cues (perceived spatial location, acoustical features, lexical information, etc.) to facilitate his/her selective attention to the target speech. However, when the listening environment is reverberant, some of the perceptual cues are limited or even abolished by reflections of sound waves (Freyman et al., 1999; Kidd, Mason, Brughera, & Hartmann, 2005; Koehnke & Besing, 1996; Zurek, Freyman, & Balakrishnan, 2004). Thus, because speech-recognition difficulties caused by maskers are augmented in reverberant environments, the perceptual integration of the direct target-speech sound wave with its own reflections becomes even more critical for segregating target speech from maskers. As mentioned above, the perceptual source/reflection integration depends on the temporal storage of acoustic details of the target source.

Previous studies have demonstrated that one of the precedence-effect components, the perceptual fusion of correlated sound waves, plays an important role in segregating target-speech signals from masker signals under simulated reverberant conditions (Brungart, Simpson, & Freyman, 2005; Freyman, Balakrishnan, & Helfer, 2001, 2004; Freyman et al., 1999; Li et al., 2004; Rakerd, Aaronson, & Hartmann, 2006; Wu et al., 2005). For example, when both the target and masker are presented by a loudspeaker to the listener's left and another loudspeaker to the listener's right, the perceived location of the target and that of the masker can be manipulated by changing the interloudspeaker interval for the target and that for the masker (Li et al., 2004). For both the target and masker, when the sound onset of the right loudspeaker leads that of the left loudspeaker by a short time (e.g., 3 ms), both a single-target image and a single-masker image are perceived as coming from the right loudspeaker. However, if the onset delay between the two loudspeakers is reversed only for the masker, the target is still perceived as coming from the right loudspeaker but the masker is perceived as coming from the left loudspeaker. The perceived colocation and perceived separation are based on perceptual integration of correlated sound waves delivered from the two loudspeakers. If the masker is speech (which causes both energetic and informational masking), the performance of recognizing target speech under the condition of perceived spatial separation is markedly better than that under the condition of perceived colocation. However, when the masker is noise (which causes energetic masking only), perceived spatial separation leads to only a slight (but significant) release (Freyman et al., 1999; Li et al., 2004; Wu et al., 2005). Thus for human listeners, the perceptual integration of the source with its reflections has evolved to be important for segregating target speech from masking speech in reverberant environments.

In a recent study (Rakerd et al., 2006), a two-speaker speech masker was presented by two spatially separated loudspeakers and the interloudspeaker time interval for the speech masker (intermasker interval) was varied in a broad range from -64 to 64 ms. At the same time the target speech was presented only by one of the two loudspeakers. When the absolute value of intermasker interval was 32 ms or shorter, there was consistent evidence of release from speech masking for target-speech recognition. However, when the intermasker interval was either -64 or 64 ms, there was no evidence of release from masking. If the masker became speech-spectrum noise, significant release occurred only at a few short intermasker intervals less than 4 ms. Thus the release of target speech from speech masking over a range of intermasker interval between 4 and 32 ms cannot be explained by a reduction in energetic masking, and perceptual integration of the leading and lagging speech maskers must play a role in reducing informational masking of target speech. More interestingly, for the masker signals, even when the loudspeaker that delivered both the target and the masker led the loudspeaker that only delivered the masker

by a time interval between 0 and 32 ms (when there was no perceived spatial separation between the target and the masker), the release was still evident, suggesting that in addition to introducing differences in perceived spatial location, introducing differences in auditory image (compactness/diffusiveness, timbre, and/or loudness) between target speech and masking speech can unmask target speech. Similarly, another study using virtual synthesis techniques (Brungart et al., 2005) also demonstrated that when the masker was one- or two-speaker speech, a significant release from masking occurred across a broad range of the intermasker interval, and when the masker was speech-spectrum noise, a significant release occurred only at a few short intermasker intervals.

It should be noted that to parse the auditory scene in a noisy, reverberant environment, perceptual integration occurs not only between correlated masking stimuli but also between the direct sound wave coming from the target source and the target reflections. Because listeners normally try to attend to target signals and ignore masking stimuli, the function of perceptually integrating target stimuli must be more important than that for masking stimuli. To our knowledge, the unmasking effect of perceptual integration of target speech with the target-reflection simulation has not been reported in the literature.

To manipulate the perceived target-masker spatial configuration with the precedence effect, it would be also important to know whether the integration of the target speech with its reflections plays a role in target-masker segregation (Nabelek & Robinette, 1978). A further important question is whether there is a functional connection between the two types of abilities. One is the ability to temporally maintain acoustic details to achieve the perceptual integration between an arbitrary noise and its delayed copy at the early auditory processing stage. The other is the ability to perceptually integrate target speech with its reflection simulation to achieve the perceptual segregation between target speech and maskers, and improve the recognition of target speech in noisy, reverberant environments.

In Experiment 1 of this study, we examined the longest IAI at which a BIC, in either wideband correlated noises or narrowband correlated noises with different center frequencies, was detectable. In Experiment 2, using the same listeners who participated in Experiment 1, we investigated whether changing the intertarget interval (ITI) in a broad range (0 to 64 ms) over two spatially separated loudspeakers can induce a release of target speech from speech masking or noise masking. Moreover, we also examined the correlation between the longest IAI and the longest ITI at which the release of speech from masking was significant.

Method

. Nineteen young university students (19 to 25 years old, 13 women) participated in this experiment. They all had normal and balanced pure-tone hearing thresholds at frequencies from 125 to 8000 Hz, confirmed by audiometry. They gave their written informed consent to participate in the experiment and were paid a modest stipend for their participation.

. Each participant was seated in a chair at the center of a sound-attenuating chamber (EMI Shielded Audiometric Examination Acoustic Suite, Beijing CA Acoustics, Beijing, China). Gaussian wideband noise, 2,000 ms in duration, including 30-ms rise-fall time, was synthesized using the "randn()" function in the MATLAB function library (the MathWorks Inc., Natick, MA) at the sampling rate of 48 kHz with 16-bit amplitude quantization. In narrowband-noise stimulation conditions, stimuli had a fixed bandwidth of 1/3 octaves and a center frequency of 200, 400, 800, 1600, or 3200 Hz. In wideband-noise stimulation conditions, Gaussian wideband noises were low-pass filtered at 10 kHz. Stimuli were transferred using the Creative Sound Blaster PCI128 (Creative SB Audigy 2 ZS, Creative Technoloy Ltd, Singapore), passed through an AURICAL system (MADSEN, Denmark), and presented to listeners by two headphones. Calibration of sound level was carried out with the Larson Davis Audiometer Calibration and Electroacoustic Testing System (AUDit and System 824, Larson Davis, Depew, NY). The sound level was set at 58 dBA SPL.

. Two 2,000-ms presentations of correlated noises were delivered over headphones. The right-headphone noise in one of the presentations was an exact copy of the left-headphone noise. The right-headphone noise in the other presentation was also identical to the left-headphone noise except for the substitution of a BIC introduced into the temporal middle of the 2,000-ms noise by simply substituting an independent noise segment to the right-headphone noise. Based on our pilot studies, the duration of the BIC was fixed at 200 ms. In each trial, the BIC was randomly assigned to one of the two presentations. The interval between the two presentations was 1,000 ms. For each presentation, noise in the

(Dynaudio Acoustics, BM6 A, Dynaudio, Denmark), which were in the frontal azimuthal plane at the left and right 45° positions with respect to the median plane. The loudspeaker height was 140 cm, which was approximately the ear level for a seated listener with average body height. The distance between the loudspeaker and the center of the participant's head was 200 cm.

Speech stimuli used in the present study were Chinese "nonsense" sentences. 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, 2001, 2004) and Li et al. (2004). Each of the Chinese nonsense sentences has three key components: subject, predicate, and object, which are also the three keywords, with two characters for each (also one syllable for each character). Note that the sentence frame did not provide any contextual support for recognition of the keywords.

Based on the database of the Chinese newspaper , published over 9 years (1994 to 2002), 6,000 double-syllable verbs, which were rated as having high frequencies of occurrence, and 12,000 double-syllable nouns, which were also rated as having high frequencies of occurrence, were selected. These words were combined randomly into 6,000 syntactically correct sentences with the frame of subject predicate object. To ensure that sentences used in experiments were not meaningful, the probability of co-occurrence of two nouns with a verb in a normal sentence was determined according to the above database. Only sentences in which probability of co-occurrence of keywords in the database was zero were used as the nonsense sentences for the present study. Because Chinese is a tonal language, further selection was made to balance syllable tones across sentences. A double-syllable pronoun was then placed before a noun, and an auxiliary verb was placed before a verb, making a selected sentence more natural. Finally, all sentences were examined by experimenters to ensure that selected sentences were nonsensical.

Eighteen lists (18 sentences/list) of nonsense sentences were used as target sentences. To balance information quantity across experimental conditions in this study, the information quantity of a keyword in a sentence was calculated as

$$= \log\left(\frac{1}{}\right),$$

where is word frequency. Information quantity of a sentence was the sum of information quantities of the three keywords. All the lists of nonsense sentences were constructed in such a way that the information quantity of each list was about the same. Target speech was spoken by a young female speaker (Speaker A).

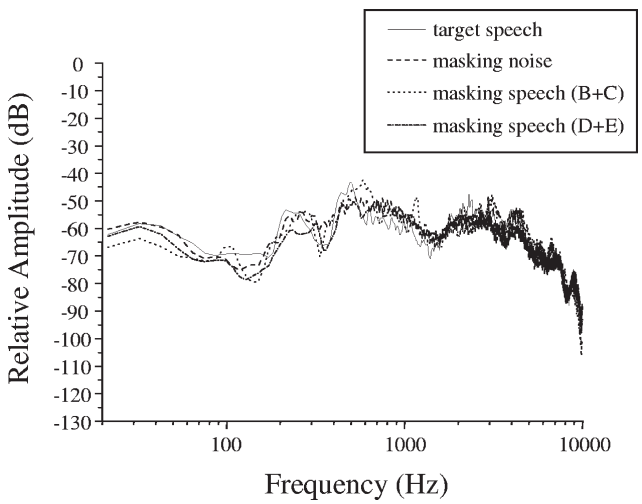
The speech masker presented from the left loudspeaker was a 47-s loop of digitally combined continuous recordings for Chinese nonsense sentences (which keywords did not appear in target sentences) spoken by two different young female speakers (Speakers B and C). The speech masker presented from the right loudspeaker was also a 47-s loop of digitally combined continuous recordings of Chinese nonsense sentences (which keywords did not appear in target sentences also) spoken by another two young female speakers (Speakers D and E). Each of the four masking speakers spoke different sentences and the sound pressure levels were the same across the four masking speakers' speech sounds within a testing session. In a trial, a speech masker started from a different point in the loop; therefore, the loop for the left loudspeaker was not in synchrony with that for the right loudspeaker on a trial-by-trial basis.

A noise masker was a stream of steady state speech-spectrum noise. The development of the noise masker is described in Yang et al. (2007). To estimate the differences in the spectrum between the noise masker, left-loudspeaker speech masker, right-loudspeaker speech masker, and target speech, a Knowles Electronic Manikin for Acoustic Research (KEMAR, Knowles Electronics, Itasca, IL) was located at the position of a participant in the anechoic chamber. Each of the stimuli was delivered by the right loudspeaker and sound waves were recorded using the right ear of the KEMAR. The spectra under these stimulus conditions, as presented in Figure 1, were very similar.

All speech stimuli were calibrated using a B&K sound level meter (Type 2230, Bruel & Kjaer, Denmark) where the microphone was placed at the central location of the listener's head when the listener was absent, using a "slow"/"RMS" meter response. During a session, the target-speech sounds were presented at a level so that each loudspeaker, playing alone, would produce a sound pressure of 56 dBA. The target-speech sound pressure level remained constant throughout the experiment. The sound pressure levels of the maskers were adjusted to 64 dBA to produce the signal-to-noise ratio (SNR) of −8 dB.

For the two-source target presentation, the two loudspeakers presented the identical target sentences, but the right loudspeaker either led or lagged behind the left loudspeaker by 0, 0.5, 1, 2, 4, 8, 16, 32, or 64 ms. In this study, positive ITI values were used to stand for conditions when the left loudspeaker led the right loudspeaker for target presentation and negative ITI values for conditions when the left loudspeaker lagged behind the right loudspeaker for target presentation. For the single-source target presentation, only the right loudspeaker presented target speech.

For the two-source masker presentation, the two loudspeakers presented either different two-speaker speech maskers (both speakers



. Comparison of the spectra between target speech, masking speech spoken by Speakers B and C, masking speech spoken by Speakers D and E, and masking noise. Each of the stimuli was presented by the right loudspeaker, and sound waves were recorded using the right ear of the Knowles Electronic Manikin for Acoustic Research (KEMAR).

and contents were different between the two loudspeakers) or independent speech-spectrum noises at the same time. For the single-source masker presentation, only the right loudspeaker presented either the two-speaker speech masker or the noise masker.

. Eighteen target sentences were used in each condition. The order of presenting masker types was counterbalanced across 18 participants. The order of ITI was arranged in a random manner. In each trial, the participant pressed a button of the response box to start the masker. About 1 s later, a single target sentence was presented along with the masker and then the masker was gated off with the target. The participant was instructed to loudly repeat the entire target sentence toward a microphone as best as he/she could immediately after the sounds were completed. The experimenters (the authors), sitting outside the anechoic chamber and listening to the participant's responses via a loudspeaker, indicated on a marking sheet whether each syllable in each keyword had been identified correctly by the participant. The number of correctly identified syllables in keywords was tallied later.

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

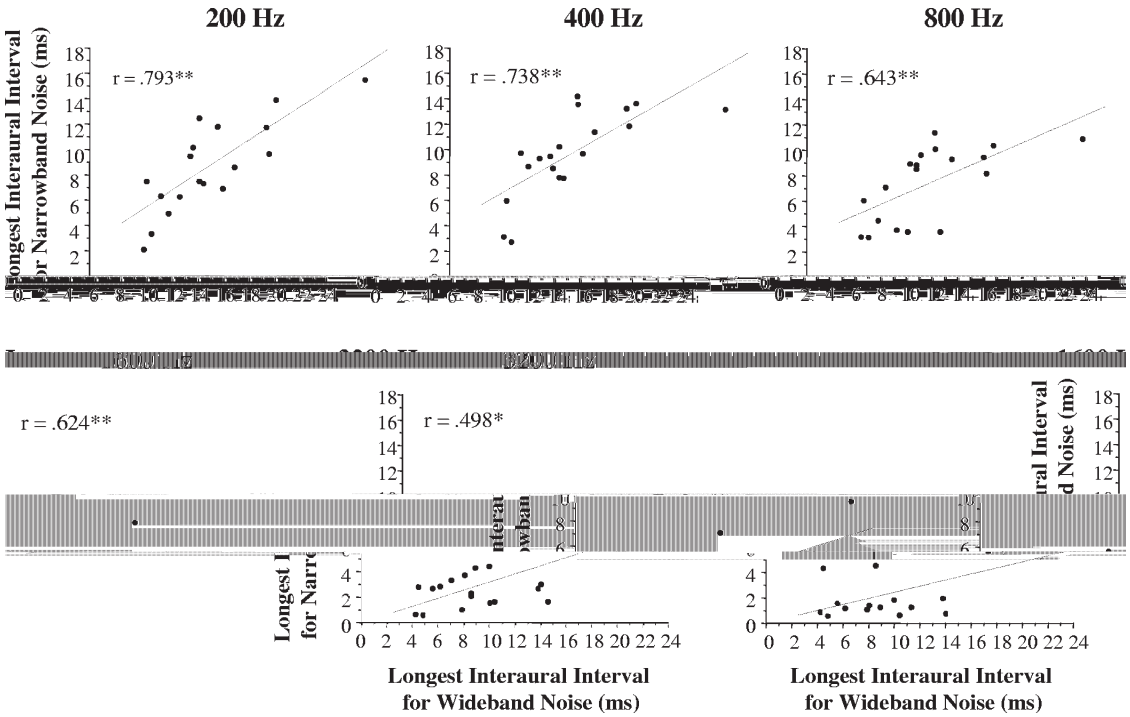
Results

Figure 2 shows the longest IAI at which each participant could detect the 200-ms BIC in narrowband noises with various center frequencies as a function of the longest IAI at which each participant could detect the BIC in wideband noises. All participants could detect the BIC under each noise types, except three participants could not detect the BIC when the center frequency of narrowband noise was at 3200 Hz. As shown in Figure 2, the longest IAI under each condition varied markedly across participants. Figure 2 also shows that participants generally performed better when the center frequency of narrowband noise was lower than when it was higher. This suggests that the temporal storage of low-frequency acoustic details lasted longer than high-frequency acoustic details.

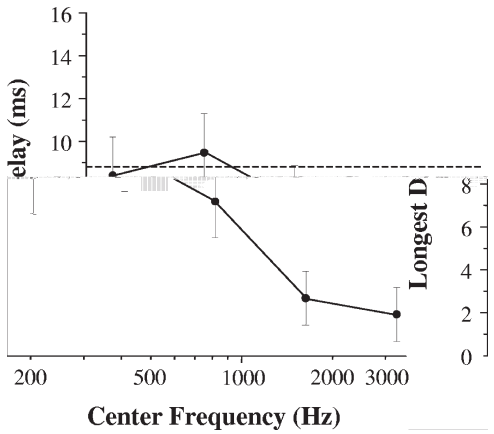
To explore whether the variability of the longest IAI for wideband noises across participants was correlated with that for narrowband noises, correlation coefficients were calculated between the longest IAI for wideband noises and those for narrowband noises (Figure 2). Significant correlations were found in all the five pairs, and the correlation coefficient decreased as the center frequency of narrowband noises increased. Thus, the persistence of the temporal storage of wideband details was contributed more by low-frequency components than by high-frequency components.

The center-frequency effect for narrowband noises can also be found in the group mean of the longest IAIs at various center frequencies (Figure 3). Clearly, participants were able to detect the BIC over longer IAIs when the center frequency was low (200, 400, or 800 Hz) than when it was high (1600 or 3200 Hz).

One-way within-participant analysis of variance (ANOVA) shows that the effect of noise type was significant, (5, 75) − 40.189, ◄ .001, ^{2} − .728. Bonferroni post hoc analyses with



. The longest interaural interval (IAI) at which a 200-ms break in correlation could be detected for each of five narrowband noises as a function for that wideband noise. The dotted straight line in each panel is the best-fitting line for the data points, and represents the correlation coefficient between the longest IAI for the narrowband noise and for that wideband noise. * − .05. ** − .01.



. The group mean of longest interaural interval (IAI) at which the break in correlation in the narrowband noise could be detected as a function of the center frequency. The broken line represents the longest IAI when the noise was wideband. The error bars represent the standard errors of the mean.

the significant level set at .05 indicate that the longest IAI for wideband noises was not significantly different from that for narrowband noises with the center frequency of 200, 400, or 800 Hz, but significantly longer than that for narrowband noises with the center frequency of 1600 or 3200 Hz. The longest IAI for the center frequency of 1600 or 3200 Hz was significantly shorter than that for the center frequency of 200, 400, or 800 Hz. The longest IAI for the center frequency of 200 Hz was not significantly different from that for the center frequency of 400 or 800 Hz. The longest IAI for the center frequency of 400 Hz was significantly longer than that for the center frequency of 800 Hz. The longest IAI for the center frequency of 1600 Hz was not significantly different from that for the center frequency of 3200 Hz.

The top panels of Figure 4 show the percentage correct identification of keyword syllables as a function of the ITI when the masker was speech (left panel) or speech-spectrum noise (right panel). Obviously, both masker type and ITI influenced the recognition of target speech. Percentage correct identification under the single source-presentation condition is also shown in Figure 4 as the broken line.

Under single source-presentation conditions, the noise masker caused a larger masking effect than the speech masker, $(1, 17) = 29.309$, $< .001$, $^2 = .633$. Under two-source-presentation conditions, participants recognized more keyword syllables under short ITIs than under long ITIs. With the change of the ITI from 64 to 0 ms, the correct percentage speech identification increased progressively. A 2 (masker type) 17 (ITI) within-participant ANOVA shows that the main effect of masker type was significant, $(1, 17) = 167.424$, $< .001$, $^2 = .908$; the main effect of ITI was significant, $(16, 272) = 51.711$, $< .001$, $^2 = .753$ the interaction between the two factors was significant $(16, 272) = 20.010$, $< .001$, $^2 = .541$.

As shown in the top panels of Figure 4, the mean percentage correct identification of target speech was similar between the left-loudspeaker leading condition and right-loudspeaker leading condition for each masker type. One-way within-participant ANOVAs and Bonferroni post hoc analyses with the significant level set at .05 show that at each of the ITIs there was no significant difference between the two leading directions for both speech- and noise-masking conditions. Thus, to evaluate the releasing effect of shortening the ITI, the mean percentage correct identification between the two leading conditions was averaged for each ITI.

When the delay between a speech sound and its copy becomes 50 ms or more, no precedence effects occur (Blauert, 1997, p. 226; also see the discussion by Rakerd et al. 2006). In other words, a target speech cannot be perceptually integrated with its reflection simulation when the IAI is 64 ms. Also, in the present study, the sound pressure level of each loudspeaker was not changed with the reduction of the ITI from 64 to 0 ms under both speech-masking conditions and noise-masking conditions. Thus it is reasonable to use the performance at the longest ITI (64 ms) as the baseline performance for two-source presentation conditions. Based on the averaged left–right percentage correct recognition, the release of speech from masking at an ITI is defined as the difference between the percentage of correct speech recognition at the ITI and the percentage of correct speech recognition at the ITI of 64 ms. The middle panels of Figure 4 show the percentage releases of target speech as a function of the absolute value of ITI under the speech-masking condition (left panel) or under the noise-masking condition (right panel). Obviously, the release increased with the decrease of the absolute value of ITI, but it was much larger under the speech-masking condition than under the noise-masking condition. A 2 (masker type) 9 (absolute value of ITI) within-participant ANOVA shows that the main effect of masker type was significant, $(1, 17) = 109.349$, $< .001$, $^2 = .865$; the main effect of ITI was significant $(8, 136) = 111.520$, $< .001$, $^2 = .868$; the interaction between the two factors was significant $(8, 136) = 37.205$, $< .001$, $^2 = .686$. One-way within-participant ANOVA was conducted separately under the speech-masking condition and under the noise-masking condition, and Bonferroni post hoc analyses with the significant level set at .05 show that under either speech- or noise-masking conditions, the group mean of the release was significant when the ITI was 32 ms or shorter.

The bottom panels of Figure 4 show the percentage release as a function of ITI for individual participants under speech-masking conditions (left panel) or noise-masking conditions (right panel). Clearly, there was a remarkable variability in the release across participants, particularly under speech-masking conditions. The critical ITI for individuals is defined as the longest ITI (among those used in the experiment) at which the correct recognition of keyword syllables was significantly better than the correct recognition of keyword syllables at the ITI of 64 ms. In this experiment, critical ITIs could be obtained in each of the participants under speech-masking conditions. For individual participants under noise-masking conditions, however, significant differences in performance between the ITI of 64 ms and any other ITIs could not be obtained, except only for one participant whose performance at the ITI of 0.5 ms was significantly different from that at the ITI of 64 ms. Thus we could not obtain reliable critical ITIs for individuals under noise-masking conditions.

Correlation coefficients were calculated between the longest IAI (for wideband and narrowband noises, Experiment 1) and the critical ITI under speech-masking conditions (Experiment 2) across the same 18 participants. The correlation coefficients are shown in Figure 5. For the two narrowband noises with the low center frequencies (200

Hz, 400 Hz), significant correlations (200 Hz: − .041; 400 Hz: − .005) were obtained between the longest IAI and the critical ITI. When the center frequency of the narrowband noise was 800 Hz or higher, no significant correlations were found (.05). Even for wideband noise, the correlation was not significant (− .075).

Discussion

details of the acoustic waveform is considerably different between listeners. The persistence of the central representation of interaurally correlated sounds also has been estimated previously using indirect measures (Blodgett, Wilbanks, & Jeffress, 1956; Cherry & Taylor, 1954; Langford & Jeffress, 1964; Mossop & Culling, 1998). Results of these early studies suggested that a representation of the waveform may persist for up to 9 to 15 ms.

The large interlistener variability in detecting the BIC under zero ITD was reported by previous studies. For example, in the study by Akeroyd and Summerfield (1999) the interlistener differences were quantified as the coefficient of variation: the across-listener standard deviation divided by the across-listener mean. The mean values for detecting the BIC (so-called binaural gap in their study) were 0.83 (under the "center frequency" conditions) or 0.76 (under the "lower cutoff" conditions) compared with values of 0.11 (under the center frequency conditions) and 0.12 (under the low-cutoff conditions) for detecting the monaural gap. Also, the great variability of binaural temporal window across listeners was reported by Boehnke et al. (2002). Thus, the large interparticipant variability in the longest IAI in the present study might be partially associated with the great variability of binaural temporal window across participants.

On the other hand, the interlistener variability in binaural processing is still marked even when an interaural delay is introduced (e.g., Blodgett et al., 1956; Mossop & Culling, 1998). For example, in the study by Blodgett et al., participants who had experience in experiments on sound localization and on the masking of tones by noise, were instructed to report the sidedness of binaurally presented identical (correlated) noises when an IAI was introduced. The results show that the sidedness was maintained (correlated noises were distinguishable from uncorrelated noises) even when the IAI was up to 20 ms. The IAI was greater with the wideband noise than with the narrowband noise, and greater with noise bands of low frequency than with bands of high frequency. Particularly, the maximal delay values varied widely from participant to participant, ranging from 7.5 to 20.7 ms for a lower frequency band (106 to 212 Hz) and from 2.5 to 14.2 ms for a higher frequency band (2400 to 4800 Hz). Blodgett et al. proposed that listeners who excelled in ability to respond to long delays also excelled in ability to make precise localizations 7f m(n229.8(in)-raural)-2JT*[(thee

over a broad range, and the improvement was much larger under the speech-masking condition than under the noise-masking condition. Particularly, the critical ITI for each of the participants was obtained under speech-masking conditions but it could not be obtained for individual participants under noise-masking conditions. Chiang and Freyman study (1998) reported that when a leading sound was delivered from a loudspeaker at 45° to the right of center and a lagging sound from 45° left, presenting background noise substantially reduced both the dominance of the leading sound on perceived location and the echo threshold for fusing the leading and lagging sounds. To our knowledge, whether there is a difference between speech masker and noise masker in weakening source-reflection integration has not been reported in the literature. However, in the present study, even when the ITI was 0 ms at which the maximum integration of target speech was achieved, the group mean release of target speech was much lower under the noise-masking condition than under the speech-masking condition. These results indicate that temporal integration of target speech with its reflection simulation increases with the reduction of ITI and predominately facilitates the release of target speech from informational masking.

Certain manipulations, as long as they help distinguish the target speech from the masker and direct selective attention toward target speech, will release target speech from masking, especially from informational masking (Brungart, 2001; Freyman et al., 2004; Kidd et al., 2005; Li et al., 2004; Yang et al., 2007). In this study, although there was no physical separation between the target and the masker that was presented from the same loudspeaker, manipulation of the ITI could cause certain changes of the perceived image of the target speech in compactness, loudness, timbre, and/or spatial location. Using one or some of these cues available at certain ITIs helped participants perceptually segregate target speech from masker, especially from the speech masker, thereby improving selective attention to target speech. Obviously, the ITI-dependent release of target speech from informational masking is based on the perceptual integration between the leading and lagging target-speech signals, and this perceptual integration is, in turn, based on the temporal storage of acoustic details of the leading target-speech signal.

A highly reverberant environment can significantly reduce the head shadow advantage and obscure IAI differences, thereby significantly reducing the spatial separation effect on releasing targets from energetic masking (e.g., Freyman et al. 1999; Kidd et al., 2005; Koehnke & Bessing, 1996; Zurek et al., 2004). However, when the target is the speech of one speaker and the masker is the speech of another speaker(s), spatial separation of the sources can still improve identification of the target through the perceptual segregation of sound images. Thus spatial separation under high reverberation can be used for separating the unmasking factors other than head shadowing and binaural processing (Kidd et al., 2005). Some studies, in which reverberation was simulated by presenting target speech and masker with two spatially separated loudspeakers, have confirmed that precedence-induced perceived spatial separation mainly releases target speech from informational masking. However, it should be noted that perceived spatial separation also produces significant release of speech from steady state speech-spectrum noise (Li et al., 2004; Wu et al., 2005; also see Freyman et al., 1999). Further investigation is still needed to verify whether the difference between the release from speech maskers and that from noise maskers can be used for estimating the "pure" informational masking effect.

[illegible]

The present study shows that there was a remarkable across-participant variability in the longest IAI (Experiment 1) and in the release of speech from speech masking (Experiment 2), and the critical ITI was significantly correlated with the longest IAI only for the two low-frequency narrowband noises (200 Hz, 400 Hz). When the center frequency of the narrowband noise was 800 Hz or higher, no significant correlation between the longest IAI and the critical ITI was found. Even for wideband noise, the correlation was not significant. These results suggest that the ability of temporal storing of low-frequency fine-structure information, which is in the range of the fundamental frequency and the first formant of female voices, is functionally associated with the ability of temporal integrating acoustic fine structures of female-voice speech for releasing speech from informational masking in reverberant environments. Thus, if a listener has a longer temporal storage of low-frequency acoustic details, she/he may have a better chance to correctly recognize the target speech in noisy, reverberant environments.

The perceptual integration of correlated sounds is sound-type dependent. For example, the echo threshold of the precedence effect varies depending on stimulus type. Previous studies have shown that for speech sounds the echo threshold can be as long as tens of milliseconds (e.g., Rakerd, Hartmann, & Hsu, 2000), which is much longer than that (about 4 ms) when stimuli are bursts of white noise (e.g., Roberts & Lister, 2004). For speech sounds, both the present study and previous studies (Brungart et al., 2005; Rakerd et al., 2006) indicated that perceptual integration can occur at the delay up to 32 ms. However, for noise sounds, the present study shows that the longest IAI for detecting the BIC was less than 20 ms for most of the participants. The significant correlation between the longest IAI for integrating low-frequency noises and the critical ITI for reducing speech masking suggests that (a) the longest IAI for detecting the BIC in noises does not perfectly reflect the real auditory persistence for various types of sounds, such as spectrum- and amplitude-modulated speech sounds, but (b) the ability to temporally maintain low-frequency fine acoustic details contributes to the perceptual integration of various sounds containing similar low-frequency components, such as speech sounds. In the future, the interaction between lower level auditory processes and higher level speech processes in noisy, reverberant environments needs further investigation.

[illegible]

The temporal storage of acoustic details for a short time (up to over 20 ms) must occur at the pre-attentive stage because listeners cannot consciously perceive individual components of acoustic details. Moreover, because the content of this type of primitive auditory memory is fine-structure information, this type of information storage must have a huge capacity. Furthermore, because the primitive auditory memory is critically functional in noisy, reverberant environments, it must be tolerant to disruptive stimuli.

It remains to be investigated how such primitive auditory memory of acoustic details develops during the individual's life span, whether it can be modified by experience, where the underlying neural circuits are located (both fMRI and MEG studies suggest an involvement of the auditory cortex in processing changes of interaural correlation [Budd et al., 2003; Chait, Poeppel, Cheveigne, & Simon, 2005]), and in particular, whether this type of memory is critical for subsequent high-level auditory processing. The present study provides only a start.

Note that the primitive auditory memory investigated in the present study is different from the transient auditory memory (echoic memory) as investigated by the mismatch negativity (MMN) of event-related potentials (Näätänen & Winkler, 1999; Ritter, Deacon, Gomes, Javitt, & Vaughan, 1995; Tiitinen, May, & Reinikalnen, 1994). The MMN-probed auditory memory can last up to 10 s, and in some circumstances, be of a long-term nature. Thus, we hypothesize that the primitive auditory memory of acoustic details is at the early end of the chain of the transient auditory memory system, and the memory probed by MMN is at the late end.

Summary

Listeners can detect the 200-ms BIC between two correlated wideband noises (0 to 10 kHz) even when the IAI is up to 21.5 ms, indicating that the temporal storage of detailed acoustic information can last over 20 ms in some listeners. However, in some other listeners with normal hearing, this temporal extent is reduced to the level of no more than 5 ms.

The temporal storage of fine-structure information is frequency dependent. The storage of low-frequency details lasts longer than that of high-frequency details.

Under the reverberation-simulation condition with the presentation of the speech masker, the reduction of the ITI from 64 to 0 ms progressively improves the recognition of target speech. Under noise-masking conditions, however, the improvement is minor. Thus, the reduction of the ITI enhances the temporal integration between target speech with its reflections and predominantly releases target speech from informational masking. There is also a remarkable variability between listeners in the masking release.

The ability of temporal integrating speech with its reflections is functionally associated with the ability of temporal storing of low-frequency acoustic details. Thus, the primitive auditory memory, which occurs at the early stage of the transient auditory memory system, is critical for later high-level segregating target speech from informational masking in noisy, reverberant environments.

References

Akeroyd, M. A., & Summerfield, A. Q. (1999). A binaural analog of gap detection. , , 2807–2820.

Arbogast, T. L., Mason, C. R., & Kidd, G. (2002). The effect of spatial separation on informational and energetic masking of speech. , , 2086–2098.

Bernstein, L. R., & Trahiotis, C. (1999). The effects of signal duration on NoSo and NoS pi thresholds at 500 Hz and 4 kHz. , , 1776–1783.

Best, V., Ozmeral, E., Gallun, F. J., Sen, K., & Shinn-Cunningham, B. G. (2005). Spatial unmasking of birdsong in human listeners: Energetic and informational factors. , , 3766–3773.

Blauert, J. (1997). . Cambridge, MA: MIT Press.

Blauert, J., & Lindemann, W. (1986). Spatial-mapping of intracranial auditory events for various degrees of interaural coherence. , , 806–813.

Blodgett, H. C., Wilbanks, W. A., & Jeffress, L. A. (1956). Effect of large interaural time differences upon the judgment of sidedness. , , 639–643.

Boehnke, S. E., Hall, S. E., & Marquardt, T. (2002). Detection of static and dynamic changes in interaural correlation. , , 1617–1626.

Bregman, A. S. (1990). . Cambridge, MA: MIT Press.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. , *0* , 1101–1109.

Brungart, D. S., Simpson, B. D., & Freyman, R. L. (2005). Precedence-based speech segregation in a virtual auditory environment. , , 3241–3251.

Budd, T. W., Hall, D. A., Goncalves, M. S., Akeroyd, M. A., Foster, J. R., Palmer, A. R., et al. (2003). Binaural specialization in human auditory cortex: An fMRI investigation of interaural correlation sensitivity. , *0,* 1783–1794.

Chait, M., Poeppel, D., Cheveigne, A. D., & Simon, J. Z. (2005). Human auditory cortical processing of changes in interaural correlation. , , 8518–8527.

Cherry, E. C., & Taylor, W. K. (1954). Some further experiments upon the recognition of speech, with one and with two ears. , , 554–559.

Chiang, Y. C., & Freyman, R. L. (1998). The influence of broadband noise on the precedence effect. , *04,* 3039–3047.

Culling, J. F. (2007). Evidence specifically favoring the equalization-cancellation theory of binaural unmasking. , , 2803–2813.

Culling, J. F., Colburn, H. S., & Spurchise, M. (2001). Interaural correlation sensitivity. , *0,* 1020–1029.

Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., & Kidd, G. (2003). Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity. , *4,* 368–379.

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2001). Spatial release from informational masking in speech recognition. , *0* , 2112–2122.

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. , , 2246–2256.

Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. , *0* , 3578–3588.

Gabriel, K. J., & Colburn, H. S. (1981). Interaural correlation discrimination: I. Bandwidth and level dependence. , , 1394–1401.

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapers from notched-noise data. , *4* , 103–138.

Goupell, M. J., & Hartmann, W. M. (2006). Interaural fluctuations and the detection of interaural incoherence: Bandwidth effects. , , 3971–3986.

Helfer, K. S. (1997). Auditory and auditory-visual perception of clear and

conversational speech. , *40,* 432–443.

Huang, Y., Kong, L.-Z., Fan, S.-L., Wu, X.-H., & Li, L. (2008). Both frequency and interaural delay affect ERP responses to binaural gap. , , 1673–1678.

Kidd, G., Jr., Mason, C. R., Brughera, A., & Hartmann, W. M. (2005). The role of reverberation in release from masking due to spatial separation of sources for speech identification. , , 526–536.

Kidd, G., Jr., Mason, C. R., Deliwala, P. S., Woods, W. S., & Colburn, H. S. (1994). Reducing informational masking by sound segregation. , , 3475–3480.

Koehnke, J., & Besing, J. M. (1996). A procedure for testing speech intelligibility in a virtual listening environment. , 211–217.

Langford, T. L., & Jeffress, L. A. (1964). Effect of noise crosscorrelation on binaural signal detection. , , 1455–1458.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. , *4* , 467–477.

Li, L., Daneman, M., Qi, G. Q., & Schneider, B. A. (2004). Does the information content of an irrelevant source differentially affect speech recognition in younger and older adults? , *0,* 1077–1091.

Li, L., Qi, J. G., He, Y., Alain, C., & Schneider, B. A. (2005). Attribute capture in the precedence effect for long-duration noise sounds. , *0* , 235–247.

Litovsky, R. Y., Colburn, H. S., Yost, W. A., & Guzman, S. J. (1999). The precedence effect. , *0* , 1633–1654.

Lutfi, R. A. (1990). How much masking is informational masking? , , 2607–2610.

Mason, R., Brookes, T., & Rumsey, F. (2005). Frequency dependency of the relationship between perceived auditory source width and the interaural cross-correlation coefficient for time-invariant stimuli. , , 1337–1350.

Mossop, J. E., & Culling, J. F. (1998). Lateralization of large interaural delays. , *04,* 1574–1579.

Näätänen, R., & Winkler, I. (1999). The concept of auditory stimulus representation in cognitive neuroscience. , , 826–859.

Nabelek, A. K., & Robinette, L. (1978). Influence of precedence effect on word identification by normally hearing and hearing-impaired subjects. , , 187–194.

Oxenham, A. J., Fligor, B. J., Mason, C. R., & Kidd, G. (2003). Informational masking and musical training. , *4,* 1543–1549.

Pollack, I., & Trittipoe, W. J. (1959). Binaural listening and interaural noise cross correlation. , , 1250–1252.

Rakerd, B., Aaronson, N. L., & Hartmann, W. M. (2006). Release from speech-on-speech masking by adding a delayed masker at a different location. , , 1597–1605.

Rakerd, B., Hartmann, W. M., & Hsu, J. (2000). Echo suppression in the horizontal and median sagittal planes. , *0* , 1061–1064.

Ritter, W., Deacon, D., Gomes, H., Javitt, D. C., & Vaughan, H. G. (1995). The mismatch negativity of event-related potentials as a probe of transient auditory memory: A review. , , 52–67.

Roberts, R. A., & Lister, J. J. (2004). Effects of age and hearing loss on gap detection and the precedence effect: Broadband stimuli. , , *4* , 965–978.

Shinn-Cunningham, B. G., Ihlefeld, A., Satyavarta, & Larson, E. (2005). Top-down and bottom-up influences on spatial unmasking. , , 967–979.

Summers, V., & Molis, M. R. (2004). Speech recognition in fluctuating and continuous maskers: Effects of hearing loss and presentation level. , *4* , 245–256.

Tiitinen, H., May, P., & Reinikalnen, K. (1994). Attentive novelty detection in humans is governed by pre-attentive sensory memory. , , 90–92.

Wallach, H., Newman, E. B., & Rosenzweig, M. R. (1949). The precedence effect in sound localization. , , 315–336.

Wu, X.-H., Wang, C., Chen, J., Qu, H.-W., Li, W.-R., Wu, Y.-H., et al. (2005). The effect of perceived spatial separation on informational masking of Chinese speech. , , 1–10.

Yang, Z.-G., Chen, J., Wu, X.-H., Wu, Y.-H., Schneider, B. A., & Li, L. (2007). The effect of voice cuing on releasing Chinese speech from informational masking. , *4* , 892–904.

Zurek, P. M., Freyman, R. L., & Balakrishnan, U. (2004). Auditory target detection in reverberation. , , 1609–1620.

Received April 30, 2007
Revision received November 26, 2008
Accepted December 8, 2008 ■