Perceptual Fusion Tendency of Speech Sounds

Ying Huang, Jingyu Li, Xuefei Zou, Tianshu Qu, Xihong Wu, Lihua Mao, Yanhong Wu, and Liang Li

Abstract

■ To discriminate and to recognize sound sources in a noisy, reverberant environment, listeners need to perceptually integrate the direct wave with the reflections of each sound source. It has been confirmed that perceptual fusion between direct and reflected waves of a speech sound helps listeners recognize this speech sound in a simulated reverberant environment with disrupting sound sources. When the delay between a direct sound wave and its reflected wave is sufficiently short, the two waves are perceptually fused into a single sound image as coming from the source location. Interestingly, compared with nonspeech sounds such as clicks and noise bursts, speech sounds have a much larger perceptual fusion tendency. This study investigated why the fusion tendency for speech sounds is so large. Here we show that when the temporal amplitude fluctuation of speech was artificially time reversed, a large perceptual fusion tendency of speech sounds disappeared, regardless of whether the speech acoustic carrier was in normal or reversed temporal order. Moreover, perceptual fusion of normal-order speech, but not that of time-reversed speech, was accompanied by increased coactivation of the attention-control-related, spatial-processing-related, and speech-processing-related cortical areas. Thus, speech-like acoustic carriers modulated by speech amplitude fluctuation selectively activate a cortical network for top–down modulations of speech processing, leading to an enhancement of perceptual fusion of speech sounds. This mechanism represents a perceptual-grouping strategy for unmasking speech under adverse conditions. ■

INTRODUCTION

Investigation of how various sound sources (including speech sounds) are discriminated and a target source is correctly recognized in a noisy, reverberant environment is critical for reaching the understanding of why the human brain is able to extract target information under conditions with sensory-input "flooding." It has been confirmed that to distinguish signals from various speech sources and to correctly recognize the target speech source in a simulated reverberant environment, listeners need to not only perceptually integrate the direct wave with the reflections of the target speech source (Huang, Huang, Chen, Wu, & Li, 2009; Huang, Huang, et al., 2008) but also perceptually integrate the direct wave with the reflections of the masking speech source (Rakerd, Aaronson, & Hartmann, 2006; Brungart, Simpson, & Freyman, 2005).

When the delay between a leading sound (such as the direct wave from a sound source) and a correlated lagging sound (such as a reflection of the direct wave) is sufficiently short, attributes of the lagging sound are perceptually captured by the leading sound (Li, Qi, He, Alain, & Schneider, 2005), causing a single fused sound image from a location near the leading source (the precedence effect, see Litovsky, Colburn, Yost, & Guzman, 1999; Freyman, Clifton, & Litovsky, 1991; Zurek, 1980; Wallach, Newman, & Rosenzweig, 1949). This perceptual fusion is able to produce *perceived* spatial separation between uncorrelated sound sources, and the *perceived* spatial separation plays a role in reducing masking for speech recognition (Huang, Huang, et al., 2008, 2009; Rakerd et al., 2006; Wu et al., 2005; Li, Daneman, Qi, & Schneider, 2004; Freyman, Helfer, McCall, & Clifton, 1999). For example, when both the target and the masker are presented by a loudspeaker to the listener's left and by 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). More specifically, for both the target and the masker, when the sound onset of the right loudspeaker leads that of the left loudspeaker by a short time (e.g., 3 msec), both a single target image and a single masker image are perceived by the human listener 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 the perceived separation are based on perceptual integration of correlated sound waves delivered from each of the two loudspeakers. It has been confirmed that perceived target-masker spatial separation facilitates the listener's selective attention to target signals and significantly improves recognition of target signals, although neither the masker

Peking University, Beijing, China

energy at each ear nor the stimulus-image compactness/diffusiveness is substantially changed (e.g., Li et al., 2004).

The minimum delay allowing a listener to perceive the lagging sound as a discrete echo (when the perceptual fusion is just broken) is called the echo threshold (Litovsky et al., 1999; Freyman et al., 1991), which indicates the perceptual fusion tendency (i.e., a large echo threshold suggests a large perceptual fusion tendency). Interestingly, speech sounds, which are the most important acoustic stimuli for human communication and contain distinct patterns of periodicities and transients, have much larger echo thresholds than other types of sounds such as clicks and noise bursts (Rakerd, Hartmann, & Hsu, 2000; Litovsky et al., 1999; Lochner & Burger, 1958; Cherry & Taylor, 1954; Wallach et al., 1949).

This study investigated why the tendency of the perceptual fusion of speech sounds is so large. Because the temporal properties of speech sounds are critical for achieving communication (Rosen, 1992), the focus of this study was particularly on the role played by the overall temporal amplitude fluctuation (TAF) of speech in maintaining the perceptual fusion tendency and in eliciting perceptual-fusion-related cortical activations.

The TAF of Chinese speech was extracted using the Hilbert transform as previously used by other investigators (Zeng et al., 2005; Smith, Delgutte, & Oxenham, 2002). When the speech TAF is artificially removed from speech, the remaining acoustic component is called the speech acoustic carrier (AC), which still contains some speech acoustic features including harmonic structures, frequency modulation, and periodic occurrences of noise-like consonants. When the TAF and the AC are artificially separated, they can be time reversed either independently or at the same time. Any of the temporal reversals does not change the long-term spectrum, the fundamental frequency, and other spectrotemporal characters, including the overall spectral modulation density, the temporal modulation frequency, and the modulation amplitude. The modified speech sound is still "speech-like" but has reduced or no semantic content.

EXPERIMENT 1

Methods

Participants

Eighteen university students (12 women and 6 men, 19–28 years old, mean age = 23 years) participated in this experiment. In this and the following three experiments (Experiments 2–4), each participant participated in only one experiment of this study, and all the participants had normal (no more than 25 dB) and balanced (no more than 15 dB difference between the two ears) pure-tone hearing thresholds at frequencies from 0.125 to 8 kHz. They all gave their written informed consent to participate in the experiment and were paid a modest stipend for their participation.

Apparatus

During a testing session, the participant was seated at the center of an anechoic chamber (Beijing CA Acoustics, Beijing, China), which was 560 cm in length, 400 cm in width, and 193 cm in height. Acoustic signals were digitized using the 24-bit Creative Sound Blaster PCI128 (which had a built-in anti-aliasing filter) (Creative Technology, Ltd., Singapore) and audio editing software [Cooledit Pro 2.0 (Syntrillium Software Corp., Phoenix, AZ)]. The analog outputs were delivered to two loudspeakers (Dynaudio Acoustics, BM6A; Risskov, Denmark) 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 seated listener's head was 200 cm.

Stimuli

Speech files used in this study were Chinese "nonsense" sentences spoken by a young female speaker. These sentences are syntactically correct but not semantically meaningful. Direct English translations of the sentences are similar but not identical to the English nonsense sentences that were developed by Helfer (1997) and also used in studies by Freyman, Balakrishnan, and Helfer (2001, 2004), Freyman et al. (1999), and Li et al. (2004). For example, the English translation of one Chinese nonsense sentence is "These war situations continually look into the workshop." The development of the Chinese nonsense sentences is described in detail by Yang et al. (2007) and Wu et al. (2005). The duration of these nonsense sentences ranges from 2400 to 2700 msec. To obtain spectrum-matched steady-state noises, which had the amplitude spectrum of the speech (except for a scale factor) and randomized phase, the long-term, complex spectrum of a speech sentence spoken by the female talker was multiplied by the long-term, complex spectrum of a wideband Gaussian noise, and the consequence was inverse-fast-Fourier transformed (Arbogast, Mason, & Kidd, 2002, 2005). The duration of the spectrum-matched noise was equal to that of the speech sentence, including 30-msec rise–fall times.

Three types of sounds with the same long-term spectrum were used in this experiment: (a) spectrum-matched steady-state noises; (b) AC of normal-order speech, which was produced by replacement of the TAF of the normal-order speech with the TAF of the spectrum-matched steady-state noise; and (c) AC of time-reversed speech, which was produced by replacement of the TAF of the time-reversed speech with the TAF of the spectrum-matched steady-state noise.

In this and in the following psychophysical experiments (Experiments 1–3), TAFs for the following three types of stimuli were extracted with the Hilbert transform (Oppenheim, Schafer, & Buck, 1999) as used by

previous investigators (Zeng et al., 2005; Smith et al., 2002): spectrum-matched steady-state noise, normal-order speech, and time-reversed speech. To obtain the normal-order speech AC, the normal-order speech was multiplied by the ratio of the TAF of the spectrum-matched steady-state noise to the TAF of the normal-order speech. Thus, the normal-order speech AC had the fine structures of the normal-order speech and the TAF of the spectrum-matched steady-state noise (Drullman, Festen, & Plomp, 1994). Similarly, to obtain the time-reversed speech AC, the time-reversed speech was multiplied by the ratio of the TAF of the spectrum-matched noise to the TAF of the time-reversed speech.

Procedure

Participants pressed a button of the response box to initiate a test session. For each of the three sound types, sounds delivered from the two loudspeakers were identical in a trial. The right loudspeaker always led the left loudspeaker. Participants were instructed to indicate whether they perceived a discrete sound from the location around the left loudspeaker by pressing the left button of the response box or nothing from the location around the left loudspeaker by pressing the right button. The time lag between the two loudspeakers (called the lead/lag delay) was decreased following the participant's three successive responses, indicating that a sound around the left loudspeaker location was perceived, and increased following one response, indicating that no sound was perceived from the left loudspeaker, using a three-down-one-up procedure (Levitt, 1971). No feedback was given to participants. The testing order for the three types of sounds was counterbalanced among the 18 participants according to the Latin square design.

Each session was started with the lead/lag delay of 72 msec. The initial step size of changing the lead/lag delay was 16 msec, and the step size was altered by a factor of .5 with each reversal of direction until the minimum size of 1 msec was reached. A test session was terminated following 10 reversals in direction, and the echo threshold for that session was defined as the averaged lead/lag delay for the last six reversals. For each participant under each condition, there were four test sessions. The averaged echo threshold of the four sessions was used as the echo threshold for the participant under the condition. During testing, the participant was instructed to keep his or her head still and face the midline in the frontal area, but the head was not physically fixed.

In this and in the following psychophysical experiments (Experiments 1–3), sounds from the loudspeakers were calibrated using a B&K sound level meter (Type 2230) whose microphone was placed at the position of the head when the listener was absent, using a "slow"/"RMS" meter response. All sounds were presented at a level such that each loudspeaker, playing alone, would produce a comfortable sound pressure of 56 dBA SPL.

Results

In Experiment 1, both TAF information of normal-order speech and TAF information of time-reversed speech were removed by replacement of these speech TAFs with the TAF of spectrum-matched steady-state noise. The echo threshold of the spectrum-matched steady-state noise was significantly lower than both the echo threshold of the normal-order speech AC ($p = .011$, two-tailed paired samples t test with the significant level of $.05/3 = .0167$) and the echo threshold of the time-reversed speech AC ($p = .006$), but there was no significant difference between the two types of speech ACs ($p > .05$) (Figure 1).

EXPERIMENT 2

Methods

Participants

Eighteen university students (13 women and 5 men, 17–27 years old, mean age = 22 years) participated in this experiment.

Apparatus

The apparatus was the same as used in Experiment 1.

Stimuli

Three types of sounds were used: (a) spectrum-matched steady-state noises, (b) spectrum-matched noises modulated by the TAF of normal-order speech, and (c) spectrum-matched noises modulated by the TAF of time-reversed

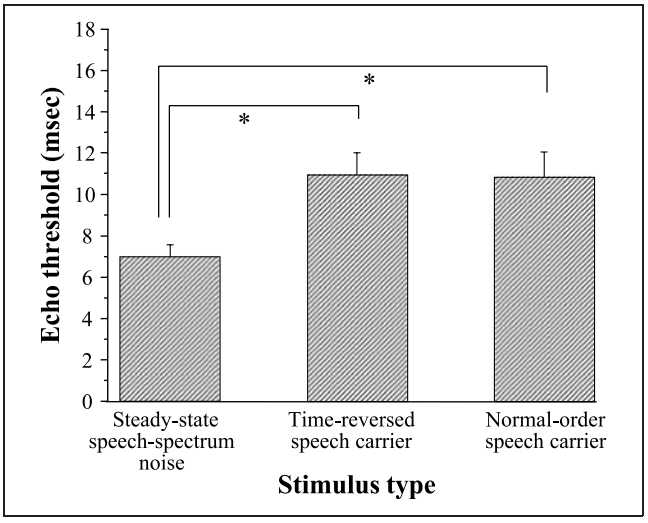

Figure 1. Comparisons of mean echo thresholds (over 18 participants) between spectrum-matched steady-state noise, time-reversed speech AC, and normal-order speech AC. These three types of stimuli had both the same long-term spectrum and the same TAF of the spectrum-matched steady-state noise. *$p < .0167$ (.05/3). The error bars represent the standard errors of the mean.

speech. To obtain the spectrum-matched noise modulated by the TAF of normal-order speech (the modulated noise had the same AC as the spectrum-matched steady-state noise and the TAF of the normal-order speech), the spectrum-matched steady-state noise was multiplied by the ratio of the TAF of normal-order speech to the TAF of the spectrum-matched steady-state noise (Drullman et al., 1994). Similarly, to obtain the spectrum-matched noise modulated by the TAF of time-reversed speech, the spectrum-matched steady-state noise was multiplied by the ratio of the TAF of time-reversed speech to the TAF of the steady-state noise.

Procedure

The procedure was the same as used in Experiment 1. The testing order for these three types of sounds was counterbalanced among the 18 participants according to the Latin square design.

Results

In Experiment 2, both the TAF of normal-order speech and the TAF of time-reversed speech were extracted, and each of the two types of TAFs was used to modulate spectrum-matched steady-state noise. The echo threshold of spectrum-matched steady-state noise significantly increased after the noise was amplitude modulated by either the TAF of normal-order speech ($p < .001$, two-tailed paired samples t test with the significant level of $.05/3 = .0167$) or the TAF of time-reversed speech ($p = .001$), and the two types of amplitude-modulated noises did not differ significantly ($p > .05$) (Figure 2).

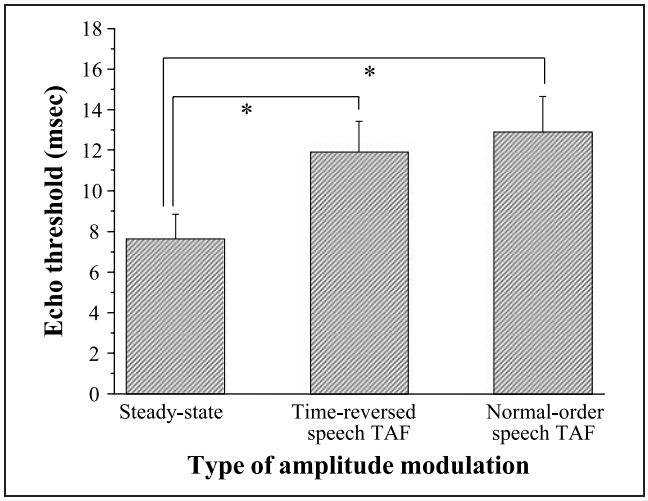

Figure 2. Comparisons of mean echo thresholds (over 18 participants) between steady-state speech-spectrum noise, the spectrum-matched noise modulated by the TAF of time-revered speech, and the spectrum-matched noise modulated by the TAF of normal-order speech. *$p < .0167$. The error bars represent the standard errors of the mean.

EXPERIMENT 3

Methods

Participants

Twenty university students (8 women and 12 men, 19–25 years old, mean age = 22 years) participated in this experiment.

Apparatus

The apparatus was the same as used in Experiment 1.

Stimuli

Four types of Chinese speech or speech-like sounds were used in this experiment, which had different TAF and AC combinations: (a) normal-order TAF and normal-order AC, (b) time-reversed TAF and time-reversed AC, (c) normal-order TAF and time-reversed AC, and (d) time-reversed TAF and normal-order AC. To obtain speech with the TAF of normal-order speech and the AC of time-reversed speech, the time-reversed speech was multiplied by the ratio of the TAF of normal-order speech to the TAF of time-reversed speech (Drullman et al., 1994). Similarly, to obtain speech with the TAF of time-reversed speech and the AC of normal-order speech, the normal-order speech was multiplied by the ratio of the TAF of time-reversed speech to the TAF of normal-order speech.

Procedure

The procedure was the same as used in Experiment 1. The testing order for the four types of sounds was counterbalanced among the 20 participants according to the Latin square design.

Results

Experiment 3 investigated whether the TAF of normal-order speech and the TAF of time-reversed speech are different in affecting the perceptual fusion tendency when the AC is of speech or speech-like. Echo-threshold comparisons were made across the following four types of Chinese speech or speech-like sounds with different TAF and AC combinations: (a) normal-order TAF and normal-order AC, (b) time-reversed TAF and time-reversed AC, (c) normal-order TAF and time-reversed AC, and (d) time-reversed TAF and normal-order AC. As shown in Figure 3, unlike the results of Experiment 2, speech sounds with the normal-order TAF had markedly larger echo thresholds than those with the time-reversed TAF.

An ANOVA confirms that the interaction between the temporal order of TAF and the temporal order of AC was not significant ($p > .05$), the main effect of temporal order of AC was not significant ($p > .05$), but the main effect of temporal order of TAF was significant, $F(1, 19) = 51.875$, $p < .001$.

EXPERIMENT 4

Methods

Participants

Eight university students (two women and six men, 19–27 years old, mean age = 24 years) participated in this experiment.

Stimuli and Apparatus

A Knowles Electronic Manikin for Acoustic Research (KEMAR) was located at the center of the anechoic chamber that is described for Experiment 1. Two types of sounds were used in this experiment: normal-order speech (which were created and reproduced as in Experiment 1) and time-reversed speech, and they were the same as used in Experiment 3. Similar to Experiment 3, speech analog signals were delivered to two loudspeakers (Dynaudio Acoustics, BM6A) in the frontal azimuthal plane at the left and right 45° positions with respect to the median plane. For a recording trial, the two loudspeakers presented identical speech (either normal order or time reversed), with the left loudspeaker lagging behind the right loudspeaker by either 2 or 40 msec.

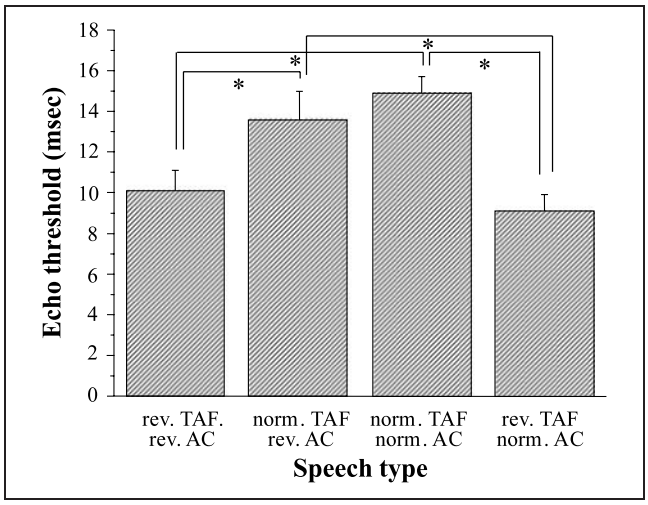

Figure 3. Comparisons of mean echo thresholds (over 20 participants) between (a) speech with time-reversed TAF and time-reversed AC, (b) speech with normal-order TAF and time-reversed AC, (c) speech with normal-order TAF and normal-order AC, and (d) speech with time-reversed TAF and normal-order AC. *$p < .0083$ (.05/6); rev. = time-reversed; norm. = normal order. The error bars represent the standard errors of the mean.

The loudspeaker height was 140 cm, which was at the ear level of KEMAR, and the distance between the loudspeaker and the center of the KEMAR was 200 cm. Sound waves were recorded using the KEMAR, which was equipped with the ear simulators (RA0045, G.R.A.S, Sound and Vibration, Holte, Denmark), a programmable front-end (BEQ II.0, Head-acoustic, Herzogenrath, Germany), and a sound-progressed software (ArtemiS 6.0, Head-acoustic). These recorded sound waves were presented to participants at the level of 77 dB SPL through two air-conduction headphones during fMRI data acquisition. A new speech sentence was used for each fMRI scanning trial.

Experimental Design

As indicated by the results of Experiment 3, the mean difference in echo threshold between normal-order speech and time-reversed speech was about 5 msec, and there were large between-participant fluctuations in echo threshold. For each of the two types of speech sounds used in Experiment 4 (the fMRI experiment), to ensure that (a) the shorter lead/lag delay (that was below the echo threshold) was the same across participants and the longer lead/lag delay (that was above the echo threshold) was also the same across participants, (b) the shorter lead/lag delay was short enough to induce perceptual fusion with a high degree of image compactness, and (c) the longer lead/lag delay was long enough to cause two discretely separated sound images, the shorter and longer lead/lag delays were set at 2 and 40 msec, respectively. Thus, there were four stimulus conditions (two sound types: normal-order speech and time-reversed speech; two lead/lag delays: 2 and 40 msec). For each condition, 60 scans were conducted, including 40 scans with speech stimuli and 20 scans with silence as the baseline. Both speech stimuli and silence presentations were delivered to each participant with a random manner. Thus, there were totally 240 scans for each participant. The order of the four conditions was counterbalanced across the eight participants.

fMRI Data Acquisition

Echo-planar brain images were acquired using a 3-T scanner (Siemens) with a standard single-channel head coil. Participants were instructed to close their eyes during scanning. A T2*-weighted gradient-echo EPI sequence was used. Each image set constituted 30 slices (slice thickness = 4 mm) to cover the whole brain. The following acquisition parameters were used: repetition time = 9 sec, echo time = 30 msec, flip angle = 90°, and voxel size = $3.75 \times 3.75 \times 5$ mm^3. Data acquisition was conducted in the last 1.8 sec of the repetition time, with a stimulus presented 6.7 sec before each scan (sparse temporal sampling, see Hall et al., 1999). Participants were instructed to press a button using their right thumb in response to the end of each scanning noise. For each participant, the individual high-resolution three-dimensional T1-weighted MR image was also collected after

functional image acquisition for coregistration and normalization of functional images.

fMRI Data Analyses

Under each stimulus condition, the first image in response to speech stimuli and both the first and the last images for the silence baseline were omitted from analyses. The remaining 228 scans for each participant were preprocessed using SPM2 software (http://www.fil.ion.ucl.ac.uk). The preprocessing of the functional images included realignment (rigid-body transformation), slice timing for adjusting head movements and slice acquisition delays, and coregistration with the anatomical data. The images were then normalized to the Montreal Neurological Institute space using nonlinear transformations and smoothed with a Gaussian kernel of 6-mm full width at half maximum. Statistical analyses at the first level implemented the general linear model that specified four variables encompassing the four stimulus conditions. Linear contrasts between stimulus conditions and baseline conditions (normal-order/2-msec delay vs. baseline, normal-order/40-msec delay vs. baseline, time-reversed/2-msec delay vs. baseline, and time-reversed/40-msec delay vs. baseline) were calculated to generate a contrast-parameter-estimate map for each participant. The contrast-parameter-estimate maps were then high-pass filtered at 256 sec to deal with the low-frequency artifacts. Then, at the second level, the high-pass filtered contrast images of the first level were submitted to a random-effect group analysis using one-way ANOVA. Activations were reported only if it exceeded a voxel-level significance threshold of $p < .05$ (uncorrected for multiple comparison) and extent threshold = 185 (Leff et al., 2008). To use an automated labeling system (Lancaster et al., 2000), the spatial coordinates of activation maxima were transformed from the Montreal Neurological Institute space into Talairach space (Brett, Johnsrude, & Owen, 2002).

Results

The results of the psychophysical experiments confirm that perceptual fusion of correlated leading and lagging sounds depends on the lead/lag delay. In Experiment 4, we used the fMRI method to investigate whether the cortical activations associated with perceptual fusion (when the lead/lag delay is sufficiently shorter than the echo threshold) are different from those associated with perceptual separation (when the lead/lag delay is sufficiently longer than the echo threshold), although the sounds presented from the leading and lagging sources are always identical. Moreover, because the results of Experiment 3 suggest that perceptual fusion of speech-like sounds also depends on whether the TAF is in normal or time-reversed order, Experiment 4 particularly investigated whether the cortical activations associated with perceptual fusion of normal-order speech are different from those associated with perceptual fusion of time-reversed speech.

Either normal-order speech or time-reversed speech was presented to the listeners at the two ears with headphones. Because the speech stimuli contained signals that were modified by the head-related transfer functions, when the stimulus at the right ear led that at the left ear by 2 msec (this lead/lag delay was far below the echo threshold), listeners experienced only one sound image as coming from the right semifield. Also, when the stimulus at the right ear led that at the left ear by 40 msec (this lead/lag delay was far above the echo threshold), listeners experienced two spatially separated sound images, one as coming from the right semifield and the other as coming from the left semifield.

Compared with normal-order speech without perceptual fusion (when the lead/lag delay was 40 msec), perceptually fused normal-order speech (when the lead/lag delay was 2 msec) was associated with increased BOLD contrast activations in both the right ACC (Brodmann area [BA] 32) (top panels in Figure 4) and the left middle temporal gyrus (BA 21) (bottom panels in Figure 4). However, compared with time-reversed speech without perceptual fusion (when the lead/lag delay was 40 msec), perceptually fused time-reversed speech (when the lead/lag delay was 2 msec) was associated with increased BOLD contrast activations only in the left superior frontal gyrus (BA 10) (Figure 5).

Moreover, although both normal-order speech and time-reversed speech were perceived as a single fused image at the lead/lag delay of 2 msec, perceptually fused normal-order speech, compared with perceptually fused time-reversed speech, was associated with more BOLD activations in the right ACC (BA 32) (top panels in Figure 6), the right inferior parietal lobule (IPL, BA 40) (middle panels in Figure 6), and the left anterior part of the superior temporal gyrus (STG, BA 38) (bottom panels in Figure 6). However, these patterns of cortical activation did not occur for perceptual separation of normal-order speech against perceptual separation of time-reversed speech (Figure 7).

CALCULATING THE SPEED OF AMPLITUDE INCREASE AND THE SPEED OF AMPLITUDE DECREASE

In all, 432 sentences were used for the statistics. For each sentence, the TAF of the original wideband speech signal (16 bits at the sampling rate of 22,050 Hz) was obtained using the Hilbert transform. To ensure that the amplitude slope was sufficiently steep, the TAF was low-pass filtered at the cutoff frequency of 500 Hz, with the filter slope of approximately −40 dB/oct, and then down sampled to the sampling rate of 1000 Hz. Finally, the down-sampled TAF was low-pass filtered at the cutoff frequency of 50 Hz for calculating amplitude rises and falls.

The maximum amplitude of each envelope with the sampling rate of 1000 Hz was normalized to 1, and differences in amplitude between adjacent sampling points were obtained as either envelope rises (positive values) or envelope falls (negative values). To eliminate both silent and flat components, the amplitude differences were removed

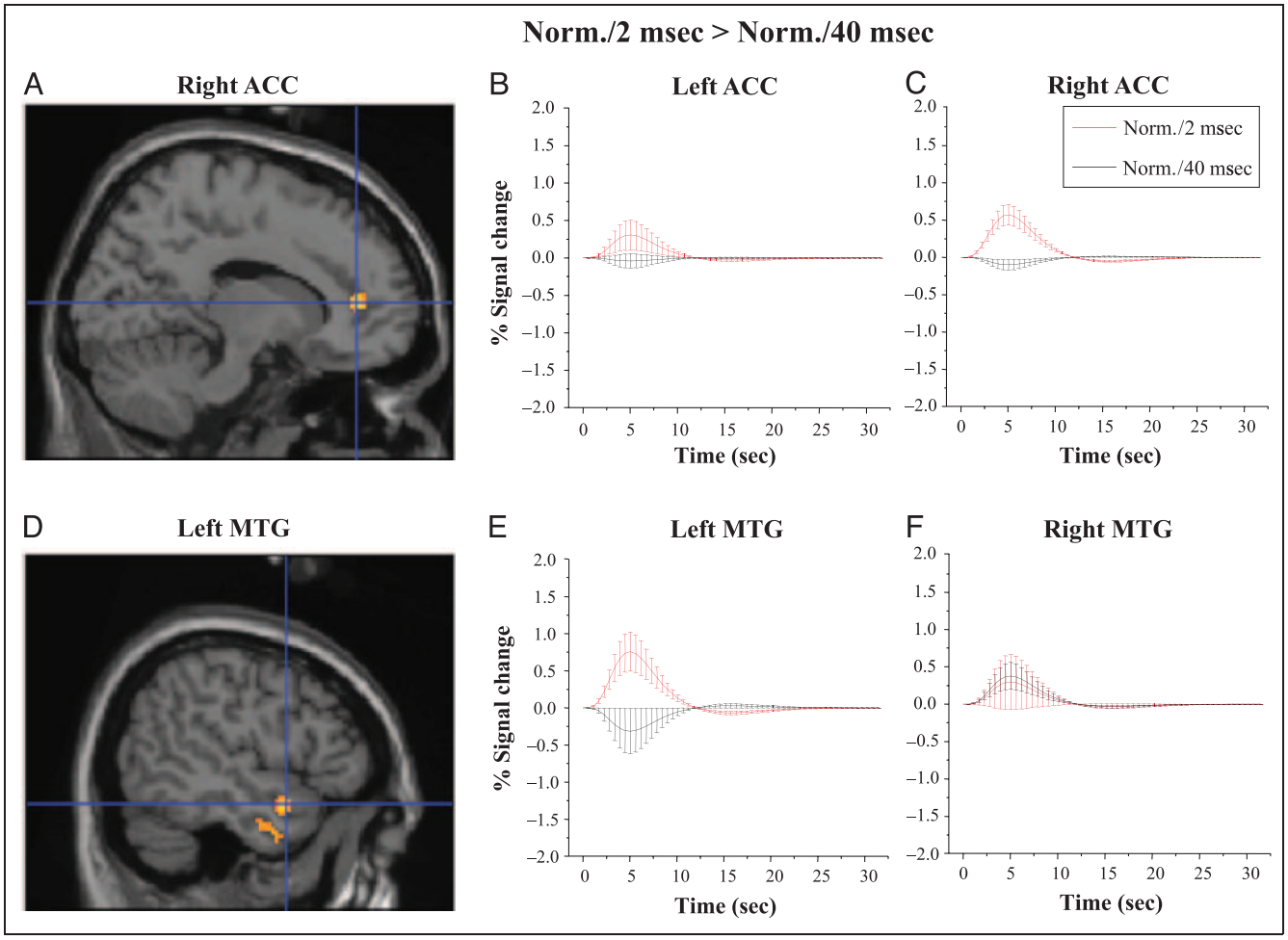

Figure 4. Statistically significant group-mean BOLD activations for perceptual fusion of normal-order speech (Norm./2 msec) against perceptual separation of normal-order speech (Norm./40 msec) in the right ACC (ACC, BA 32) (top panels) and in the left middle temporal gyrus (MTG, BA 21) (bottom panels). (A) Increased activations of the right ACC. (B) The time courses of BOLD signal changes in the left ACC (contralateral to the right ACC depicted in Panel A). (C) The time courses of BOLD signal changes in the right ACC depicted in Panel A. (D) Increased activations of the left MTG. (E) The time courses of BOLD signal changes in the left MTG depicted in Panel D. (F) The time courses of BOLD signal changes in the right MTG (contralateral to the left MTG depicted in Panel D). Error bars indicate the standard errors of the mean across participants. In this and in the following figures, the image threshold is set at $p < .05$ (voxel level, uncorrected) and the extent threshold is 185 (Leff et al., 2008).

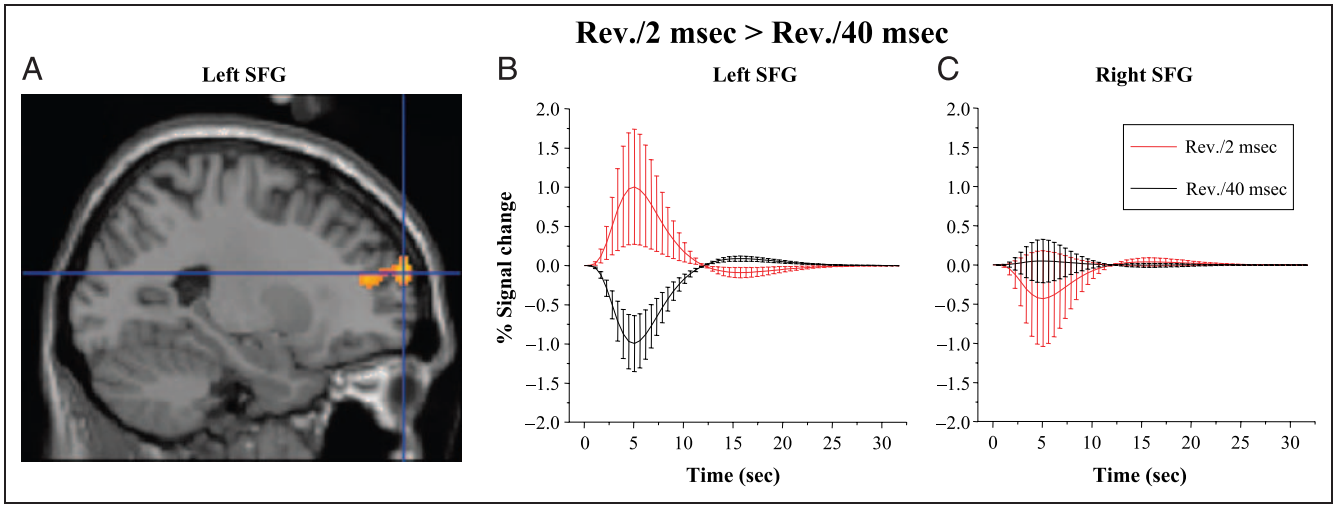

Figure 5. Statistically significant group-mean BOLD activations for perceptual fusion of time-reversed speech (Rev./2 msec) against perceptual separation of time-reversed speech (Rev./40 msec) in the left superior frontal gyrus (SFG, BA 10). (A) Increased activations of the left SFG. (B) The time courses of BOLD signal changes in the left SFG depicted in Panel A. (C) The time courses of BOLD signal changes in the right SFG (contralateral to the left SFG depicted in Panel A). Error bars indicate the standard errors of the mean across participants.

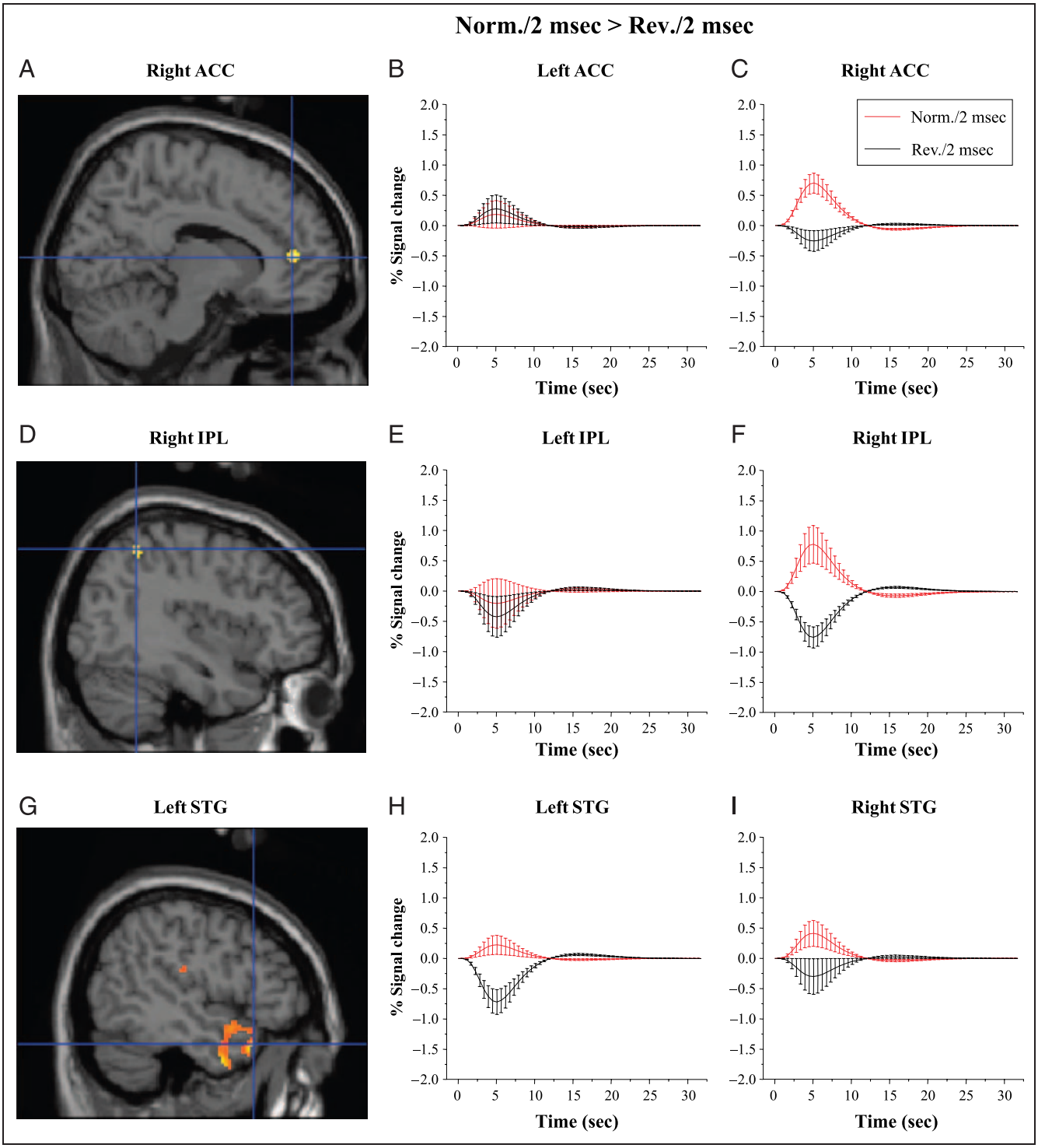

Figure 6. Statistically significant group-mean BOLD activations for perceptual fusion of normal-order speech (Norm./2 msec) against perceptual fusion of time-reversed speech (Rev./2 msec) in the right ACC (BA 32) (top panels), the right inferior parietal lobule (IPL, BA 40) (middle panels), and the left superior temporal gyrus (STG, BA 38) (bottom panels). (A) Increased activations of the right ACC. (B) The time courses of BOLD signal changes in the left ACC (contralateral to the right ACC depicted in Panel A). (C) The time courses of BOLD signal changes in the right ACC depicted in Panel A. (D) Increased activations of the right IPL. (E) The time courses of BOLD signal changes in the left IPL (contralateral to the right IPL depicted in Panel D). (F) The time courses of BOLD signal changes in the right IPL depicted in Panel D. (G) Increased activations of the left STG. (H) The time courses of BOLD signal changes in the left STG depicted in Panel G. (I) The time courses of BOLD signal changes in the right STG (contralateral to the left STG depicted in Panel G). Error bars indicate the standard errors of the mean across participants.

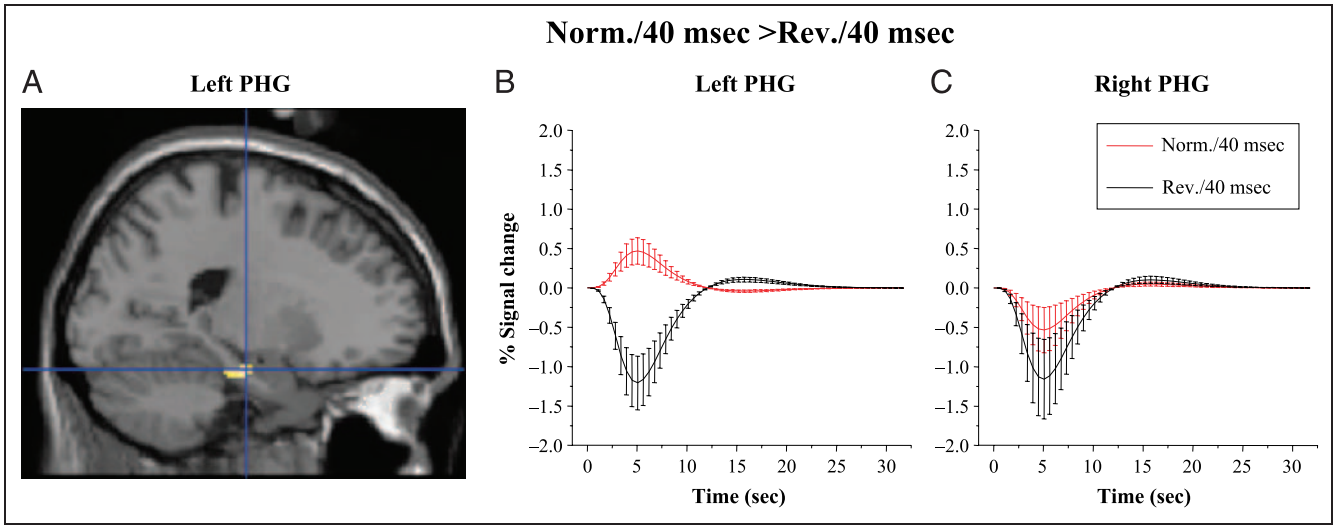

Figure 7. Statistically significant group-mean BOLD activations for perceptual separation of normal-order speech (Norm./40 msec) against perceptual separation of time-reversed speech (Rev./40 msec) in the left parahippocampal gyrus (PHG, BA 28). (A) Increased activations of the left PHG. (B) The time courses of BOLD signal changes in the left PHG depicted in Panel A. (C) The time courses of BOLD signal changes in the right PHG (contralateral to the left PHG depicted in Panel A). Error bars indicate the standard errors of the mean across participants.

from statistics if their absolute values were less than 0.0014. The results of the statistics show that the ratio of the mean positive rates to the mean negative rates was 1.252, which indicates that the amplitude rise was generally 25.2% faster than the fall.

DISCUSSION

In the present study, when the overall TAF of Chinese speech was extracted using the Hilbert transform (Zeng et al., 2005; Smith et al., 2002), the TAF and the AC could be artificially separated and manipulated independently. Although temporally reversing either the TAF or the AC does not change the long-term spectrum, the fundamental frequency, and other spectrotemporal features (including the overall spectral modulation density, temporal modulation frequency, and modulation amplitude), the speed of amplitude increase (the onset rate) for the TAF of normal-order Chinese speech is generally 25.2% faster than that for the TAF of time-reversed Chinese speech.

The results of Experiment 1 show that the echo threshold of the spectrum-matched steady-state noise, which did not contain any spectrotemporal modulations, was significantly lower than both the echo threshold of the normal-order speech AC and the echo threshold of the time-reversed speech AC, indicating that certain structural characters of speech AC, including harmonic structures, frequency modulation, and/or periodic occurrences of noise-like consonants, are able to enhance the tendency of the perceptual fusion. Because there was no significant difference between the two types of speech ACs, the temporal order of speech AC does not affect the enhancing effect.

The results of Experiment 2 show that when the AC was noise, amplitude modulation of the noise AC with either the TAF of normal-order speech or the TAF of time-reversed speech equally enhanced the perceptual fusion tendency, indicating a general amplitude-modulation effect (Schubert & Wernick, 1969). However, on the basis of the Rakerd and Hartmann (1986) study using tones with various onset durations and other parameters, when the tone-onset duration is not longer than 100 msec, shortening the tone-onset duration (i.e., increasing the onset rate) can monotonically facilitate the precedence effect. In future studies, it is worth investigating whether artificially manipulating the general contrast between the onset rate and the offset rate of the TAF in a larger range can significantly affect the echo threshold of the noise that is amplitude modulated by the TAF.

The results of Experiment 3 show that speech sounds with the normal-order TAF had markedly larger echo thresholds than those with the time-reversed TAF. Thus, when the AC is of speech or speech-like, regardless of its temporal order, the TAF of normal-order speech plays a special role in maintaining the high tendency of speech fusion.

In summary, the results of these three psychophysical experiments indicate that the markedly high perceptual fusion tendency of speech sounds is caused by the amalgamation of at least three effects: (a) the temporally bidirectional effect of speech (or speech-like) AC (Experiment 1), (b) the temporally bidirectional effect of amplitude modulation (Experiment 2), and (c) the specific effect of amplitude modulation induced by the TAF of normal-order speech only when the AC is of speech (or speech-like) (Experiment 3).

The results of Experiment 3 indicate that relative to time-reversed speech, normal-order speech has a larger perceptual fusion tendency. The main objective of Experiment 4 was to examine whether the BOLD activations associated with perceptual fusion of normal-order speech are different

from those associated with perceptual fusion of time-reversed speech. The results of the fMRI experiment (Experiment 4) of this study for the first time show that compared with normal-order speech without perceptual fusion (when the lead/lag delay was 40 msec), perceptually fused normal-order speech (when the lead/lag delay was 2 msec) was associated with increased BOLD contrast activations in both the right ACC, which is involved in top–down attentional modulation of auditory processing (Crottaz-Herbette & Menon, 2006), and the left middle temporal gyrus, which is involved in both phonemic perception (Liebenthal, Binder, Spitzer, Possing, & Medler, 2005) and semantic processing (Friederici, 2002). However, compared with time-reversed speech without perceptual fusion, perceptually fused time-reversed speech was associated with increased BOLD contrast activations only in the left superior frontal gyrus.

Compared with perceptually fused time-reversed speech, perceptually fused normal-order speech was associated with more BOLD activations in the attention-control-related right ACC (BA 32), the right inferior parietal lobule (BA 40), which is a cortical area in the "where" pathway processing auditory spatial information (Wang, Wu, & Li, 2008; Arnott, Binns, Grady, & Alain, 2004; Alain, Arnot, Hevenor, Graham, & Grady, 2001; Weeks et al., 1999), and the left anterior part of the superior temporal gyrus (BA 38), which is involved in processing speech signals for comprehension (Hickok & Poeppel, 2007; Binder et al., 1997, 2000; Scott, Blank, Rosen, & Wise, 2000). These patterns of cortical activations did not occur for perceptual separation of normal-order speech against perceptual separation of time-reversed speech. These results suggest that when two correlated normal-order speech sounds become perceptually fused, both the cortical processing of speech-spatial information and that of speech-content information are enhanced. Moreover, these cortical processing enhancements are accompanied with a facilitation of cortical top–down attentional modulations of auditory processing. However, it should be noted that because of both the limitation in trial numbers and the long time of a trial cycle (9000 msec) in the fMRI experiment, the threshold-tracking procedure used in the psychophysical experiments was not applicable in the fMRI experiment. Thus, the temporally analytical level of the fMRI experiment did not match that of the psychophysical experiments.

Perceptual fusion of correlated long-lasting sounds is an interesting auditory illusion that is based on perceptual capture of attributes of the lagging sounds, and the capture tendency is attribute dependent (Li et al., 2005). However, very few reports have been found in the literature addressing the neural mechanisms underlying this auditory illusion in humans, particularly for speech or speech-like sounds under conditions simulating reverberant environments. In this line of brain-imaging studies, sounds presented from the two simulated spatially separated sources should be always identical or highly correlated across stimulus conditions with various lead/lag delays, simulating the direct wave and the reflection. Also, some acoustic features of the stimuli, including the long-term spectrum, fundamental frequency, spectral modulation density, temporal modulation frequency, and modulation amplitude, should be maintained constant across conditions.

As mentioned in the Introduction, speech sounds have much larger echo thresholds than other types of sounds such as clicks and noise bursts (Rakerd et al., 2000; Litovsky et al., 1999; Lochner & Burger, 1958; Cherry & Taylor, 1954; Wallach et al., 1949). The distinctly large perceptual fusion tendency of speech sounds suggests that speech sounds are very special acoustic stimuli for human listeners because of the importance for communication. In a noisy, reverberant environment, because perceptual fusion of correlated sounds normally co-occurs with perceived spatial separation between uncorrelated sound sources, which in turn facilitates listeners' selective attention to target speech and consequently releases target speech from masking (Huang, Huang, et al., 2008, 2009; Rakerd et al., 2006; Wu et al., 2005; Li et al., 2004; Freyman et al., 1999), the enhanced perceptual fusion tendency of speech sounds must play an important role in improving speech recognition under the adverse acoustic condition. The results of the present study emphasize the notion that perceptual fusion and recognition of speech sounds under adverse conditions are interrelated, and they are all associated with higher-order top–down modulations of speech-sound processing. It should be noted that according to the Wallach et al. (1949) study, the echo threshold of piano music is comparable to that of speech. Because musical pieces also contain harmonic structures and frequency/amplitude temporal modulations and are important for humans' communication (particularly those with emotional, i.e., romantic, components), in future studies it is well worth investigating the acoustic features that can lead to the high perceptual fusion tendency of music.

It is of interest to know whether the difference in perceptual fusion tendency between normal-order speech and time-reversed speech is associated with the difference in informational-masking effectiveness between normal-order speech masker and time-reversed speech masker. In the Freyman et al. (2001) study, the two-talker normal-order speech masker produced larger informational masking of target speech than the two-talker time-reversed speech masker because a shift of the stimulus condition from perceived target/masker separation (which was induced by perception fusion of the masker stimuli that were delivered from two spatially separated loudspeakers) to real target/masker colocation at a single loudspeaker caused a larger reduction in recognizing target speech when the masker was normal-order speech than when the masker was time-reversed speech. On the other hand, in the Rakerd et al. (2006) study, a two-talker (normal-order) speech masker was presented by two spatially separated loudspeakers and the interloudspeaker time interval for the masker was varied in a broad range from −64 to +64 msec. At the same time, the target speech was presented only by

one of the loudspeakers. When the absolute value of intermasker interval was 32 msec or shorter, there was consistent evidence of release from speech masking for target speech recognition. If the masker became speech-spectrum noise, significant release of target speech occurred only at a few short intermasker intervals less than 4 msec. Thus, the release of target speech from speech masking over a range of intermasker intervals between 4 and 32 msec 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. Although in the Rakerd et al. study the time-reversed speech was not used, on the basis of both the results of the present study and the results of the Freyman et al. (2001) study, it can be expected that if the masker becomes time-reversed speech under the experimental conditions used in the Rakerd et al. study, the maximum intermasker interval for significantly releasing target speech must be shorter than 32 msec because time-reversed speech has both smaller perceptual fusion tendency and smaller informational-masking effectiveness than normal-order speech.

This fusion-enhancing mechanism underlying the improvement of speech recognition in noisy, reverberant environments may reflect a more general cross-time perceptual "grouping" strategy of the brain for unmasking speech signals. As summarized by Moore (2003), temporal integration of acoustic signals over time, which improves detection and discrimination under masking conditions, does not involve a simple summation or accumulation process but is based on the formation of the internal representation of acoustic inputs. Because temporal integration is also based on temporal storage of acoustic details (Huang, Huang, et al., 2009; Huang, Wu, & Li, 2009; Li, Huang, Wu, Qi, & Schneider, 2009; Huang, Kong, Fan, Wu, & Li, 2008), the fusion-tendency enhancement specifically induced by the normal-order speech TAF may be due to an expansion of the temporal storage of the internal representation of speech signals when top–down modulations of the speech-processing cortical areas are elicited.

Acknowledgment

This study was supported by the National Natural Science Foundation of China (30670704; 30711120563; 60535030), the "973" National Basic Research Program of China (2009CB320901), and "985" grants from Peking University.

Reprint requests should be sent to Liang Li, Department of Psychology, Peking University, Beijing, 100871, China, or via e-mail: liangli@pku.edu.cn.

REFERENCES

Alain, C., Arnot, S. R., Hevenor, S., Graham, S., & Grady, C. (2001). "What" and "where" in the human auditory system. *Proceedings of the National Academy of Sciences, U.S.A., 98,* 12301–12306.

Arbogast, T. L., Mason, C. R., & Kidd, G. J. R. (2002). The effect of spatial separation on informational and energetic masking of speech. *Journal of the Acoustical Society of America, 112,* 2086–2098.

Arbogast, T. L., Mason, C. R., & Kidd, G. J. R. (2005). The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America, 117,* 2169–2180.

Arnott, S. R., Binns, M. A., Grady, C. L., & Alain, C. (2004). Assessing the auditory dual-pathway model in humans. *Neuroimage, 22,* 401–408.

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S. F., Springer, J. A., Kaufman, J. N., et al. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex, 10,* 512–528.

Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M., & Prieto, T. (1997). Human brain language areas identified by functional magnetic resonance imaging. *Journal of Neuroscience, 17,* 353–362.

Brett, M., Johnsrude, I. S., & Owen, A. M. (2002). The problem of functional localization in the human brain. *Nature Reviews Neuroscience, 3,* 243–250.

Brungart, D. S., Simpson, B. D., & Freyman, R. L. (2005). Precedence-based speech segregation in a virtual auditory environment. *Journal of the Acoustical Society of America, 118,* 3241–3251.

Cherry, E. C., & Taylor, W. K. (1954). Some further experiments upon the recognition of speech with one and with two ears. *Journal of the Acoustical Society of America, 26,* 554–559.

Crottaz-Herbette, S., & Menon, V. (2006). Where and when the anterior cingulate cortex modulates attentional response: Combined fMRI and ERP evidence. *Journal of Cognitive Neuroscience, 18,* 766–780.

Drullman, R., Festen, J. M., & Plomp, R. (1994). Effect of temporal envelope smearing on speech reception. *Journal of the Acoustical Society of America, 95,* 1053–1064.

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2001). Spatial release from informational masking in speech recognition. *Journal of the Acoustical Society of America, 109,* 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. *Journal of the Acoustical Society of America, 115,* 2246–2256.

Freyman, R. L., Clifton, R. K., & Litovsky, R. Y. (1991). Dynamic processes in the precedence effect. *Journal of the Acoustical Society of America, 90,* 874–884.

Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *Journal of the Acoustical Society of America, 106,* 3578–3588.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences, 6,* 78–84.

Hall, D. A., Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliott, M. R., et al. (1999). "Sparse" temporal sampling in auditory fMRI. *Human Brain Mapping, 7,* 213–223.

Helfer, K. S. (1997). Auditory and auditory-visual perception of clear and conversational speech. *Journal of Speech, Language, and Hearing Research, 40,* 432–443.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8,* 393–402.

Huang, Y., Huang, Q., Chen, X., Qu, T. S., Wu, X. H., & Li, L. (2008). Perceptual integration between target speech and target-speech reflection reduces masking for target-speech recognition in younger adults and older adults. *Hearing Research, 244,* 51–65.

Huang, Y., Huang, Q., Chen, X., Wu, X. H., & Li, L. (2009). Transient auditory storage of acoustic details is associated with release of speech from informational masking in reverberant conditions. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 1618–1628.

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. *NeuroReport, 19,* 1673–1678.

Huang, Y., Wu, X. H., & Li, L. (2009). Detection of the break in interaural correlation is affected by interaural delay, aging, and center frequency. *Journal of the Acoustical Society of America, 126,* 300–309.

Lancaster, J. L., Woldorff, M. G., Parsons, L. M., Liotti, M., Freitas, C. S., Rainey, L., et al. (2000). Automated Talairach atlas labels for functional brain mapping. *Human Brain Mapping, 10,* 120–131.

Leff, A. P., Schofield, T. M., Stephan, K. E., Crinion, J. T., Friston, K. J., & Price, C. J. (2008). The cortical dynamics of intelligible speech. *Journal of Neuroscience, 28,* 13209–13215.

Levitt, H. (1971). Transformed up–down methods in psychoacoustics. *Journal of the Acoustical Society of America, 49,* 467–477.

Li, L., Daneman, M., Qi, J. G., & Schneider, B. A. (2004). Does the information content of an irrelevant source differentially affect speech recognition in younger and older adults? *Journal of Experimental Psychology: Human Perception and Performance, 30,* 1077–1091.

Li, L., Huang, J., Wu, X. H., Qi, J. G., & Schneider, B. (2009). The effects of aging and interaural delay on the detection of a break in the interaural correlation between two sounds. *Ear and Hearing, 30,* 273–286.

Li, L., Qi, J. G., He, Y., Alain, C., & Schneider, B. (2005). Attribute capture in the precedence effect for long-duration noise sounds. *Hearing Research, 202,* 235–247.

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., & Medler, D. A. (2005). Neural substrates of phonemic perception. *Cerebral Cortex, 15,* 1621–1631.

Litovsky, R. Y., Colburn, H. S., Yost, W. A., & Guzman, S. J. (1999). The precedence effect. *Journal of the Acoustical Society of America, 106,* 1633–1654.

Lochner, J. P. A., & Burger, J. F. (1958). The subjective masking of short time delayed echoes, their primary sounds, and their contribution to the intelligibility of speech. *Acustica, 8,* 1–10.

Moore, B. C. J. (2003). Temporal integration and context effects in hearing. *Journal of Phonetics, 31,* 563–574.

Oppenheim, A. V., Schafer, R. W., & Buck, J. R. (1999). *Discrete-time signal processing*. Upper Saddle River, NJ: Prentice-Hall Press.

Rakerd, B., Aaronson, N. L., & Hartmann, W. M. (2006). Release from speech-on-speech masking by adding a delayed masker at a different location. *Journal of the Acoustical Society of America, 119,* 1597–1605.

Rakerd, B., & Hartmann, W. M. (1986). Localization of sound in rooms: III. Onset and duration effects. *Journal of the Acoustical Society of America, 80,* 1695–1706.

Rakerd, B., Hartmann, W. M., & Hsu, J. (2000). Echo suppression in the horizontal and median sagittal planes. *Journal of the Acoustical Society of America, 107,* 1061–1064.

Rosen, S. (1992). Temporal information in speech—Acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences, 336,* 367–373.

Schubert, E. D., & Wernick, J. (1969). Envelope versus microstructure in the fusion of dichotic signals. *Journal of the Acoustical Society of America, 45,* 1525–1531.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. S. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain, 123,* 2400–2406.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature, 416,* 87–90.

Wallach, H., Newman, E. B., & Rosenzweig, M. R. (1949). The precedence effect in sound localization. *American Journal of Psychology, 62,* 315–336.

Wang, W. J., Wu, X. H., & Li, L. (2008). The dual-pathway model of auditory signal processing. *Neuroscience Bulletin, 24,* 173–182.

Weeks, R. A., Aziz-Sultan, A., Bushara, K. O., Tian, B., Wessinger, C. M., Dang, N., et al. (1999). A PET study of human auditory spatial processing. *Neuroscience Letters, 262,* 155–158.

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. *Hearing Research, 199,* 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. *Speech Communication, 49,* 892–904.

Zeng, F. G., Nie, K. B., Stickney, G. S., Kong, Y. Y., Vongphoe, M., Bhargave, A., et al. (2005). Speech recognition with amplitude and frequency modulations. *Proceedings of the National Academy of Sciences, U.S.A., 102,* 2293–2298.

Zurek, P. M. (1980). The precedence effect and its possible role in the avoidance of interaural ambiguities. *Journal of the Acoustical Society of America, 67,* 953–964.