Neuroscience 161 (2009) 59–66

IMMEDIATE INTEGRATION OF PROSODIC INFORMATION FROM SPEECH AND VISUAL INFORMATION FROM PICTURES IN THE ABSENCE OF FOCUSED ATTENTION: A MISMATCH NEGATIVITY STUDY

X. LI, Y. YANG* AND G. REN

State Key Laboratory of Brain and Cognitive Science, Institute of Psychology, Chinese Academy of Sciences, 4A Datun Road, Chaoyang District, Beijing 100101, PR China

Abstract—Language is often perceived together with visual information. Recent experimental evidences indicated that, during spoken language comprehension, the brain can immediately integrate visual information with semantic or syntactic information from speech. Here we used the mismatch negativity to further investigate whether prosodic information from speech could be immediately integrated into a visual scene context or not, and especially the time course and automaticity of this integration process. Sixteen Chinese native speakers participated in the study. The materials included Chinese spoken sentences and picture pairs. In the audiovisual situation, relative to the concomitant pictures, the spoken sentence was appropriately accented in the standard stimuli, but inappropriately accented in the two kinds of deviant stimuli. In the purely auditory situation, the speech sentences were presented without pictures. It was found that the deviants evoked mismatch responses in both audiovisual and purely auditory situations; the mismatch negativity in the purely auditory situation peaked at the same time as, but was weaker than that evoked by the same deviant speech sounds in the audiovisual situation. This pattern of results suggested immediate integration of prosodic information from speech and visual information from pictures in the absence of focused attention.

Key words: event-related potential, mismatch negativity, accentuation, audiovisual integration, speech processing.

Everyday tasks involve the integration of information from multiple sensory modalities. One such integration is that of auditory inputs and the concomitant information from visual modality. Perceptual judgments can reflect combined information from auditory and visual senses (Mishra et al., 2007; McDonald et al., 2003; Shams et al., 2000; McGurk and Macdonald, 1976). Speech comprehension is also influenced by both linguistic information from speech and visual information from the situational context. This raises the question on how and when auditory and visual information combines to form a coherent interpretation of the external world.

The integration of speech with visual information has been explored in a variety of ways in eye-tracking, event-related potential (ERP), or functional magnetic resonance imaging (fMRI) studies. After reviewing these results, we identified two major strands of previous research. The first line of research found evidence for the on-line influence of visual scene context on speech processing (e.g. Knoeferle et al., 2005; Tanenhaus et al., 1995). For example, the influential eye-tracking studies by Tanenhaus et al. (1995) revealed that the referential visual contrast between two objects affected on-line resolution of local structure ambiguity in a spoken sentence (e.g. Put the apple on the towel in the box). The second line of research looked into the integration of picture information into spoken sentence context. Some ERP studies manipulating the semantic fit of a picture with respect to the preceding part of a spoken sentence reported similar N400 amplitudes and onset latencies as found for integration of semantic information conveyed through a word (e.g. Willems et al., 2008; Özyürek et al., 2007; Federmeier and Kutas, 2001; Ganis et al., 1996). Recently, Willems et al. (2008) further found that, despite obvious differences in representational format, picture and word integration in a spoken sentence evoked similar N400 effect and led to overlapping activations in the left inferior frontal cortex.

All of those mentioned studies revealed that linguistic information from speech and visual information from pictures can be immediately integrated together to form meaning interpretation. However, those studies only demonstrated how visual information integrated with semantic or syntactic information from speech. Besides semantic or syntactic information, language also involves an additional dimension of information, namely prosody. Prosody is one kind of suprasegmental phonological information in the speech signal, which is usually distinguished as prosodic structure, intonation, and accentuation. Different from semantic or syntactic information, prosody cannot be derived from the underlying sequence of single phonetic segments, since it usually involves two or more segments and occurs simultaneously with those segments. In the past decades, there have been some studies conducted to investigate the role of prosody in spoken language comprehension.

Accentuation is one aspect of prosody, which refers to the relative prominence of a particular syllable, word, or

*Corresponding author. Tel: +86-10-64888629; fax: +86-10-64872070.
E-mail address: yangyf@psych.ac.cn (Y. Yang).
Abbreviations: CW, critical word; EEG, electroencephalogram; EOG, electro-oculogram; ERP, event-related potential; fMRI, functional magnetic resonance imaging; Given-Accented, given information that was accented; MMN, mismatch negativity; New-Accented, new information that was accented; New-Deaccented, new information that was deaccented.

0306-4522/09 $ - see front matter © 2009 IBRO. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.neuroscience.2009.01.070

phrase in a certain prosodic structure realized mainly by modulation of pitch. It has been found that there is some correspondence between accentuation and information structure in a discourse context; speakers tend to place a pitch accent on new information, while leaving given information deaccented (e.g. Wang et al., 2002; Bock and Mazzella, 1983). Considerable studies have also investigated the role of accentuation in spoken language comprehension. Psycholinguistic studies using behavioral measures found that speech processing was facilitated when new information is accented and given information deaccented (e.g. Li and Yang, 2004, 2005; Dahan et al., 2002; Terken and Nooteboon, 1987; Bock and Mazzella, 1983). Several electroencephalogram (EEG) studies revealed that when a discourse constituent was inconsistently accented, a negative or positive ERP effect was evoked immediately (Li et al., 2008a; Magne et al., 2005; Johnson et al., 2003; Hruska et al., 2000). Recently, using ERP technique, Li et al. (2008b) further demonstrated that accentuation conveys sentence-level meaning, and is processed linguistically in a similar way to lexical semantics in a discourse context. Therefore, all of those behavioral and ERP studies provided firm evidence for the fact that accentuation plays an important role in spoken discourse comprehension.

In summary, on the one hand, previous studies suggested that semantic or syntactic information (segmental information) from speech can be immediately integrated with visual information during meaning interpretation. On the other hand, besides semantic and syntactic information, spoken language involves prosody (suprasegmental information), such as accentuation. It was found that accentuation can indicate the information state of the corresponding discourse constituent, and is processed linguistically during language comprehension. That raised the question of how the prosodic information, such as accentuation, was processed in a visual scene context. That is, can the prosodic information be immediately integrated into a visual scene context during language processing? At what point along the language processing pathway does this audiovisual interaction take place? In addition, previous studies concerning accentuation processing or related audiovisual integration mainly explored the process of meaning understanding in the focused attention. Not much is known about the speech–picture integration under non-attentional condition. Given those mentioned questions, the present study aimed to investigate the integration of prosodic information from speech and visual information from pictures in the absence of focused attention, and especially the timing characteristics of this integration process.

These research questions were investigated by using scalp recorded ERPs. The ERP component of interest is the mismatch negativity (MMN), which is an event-related response elicited by infrequent acoustic events (deviant) occasionally occurring among frequently repeated sounds (standard). The averaged ERP elicited by the deviant stimuli deflects negatively relative to that elicited by the standard stimuli. The MMN, usually peaking around 100–250 ms after a stimulus onset, is considered to be the outcome of an automatic comparison process between the deviant stimulus and the memory trace formed by the repeated representation of the standard stimuli (e.g. Näätänen, 2001; Näätänen et al., 1993). Some neurophysiological studies have shown that the MMN can arise during audiovisual speech perception (Froyen et al., 2008; Colin et al., 2002; Möttönen et al., 2002; Sams et al., 1991). That indicated that the MMN response characteristics are sensitive not only to the acoustic changes but also to the change in the audiovisual matching. Given its automaticity and sensitivity to audiovisual matching, the MMN was used to investigate the implicit integration of prosodic and visual information in the current study.

In this study, the cortical evoked potentials were recorded using an oddball paradigm in two situations: purely auditory and audiovisual situations. In the audiovisual situation, relative to the concomitant pictures, the spoken sentence included appropriate accentuation in the standard stimuli, but inappropriate accentuation in the two kinds of deviant stimuli. In the purely auditory situation, the same speech sentences from audiovisual situation were presented without pictures. If accentuation coded information can be automatically integrated with visual information from pictures, the deviants in the audiovisual situation would evoke the MMN effect, and the audiovisual MMN would be stronger than that evoked by the pure speech changes in the auditory situation. In addition, by comparing the time characteristics of the MMNs evoked by the same deviant speech sounds with and without pictures, we could examine the relative time course of this audiovisual integration.

EXPERIMENTAL PROCEDURES

Subjects

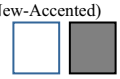
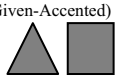
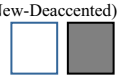
Sixteen right-handed subjects (22–30 years old; five females) participated in the experiment; all were native speakers of mandarin Chinese. None of them had any neurological impairment or history of neurological or psychiatric problems.

Stimuli

The auditory stimulus was a Chinese spoken sentence “then a red rectangle appears” (随后红色的方形出现了), which was produced by a male speaker and recorded at a sampling rate of 4410 Hz. The critical word (CW) red (红色) was accented (maximum pitch=261 Hz, duration=580 ms) in one version of the spoken sentence and deaccented (maximum pitch=186 Hz, duration=390 ms) in another version. Afterwards, the digitized sentences were edited using Praat software (Boersma and Weenink, 2004, from http://www.praat.org/) to equate the mean intensities of the CW in the two versions of spoken sentence and to replace the first word then (随后) in the two versions with the same digitized *then*. Therefore, the CW in the two versions had the same onset time (726 ms) measured from the beginning of the sentences. The visual stimuli were two pairs of pictures (“a blue rectangle and a red rectangle,” or “a red triangle and a red rectangle”). Two pictures in every pair were presented in succession. The spoken sentence was presented as soon as the second picture appeared.

In the “Audiovisual” situation, the CW *red* in the spoken sentence was consistently accented relative to the picture context in the standard stimulus. That is, a blue rectangle and a red rectan-

Table 1. Illustration for experimental materials

Audiovisual situation		
Standard (New-Accented)	 Then a **(RED)** rectangle appears. 随后(**红色**)的方形出现了	(75%)
Deviant1 (Given-Accented)	 Then a **"RED"** rectangle appears. 随后"**红色**"的方形出现了	(12.5%)
Deviant2 (New-Deaccented)	 Then a (red) **RECTANGLE** appears. 随后(红色)的**方形**出现了	(12.5%)
Auditory situation		
Standard (Accented)	Then a **RED** rectangle appears. 随后**红色**的方形出现了	
Deviant (Deaccented)	Then a red **RECTANGLE** appears. 随后红色的**方形**出现了	

Note. Two pictures were presented in succession. As soon as the second picture appeared, the spoken sentence was presented by auditory means. The picture □ indicates blue rectangle; the picture ■ indicates red rectangle; the picture ▲ indicates red triangle. Underline indicates the CWs in the sentence; brackets indicate new information relative to the picture context; quotes indicate old information relative to the picture context; capitalization and bold indicate the presence of accentuation.

gle, which appeared in succession, assigned new information status to the CW; meanwhile, the CW was consistently accented (New-Accented) (75%, see Table 1). In contrast, there was mismatch between accentuation and picture context in the two kinds of deviant stimuli. In one kind of deviant stimulus, the spoken sentence was the same as in the standard stimulus, but the picture context (a red triangle and a red rectangle) set the CW as given information, hence inconsistent with the accentuation on the CW (Given-Accented) (12.5%). In another kind of deviant stimulus, the picture context was the same as in the standard stimulus, but the CW was inconsistently deaccented (New-Deaccented) (12.5%).

In the "Purely Auditory" situation, the same spoken sentences from audiovisual situation were presented without pictures. The CW was accented in the standard stimulus (Accented) and deaccented in the deviant stimulus (Deaccented).

Procedure

After the electrodes were positioned, subjects were seated facing a computer screen in an acoustically and electrically shielded room. In all situations, they passively listened to the spoken sentences. In the audiovisual situation, each trial began with a 300 ms auditory tone, followed by the first picture (e.g. a blue rectangle) which was presented on the left side of the center screen. Then, 1000 ms after the appearance of the first picture, the second picture (e.g. a red rectangle) was presented on the right side of the center screen. The spoken sentence was presented in auditory means as soon as the second picture appeared. The two pictures did not disappear from the screen until the end of the spoken sentence. To ensure that subjects were at all times focusing the screen, they were asked to detect the changes in the pictures. In the purely auditory situation, the same spoken sentences as in the audiovisual situation were presented, but without the picture pairs. A white fixation cross (+) was presented simultaneously with the spoken sentence in the center of screen. The size of the fixation cross changed across trials. The subjects were asked to detect the changes in the fixation crosses. In both audiovisual and purely auditory situations, the subjects were informed that questions about what they saw on the screen would be asked afterword.

The effective measuring duration of about 80 min was divided into two sessions with five blocks for audiovisual situation and two blocks for auditory situation. The order of audiovisual and auditory situations was counterbalanced between subjects. In the audiovisual situation, each block contained 75% (96 trials) of the standard stimuli and 25% (32 trials) of the deviant stimuli with 12.5% for each type of inconsistent accentuation. In the auditory situation, only the spoken sentences were presented. In both situations, stimuli were presented with an inter-stimulus interval of 650 ms. In addition, every block began with additional 10 trials which were all standard stimuli.

EEG acquisition

EEG was recorded (0.01–40 Hz, sampling rate 500 Hz) from 64 Ag/AgCl electrodes mounted in an elastic cap. EEG and electrooculogram (EOG) data were amplified with a.c. amplifiers (Neuroscan). All electrode impedance levels (EEG and EOG) were kept below 5 kΩ. All electrodes were referenced to the left mastoid on-line. The EEG electrodes were re-referenced off-line to linked mastoids. Vertical eye movements were monitored via a supra-to sub-orbital bipolar montage. A right to left canthal bipolar montage was used to monitor horizontal eye movements.

ERP analysis

The first 10 trials of all blocks and the standards immediately following deviants were not included in the analysis. Since the number of deviant trials was necessarily smaller than the number of standard trials, in each block the same number of standard trials was randomly selected to match the number of deviants per condition. The raw EEG data were 0.1–40 Hz band-pass filtered and corrected for blink artifacts. Subsequently, EEG data were epoched from −100 to 700 ms relative to the acoustic onset of the CW and baseline corrected (100 ms pre-stimulus interval). Trials containing data exceeding a voltage criterion of ±75 μV were rejected (5.5% overall).

The mismatch response was obtained by subtracting the averaged response to the standard waveform from the averaged response to its corresponding deviant waveform respectively per subject, resulting in three difference waveforms: two for the audiovisual situation ("New-Deaccented" minus "New-Accented," "Given-Accented" minus "New-Accented") and one for the purely auditory situation (Deaccented minus Accented). As seen from Fig. 1, the negative deflections consisted clearly of two peaks: one peaking at about 100 ms and another peaking at about 210 ms from the acoustic onset of the CW. Given their latency and topography (see Fig. 1 and Fig. 2), we classified the first negative deflection as N1 effect, the second as MMN effect (Maess et al., 2007; Rinne et al., 2006).

Statistical analyses were done on the mean amplitudes in the following two latency windows: 80–120 ms (N1) and 190–230 ms (MMN) latency ranges following the acoustic onset of the CW (red). Analyses of variance were conducted on a selection of three midline electrodes (Fz, Cz, Pz) and six lateral electrodes (F3/F4; C3/C4; P3/P4). First, to estimate the three MMNs (two for the audiovisual situation and one for the purely auditory situation) respectively, three separate original ANOVAs were conducted with Stimulus type (deviant, standard), Laterality (left, midline, right) and Anteriority (F3/Fz/F4, C3/Cz/C4, P3/Pz/P4) as independent factors. Second, to compare the MMN elicited by the deviant speech in the purely auditory situation with the MMN elicited by the same deviant speech in the audiovisual situation, the difference ANOVAs were performed based on two difference waveforms

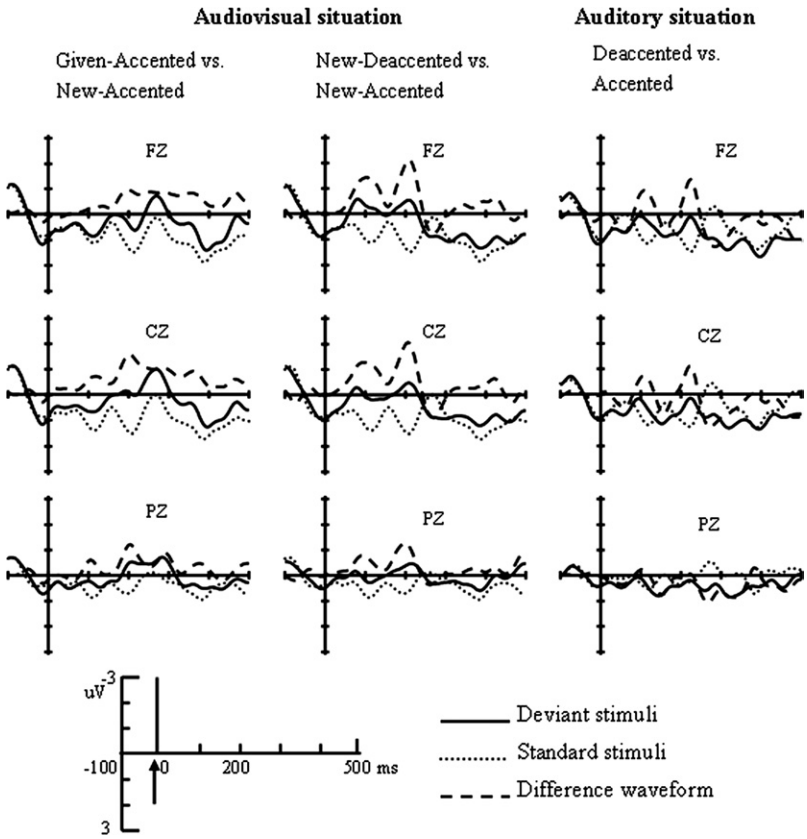


Fig. 1. Grand average ERPs at Fz, Cz, and Pz for the standard stimuli, deviant stimuli, and difference waveforms. Left, Given-Accented (deviant) vs. New-Accented (standard) in the audiovisual situation; middle, New-Deaccented (deviant) vs. New-Accented (standard) in the audiovisual situation; right, Deaccented (deviant) vs. Accented (standard) in the auditory situation.

("New-Deaccented" minus "New-Accented," Deaccented minus Accented). When the degree of freedom in the numerator was larger than 1, the Greenhouse-Geisser correction was applied.

RESULTS

Results in the 190–230 ms latency range (MMN)

In the time window of 190–230 ms, the original ANOVA, which compared New-Deaccented with New-Accented, revealed a significant main effect of Stimulus type [$F_{(1,15)}$= 30.73, $P<0.0001$], indicating that the deviant New-Deaccented evoked a larger negative deflection (MMN) than the standard New-Accented (effect size: -1.29 μV). In addition, there was a significant interaction between Stimulus type and Anteriority [$F_{(2,15)}$=9.67, $P<0.005$]. Subsequently simple-analyses showed that the MMN effect was significant at all of the three levels of Anteriority (frontal, central and parietal) [$F_{(2,15)}$=24.65, $P<0.0001$; $F_{(2,15)}$=40.62, $P<0.0001$; $F_{(2,15)}$=12.79, $P<0.005$, respectively], but reached maximum on the central sites. The original ANOVA comparing Given-Accented with New-Accented also results in a significant main effect of Stimulus type [$F_{(1,15)}$=22.26, $P<0.0001$], due to the fact that the deviant Given-Accented elicited a larger negative deflection than the standard New-Accented (effect size: -0.98 μV). In the purely auditory situation, the original ANOVA demonstrated that the deviant "Deaccented" evoked a significant larger negative deflection that the standard "Accentuated" [$F_{(1,15)}$=6.18, $P<0.05$] (effect size: -0.52 μV). The Stimulus type by Anteriority interaction was found to be significant too, due to the fact that the Stimulus type effect only reached significance on frontal and central sites [$F_{(2,15)}$= 13.04, $P<0.005$; $F_{(2,15)}$=6.65, $P<0.05$, respectively].

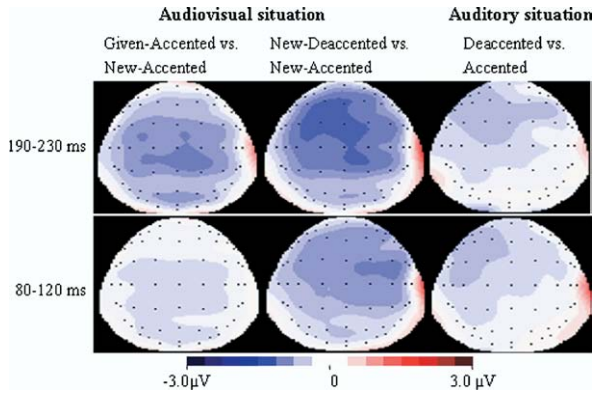


Fig. 2. Topography of the ERP effects for the deviant stimuli in the audiovisual and auditory situations in the 190–230 and 80–120 ms latency windows after the CW onset.

Then we compared the MMN evoked in the purely auditory situation with the MMN evoked by the same de-

viant speech sound in the audiovisual situation. First, mean amplitude in the 190–230 ms latency after CW onset was computed based on the two difference waveforms: Deaccented minus Accented, "New-Deaccented" minus "New-Accented." Second, timing of the negative deflection in the two difference waveforms was measured by determining individual peak latencies between 150 and 270 ms after CW onset. This time-window was set after visual inspection of the data and included the time window (190–230 ms) for the MMN in the present result. The mean amplitude or peak latency was submitted to difference ANOVAs with Situation (audiovisual, purely auditory), Laterality and Anteriority as independent factors. The difference ANOVA with mean amplitude as dependent factor resulted in a main effect of Situation, indicating that the MMN evoked in the audiovisual situation was stronger than that evoked by the same deviant speech stimuli in the purely auditory situation [$F_{(1,15)}=4.73$, $P<0.05$] (effect size: -0.78 μV). However, the difference ANOVA with peak latency as dependent factor found neither main effect of Situation [$F_{(1,15)}=1.48$, $P=0.234$] nor the interaction between Situation and other factors.

Results in the 80–120 ms latency range (N1)

In the time window of 80–120 ms, the three original ANOVAs revealed that: the deviant "Deaccented" elicited a marginally significant larger negative deflection that the standard "Accentuated" in the purely auditory situation [$F_{(1,15)}=3.86$, $P=0.068$; effect size: -0.50 μV]; in the audiovisual situation, the negative deflection evoked by the deviant New-Deaccented was significantly larger than that evoked by the standard New-Accented [$F_{(1,15)}=23.63$, $P<0.0001$; effect size: -0.96 μV]; there was no significant difference between the deflection evoked by the deviant Given-Accented and standard New-Accented [$F_{(1,15)}=1.97$, $P=0.181$].

The difference ANOVA with mean amplitude as dependent factor only resulted in a significant Situation by Laterality interaction [$F_{(2,15)}=5.32$, $P<0.05$]. Further analysis revealed that on all of the three levels of Laterality (left, midline and right), there was no significant difference between audiovisual and purely auditory situations.

DISCUSSION

This current study investigated the integration of prosodic information from speech with visual information from pictures in the absence of focused attention. The results demonstrated that: in the purely auditory situation, prosodic changes in speech evoked a MMN effect; in the audiovisual situation, changes in prosody-picture correspondence also evoked significant MMN effects. When comparing the purely auditory MMN with the MMN evoked by the same prosodic changes in the audiovisual situation, we found an enhancement of the MMN amplitude in the latter. Furthermore, the prosodic changes were processed at the same latency in the audiovisual situation as in the purely auditory situation. We interpreted this pattern of MMN effects as reflecting immediate integration of the prosodic information from speech with visual information from pictures. In the subsequent sections, we discussed those results in more detail.

On-line integration of prosodic and visual information under non-attentional condition

In the audiovisual situation, relative to the standard stimuli (consistent accentuation: "New-Accented"), both kinds of deviant stimuli (inconsistent accentuation: "Given-Accented" and "New-Deaccented") evoked MMN effects respectively. We claimed that those MMN effects were specific to the audiovisual situation. On the one hand, the MMN evoked by the deviant speech sound in the audiovisual situation (New-Deaccented vs. New-Accented) was significant larger than that evoked by the same deviant speech sound in the purely auditory situation (Deaccented vs. Accented). This enhancement could not be explained by the mere auditory difference, but might be also related to the speech–picture correspondence. On the other hand, for the deviant "Given-Accented" and standard "New-Accented," although the speech sentence was exactly the same, their picture contexts were different. Thus, the speech–picture correspondence differed between them. We claimed that the MMN evoked by "Given-Accented" was also related to the speech–picture correspondence. An alternative explanation is that the MMN was evoked by the pure picture contrast, but not by the audiovisual contrast. This was disproved by the convergence of the ERPs to the word "then" (seen Fig. 3), since the picture context had already deviated at the beginning of the sentence (namely, the word "then"). In summary, since the MMN evoked by the two kinds of deviant in the audiovisual situation was neither pure auditory nor pure visual, we interpreted it as reflecting the on-line audiovisual integration. The prosodic information from speech sentence could be integrated with visual information from pictures immediately during speech processing.

The MMN is known to be evoked automatically (Näätänen et al., 1993). In the current study, there were no task requirements related to the speech sentence or the speech–picture correspondence. Thus, the current results might reflect an automatic detection of the changes in speech–picture correspondence. It was indicated that the human brain could integrate the prosodic information from speech with visual information from pictures under non-attentional condition.

The time characteristics of the audiovisual integration process

This finding of speech–picture integration contributed to previous evidence of audiovisual interplay. Considerable behavioral, ERP, and fMRI studies have already demonstrated the audiovisual effect on perceptual processing (e.g. Mishra et al., 2007; Watkins et al., 2006; McDonald et al., 2003; Molholm et al., 2002; Shams et al., 2000; Giard and Peronnet, 1999). For example, Shams et al. (2000) found that a single flash can be misperceived as two flashes if paired with two beeps. McDonald et al. (2003) reported that the visual P1 component could be modified

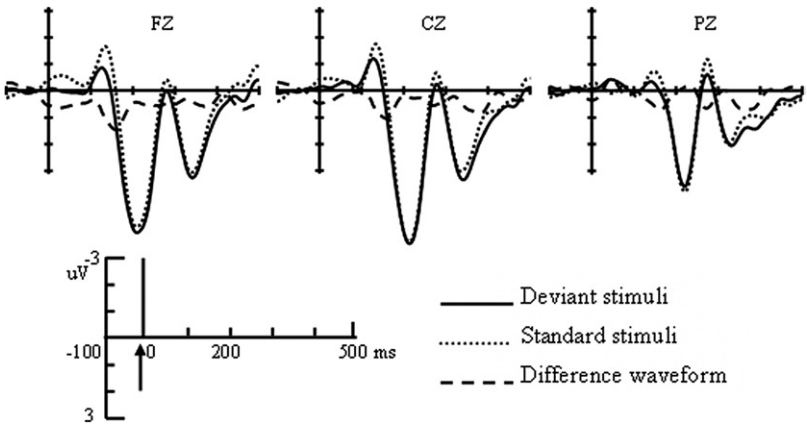


Fig. 3. Grand average ERPs aligned to words that immediately precede the CWs in the deviant (Given-Accented) and standard (New-accented) stimuli in the audiovisual situation.

by the relative location of a task-irrelevant sound with respect to the visual event. The current results added further evidence to this literature by showing that the early audiovisual integration could occur when people are presented with more natural and meaningful combinations of auditory and visual stimuli.

A key issue in the research on audiovisual perception is at which stage of information processing acoustic and visual inputs are integrated. One hypothesis is that multisensory integration occurs relatively late, following extensive processing of single sensory input. Feedback connections may exist from higher-level multisensory regions, back to lower-level areas. (e.g. Bonath et al., 2007; McDonald et al., 2003; Calvert et al., 2000; Massaro, 1999.) An alternative proposal assumed that the multisensory convergence exists at all levels of cortical hierarchy. There are direct feedforward influences between visual and auditory processing (e.g. Driver and Noesselt, 2008; Foxe and Schroeder, 2005). In the current study, changes in speech–picture correspondence elicited MMNs with around 210 ms peak latency. Importantly, this audiovisual combining process was as fast as the purely auditory processing, as indicated by the same peak latencies in the audiovisual and purely auditory situations. Those results could not be explained by the later-integration hypothesis, because according to the later-integration hypothesis, the detection of the audiovisual changes would occur later than the detection of purely auditory changes. In contrast, the current results support that the auditory and visual input might be integrated at an early stage of perceptual processing, and that the auditory inputs to this audiovisual process arrive through feedforward connections. This early audiovisual integration is in agreement with ERP studies which have demonstrated early modulation of activity in the auditory cortex caused by visual components of audiovisual stimuli (e.g. Giard and Peronnet, 1999; Möttönen et al., 2002; Molholm et al., 2002). All of those results indicated that the audiovisual integration could occur at the early stage of perceptual processing, and it utilizes the full range of anatomical connections types, including feedforward as well as feedback and lateral.

Indication in the literature of language comprehension

The current results had also important indications in the literature of language comprehension. As introduced in the introduction, previous studies have already demonstrated that the semantic or syntactic information from speech can be integrated with visual information immediately during language comprehension (e.g. Knoeferle et al., 2005; West and Holcomb, 2002; Tanenhaus et al., 1995; Willems et al., 2008; Özyürek et al., 2007). The current experiment extended those results by showing that not only semantic or syntactic information (segmental information) but also prosodic information (supersegmental information) can be immediately integrated with the visual information. Furthermore, the current results suggested that, even in the absence of focused attention, the language system still can make combined use of different sources of information, such as prosodic information from speech and visual information from pictures.

In addition, the current results adds to the understanding of the language processing system as taking different sources of information into account at the same time (e.g. Willems et al., 2008; Tanenhaus and Trueswell, 1995). Within the study of the cognition of language, the issue of how linguistic and extralinguistic information is integrated is reflected in the distinction between one-step and two-step models of language comprehension. The implication of two-step models is that, first, the meaning of a sentence is computed and second, the sentence meaning is integrated with extralinguistic information (e.g. Lattner and Friederici, 2003). In contrast, underlying the one-step model is the “immediacy assumption,” according to which every source of information that constrains the interpretation of an utterance can, in principle, do so immediately. In the current study, the detection of the changes in speech–picture correspondence was as fast as the detection of purely speech changes. This pattern of results was inconsistent with the classic two-step model of interpretation, because the two-step model would expect the integration of mere linguistic information to precede the integration of nonlinguistic information. Instead, the current result was in

line with the immediacy assumption, all available information is used immediately to co-determine the interpretation of linguistic expressions.

CONCLUSION

Our results showed that changes in prosody-picture correspondence influences the magnitude of MMN response. Importantly, the detection of this audiovisual change was as fast as the detection of purely auditory change. We interpreted these results as reflecting immediate integration of prosodic information from speech with visual information from pictures under a non-attentional condition. These findings bear relevance to audiovisual integration research by suggesting that the auditory and visual input can be integrated at an early stage of processing, and that there might be direct feedforward influences between visual and auditory processing. In addition, the current results add to the understanding of the language processing system as taking different sources of information into account at the same time.

Acknowledgments—This research was supported by Grants from the National Natural Science Foundation of China (30800296), Project for Young Scientist Fund, Institute of Psychology, Chinese Academy of Sciences (07CX122012). We thank the anonymous reviewer for critical comments and helpful suggestions on an earlier draft.

REFERENCES

Bock JK, Mazzella JR (1983) Intonational marking of given and new information: some consequences for comprehension. Mem Cogn 11:64–76.

Bonath B, Noesselt T, Martinez A, Mishra J, Schwiecker K, Heinze HJ, Hillyard SA (2007) Neural basis of the ventriloquist illusion. Curr Biol 17:1697–1703.

Calvert GA, Campbell R, Brammer MJ (2000) Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Curr Biol 10:649–657.

Colin C, Radeau M, Soquet A, Demolin D, Colin F, Deltenre P (2002) Mismatch negativity evoked by the McGurk–MacDonald effect: a phonetic representation within short-term memory. Clin Neurophysiol 113:495–506.

Dahan D, Tanenhaus MK, Chambers CG (2002) Accent and reference resolution in spoken-language comprehension. J Mem Lang 47: 292–314.

Driver J, Noesselt T (2008) Multisensory interplay reveals crossmodal influences on "sensory-specific" brain regions, neural responses, and judgments. Neuron 57:10–23.

Federmeier KD, Kutas M (2001) Meaning and modality: influences of context, semantic memory organization, and perceptual predictability on picture processing. J Exp Psychol Learn 27:202–224.

Foxe JJ, Schroeder CE (2005) The case for feedforward multisensory convergence during early cortical processing. Neuroreport 16: 419–423.

Froyen D, Van Atteveldt N, Bonte M, Blomert L (2008) Cross-modal enhancement of the MMN to speech-sounds indicates early and automatic integration of letters and speech-sounds. Neurosci Lett 430:23–28.

Ganis G, Kutas M, Sereno MI (1996) The search for "common sense": an electrophysiological study of the comprehension of words and pictures in reading. J Cogn Neurosci 8:89–106.

Giard MH, Peronnet F (1999) Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. J Cogn Neurosci 11:473–490.

Hruska C, Alter K, Steinhauer K, Steube A (2000) ERP effects of sentence accents and violations of the information structure. In Poster presented at the 13th annual CUNY conference on human sentence processing, San Diego, CA.

Johnson SM, Clifton CE, Breen ME, Martin AE, Florak JM (2003) ERP investigation of prosodic and semantic focus. Poster presented at Cognitive Neuroscience meeting, New York City.

Knoeferle P, Crockera MW, Scheepersb C, Pickering MJ (2005) The influence of the immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events. Cognition 95:95–127.

Lattner S, Friederici AD (2003) Talker's voice and gender stereotype in human auditory sentence processing-evidence from event-related brain potentials. Neurosci Lett 339:191–194.

Li XQ, Hagoort P, Yang YF (2008a) Event-related potential evidence on the influence of accentuation in spoken discourse comprehension in Chinese. J Cogn Neurosci 20:1–10.

Li XQ, Yang YF, Hagoort P (2008b) Pitch accent and lexical tone processing in Chinese discourse comprehension: an ERP study. Brain Res 1222:192–200.

Li XQ, Yang YF (2004) The role of accentuation in spoken discourse comprehension. Acta Psychol Sin 36:393–399.

Li XQ, Yang YF (2005) The influence of inconsistent accentuation on activation of information during spoken discourse processing. Acta Psychol Sin 37:285–290.

Maess B, Jacobsen T, Schröger E, Friederici AD (2007) Localizing pre-attentive auditory memory-based comparison: magnetic mismatch negativity to pitch change. Neuroimage 37:561–571.

Magne C, Astésano C, Lacheret-Dujour A, Morel M, Alter K, Besson M (2005) On-line processing of "pop-out" words in spoken French dialogues. J Cogn Neurosci 17:740–756.

Massaro DW (1999) Speechreading: illusion or window into pattern recognition. Trends Cogn Sci 3:310–317.

McDonald JJ, Teder-Salejarvi WA, Di Russo F, Hillyard SA (2003) Neural substrates of perceptual enhancement by cross-modal spatial attention. J Cogn Neurosci 15:10–19.

McGurk, MacDonald JW (1976) Hearing lips and seeing voices. Nature 264:746–748.

Mishra J, Martinez A, Sejnowski TJ, Hillyard SA (2007) Early cross-modal interactions in auditory and visual cortex underlie a sound-induced visual illusion. J Neurosci 27:4120–4131.

Molholm S, Ritter W, Murray MM, Javitt DC, Schroeder CE, Foxe JJ (2002) Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study. Cogn Brain Res 14:115–128.

Möttönen R, Krause CM, Tiippana K, Sams M (2002) Processing of changes in visual speech in the human auditory cortex. Cogn Brain Res 13:417–425.

Näätänen PP, Tiitinen H, Jiang D, Alho K (1993) Attention and mismatch negativity. Psychophysiology 30:336–350.

Näätänen R (2001) The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). Psychophysiology 38:1–21.

Özyürek A, Willems RM, Kita S, Hagoort P (2007) On-line integration of semantic information from speech and gesture: insights from event-related brain potentials. J Cogn Neurosci 19:605–616.

Rinne T, Särkkä A, Degerman A, Schröger E, Alho K (2006) Two separate mechanisms underlie auditory change detection and involuntary control of attention. Brain Res 1077:135–143.

Sams M, Aulanko R, Hamalainen M, Hari R, Lounasmaa OV, Lu ST, et al. (1991) Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. Neurosci Lett 127(1): 141–145.

Shams L, Kamitani Y, Shimojo S (2000) Illusions. What you see is what you hear. Nature 408:788.

Spivey MJ, Tanenhaus MK, Eberhard KM, Sedivy JC (2002) Eye-movements and spoken language comprehension: effects of visual context on syntactic ambiguity resolution. Cogn Psychol 45:447–481.

Tanenhaus MK, Spivey MJ, Eberhard KM, Sedivy JC (1995) Integration of visual and linguistic information in spoken language comprehension. Science 268:1632–1634.

Terken J, Noteboom SD (1987) Opposite effects of accentuation and deaccentuation on verification. Latencies for given and new information. Lang Cogn Proc 2:145–163.

Wang B, Lü SN, Yang YF (2002) The pitch movement of stressed syllable in Chinese sentences. Acta Acustica 27(3):234–240.

Watkins S, Shams L, Tanaka S, Haynes JD, Rees G (2006) Sound alters activity in human V1 in association with illusory visual perception. Neuroimage 31:1247–1256.

West WC, Holcomb PJ (2002) Event-related potentials during discourse-level semantic integration of complex pictures. Cogn Brain Res 13:363–375.

Willems RM, Özyürek A, Hagoort P (2008) Seeing and hearing meaning: ERP and fMRI evidence of word versus picture integration into a sentence context. J Cogn Neurosci 20:1235–1249.

(Accepted 19 January 2009)
(Available online 4 February 2009)