Neuroscience 180 (2011) 181–190

COGNITIVE INTEGRATION OF ASYNCHRONOUS NATURAL OR NON-NATURAL AUDITORY AND VISUAL INFORMATION IN VIDEOS OF REAL-WORLD EVENTS: AN EVENT-RELATED POTENTIAL STUDY

B. LIU,* Z. WANG, G. WU AND X. MENG

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, PR China

Abstract—In this paper, we aim to study the cognitive integration of asynchronous natural or non-natural auditory and visual information in videos of real-world events. Videos with asynchronous semantically consistent or inconsistent natural sound or speech were used as stimuli in order to compare the difference and similarity between multisensory integrations of videos with asynchronous natural sound and speech. The event-related potential (ERP) results showed that N1 and P250 components were elicited irrespective of whether natural sounds were consistent or inconsistent with critical actions in videos. Videos with inconsistent natural sound could elicit N400-P600 effects compared to videos with consistent natural sound, which was similar to the results from unisensory visual studies. Videos with semantically consistent or inconsistent speech could both elicit N1 components. Meanwhile, videos with inconsistent speech would elicit N400-LPN effects in comparison with videos with consistent speech, which showed that this semantic processing was probably related to recognition memory. Moreover, the N400 effect elicited by videos with semantically inconsistent speech was larger and later than that elicited by videos with semantically inconsistent natural sound. Overall, multisensory integration of videos with natural sound or speech could be roughly divided into two stages. For the videos with natural sound, the first stage might reflect the connection between the received information and the stored information in memory; and the second one might stand for the evaluation process of inconsistent semantic information. For the videos with speech, the first stage was similar to the first stage of videos with natural sound; while the second one might be related to recognition memory process.

Key words: video, sound, speech, N400-P600, N400-LPN.

In daily life, most external information is received from visual and auditory senses. Visual and auditory information are received separately and integrated in the human brain, and thus people get a comprehensive understanding of the outside world. Therefore, the study of visual and auditory information integration processing is important to reveal the cognitive mechanism in the human brain.

In recent years, studies on unisensory visual semantic perceptions of natural videos have been reported. In 2003, Sitnikova et al. studied semantic integration processing in videos. A strong negative wave was found in the ERP record when inappropriate objects appeared in video clips of common activities. Meanwhile, a central and posterior distributed positive wave was found after the negative wave (Sitnikova et al., 2003). These two waves were similar to the N400-P600 pattern found in language cognitive studies (Liu et al., 2009a, 2010b,c, 2011a). It is suggested that the cognitive processing of language and natural videos was similar. In 2008, Sitnikova et al. conducted a further experiment on semantic integration in videos. They found that completely unexpected endings in videos could elicit an N400 effect, and goal-related unexpected endings could elicit an N400 effect and a late positive component. They suggested that there might be two independent cognitive processes in the human brain involved in the comprehension of the visual actual world. The first cognitive processing is reflected by the N400 effect and stands for the connection between the received information and the semantic memory in the human brain. The second cognitive processing is reflected by the positive wave and stands for the evaluation of received information which is inconsistent with actual world actions (Sitnikova et al., 2008). In these studies, some of the experimental stimuli contained switches of shooting angles, but it is found that the switches of shooting angles in videos could not influence the semantic integration of videos (Liu et al., 2010d).

In studies on auditory semantic information, speech and natural sounds are also two major research points. In Van Pettern and Rheinfelder's study, they found that semantic inconsistent speech or natural sounds could both elicit similar N400 effects, and that there was no significant difference between these two N400 effects (Van Petten and Rheinfelder, 1995). Therefore, the semantic processing of speech and natural sounds is similar. However, speech and natural sounds are still two different kinds of auditory information, and their semantic functions might be different.

In recent years, many studies on multisensory information have been reported. In 2007, Puce et al. used the human face, monkey face and house as visual stimuli, and the human burp, monkey scream and house creak as auditory stimuli to study cognitive processing on matched or mismatched pictures and sounds in the human brain (Puce et al., 2007). They found that a significant P400 effect was found when the semantic information of pictures and sounds were inconsistent. No N400-P600 waveforms were found in their experiment. In 2009, we improved on their experimental method. Videos with semantic consistent or inconsistent natural sounds and critical actions were used as stimuli in our experiment. We also found that

*Corresponding author. Tel: +86-10-62781789; fax: +86-10-62771138.
E-mail address: liubaolin@tsinghua.edu.cn (B. Liu).

Abbreviation: EEG, electroencephalogram; ERP, event-related potential; LPN, late posterior negativity; VN+, video with consistent natural sound; VN-, video with inconsistent natural sound; VS+, video with consistent speech; VS-, video with inconsistent speech.

0306-4522/11 $ - see front matter © 2011 IBRO. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.neuroscience.2011.01.066

inconsistent audio-visual information could elicit the P400 effect (Liu et al., 2009b). It is suggested that the P400 effect might reflect the semantic integration of synchronous audio-visual information. However, different results have been found in studies on asynchronous audio-visual information (Cummings et al., 2008; Plante et al., 2000). In Cummings et al.'s experiment, pictures were presented first, followed by speech or natural sounds which were semantically consistent or inconsistent with the pictures that were presented. They found that semantically inconsistent speech or natural sounds with pictures could elicit N400 effects (Cummings et al., 2008).

The question arises over whether the integration processes of videos with speech and natural sound are similar. As an inevitable consequence, the important issue is whether both semantically inconsistent natural sounds and speech with critical actions could elicit typical N400-P600 effects. Therefore, in this experiment, we would use videos of real-world events as experimental stimuli to study the integration of speech or natural sounds and critical actions of videos in the human brain. We would observe the ERP waveforms elicited by videos with speech and natural sound in order to determine the difference and similarity of their semantic integrations in the human brain.

EXPERIMENTAL PROCEDURES

Participants

Eighteen students from Tsinghua University (nine males, nine females, mean age 21.2 years (SD=1.8)) participated in the experiment. They had normal or corrected-to-normal vision and normal hearing. None of them were color blind.

All subjects had no history of neurological diseases, and were free of medication for at least 1 week before the experiment. They were all judged to be right-handed according to the Edinburgh Handedness Inventory (Oldfield, 1971). Before the experiment, they were told that our experiment was conducted according to the Declaration of Helsinki, and signed the Researchers' Consent Form. Each participant was paid 35 Yuan (RMB) per hour for his/her participation.

Materials

Video clips of real-world events were selected as visual stimulus materials in our experiment. Each of them contained a critical action, and expressed a natural scene. Natural sounds and speech which were semantically consistent or inconsistent with the critical actions were selected as auditory stimulus materials (average length 293 ms (SD=53)). For speech, we first selected Chinese verbs (consisting of two Chinese characters) which were semantically consistent or inconsistent with the critical actions, and then the verbs were recorded as speech by a male voice reading. Finally, the natural sounds or speech were integrated with the original videos respectively, where the auditory information was semantically consistent or inconsistent with the critical action and occurred 1000 ms after the onset of critical action.

The stimulus materials consisted of 24 video clips with durations of between 3600 ms to 4000 ms (mean duration 3805 ms (SD=168)), and were divided into six groups according to different events. Each group included four video clips, from the same original color videos, which were respectively: (1) Video with consistent natural sound, where natural sound is semantically consistent with critical action. This kind of video is defined as (VN+). (2) Video with inconsistent natural sound, in which natural sound is semantically inconsistent with critical action. This kind of video is defined as (VN−). (3) Video with consistent speech, where speech is semantically consistent with critical action. This kind of video is defined as (VS+). (4) Video with inconsistent speech, in which speech is semantically inconsistent with critical action. This kind of video is defined as (VS−).

The six original color video clips were as follows: (1) a large display of festive fireworks exploding in the sky; (2) a glass falling from a height and shattering; (3) a small ball on a horizontal board knocking another ball; (4) a cue-stick striking a white billiard ball; (5) a person with his back to the screen going out and closing the door; (6) a water droplet falling down with a splash. A critical action was defined in each video; for example, the moment when the festive fireworks explode in the sky in the first video. Fig. 1 illustrates video frames of naturalistic motion stimuli.

In the selection of the original video materials, we complied with the following principles: (1) we featured commonplace actions and sounds in the videos, in order to ensure that participants were familiar with them; (2) the videos were clear, with no pauses or switches in shooting angles (Liu et al., 2010d); (3) the critical actions were instantaneous actions, as shown in Fig. 1, with the critical action being the moment when the festive fireworks explode in the sky; (4) in the moments preceding the critical action, subjects would be guided by the context of the video to have an expectation consistent to the following critical action; (5) the difference between the inconsistent sounds and expected sounds should be significant; (6) the onset of the critical action was at least 1000 ms after the beginning of the video and at least 1000 ms before the end of the video, in order to avoid influence from the presentation and conclusion of the video clips.

Adobe Premiere Pro software was used to edit the original video clips, using the NTSC standard with a resolution of

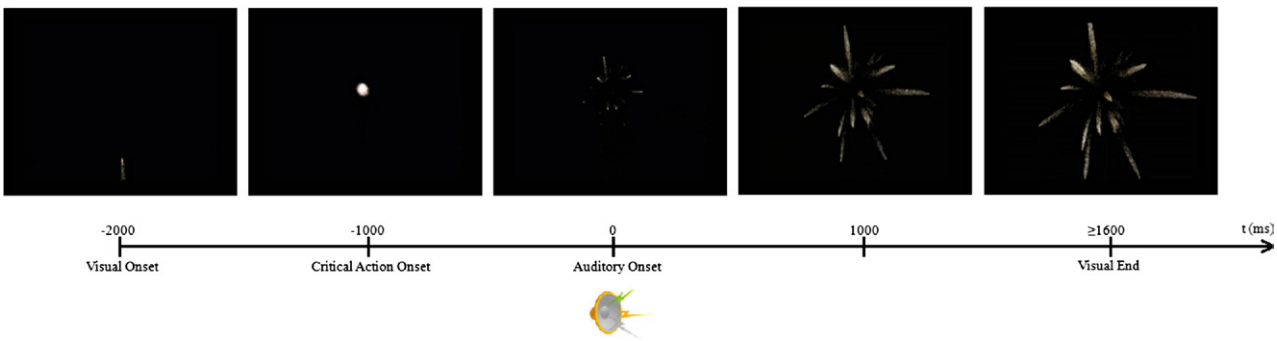

Fig. 1. Illustrational video frames of naturalistic motion stimuli. The critical action onset precedes the auditory onset by 1000 ms. For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.