Acta Psychologica 171 (2016) 1–16

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy

Modulation of scene consistency and task demand on language-driven eye movements for audio-visual integration

Wan-Yun Yu [a], Jie-Li Tsai [a,b,*]

[a] Department of Psychology, National Chengchi University, Taiwan
[b] Research Center for Mind, Brain, and Learning, National Chengchi University, Taiwan

ARTICLE INFO

Article history:
Received 7 May 2015
Received in revised form 8 September 2016
Accepted 11 September 2016
Available online 15 September 2016

Keywords:
Audio-visual integration
Comprehension
Scene consistency
Spoken language

ABSTRACT

Previous psycholinguistic studies have demonstrated that people tend to direct fixations toward the visual object to which spoken input refers during language comprehension. However, it is still unclear how the visual scene, especially the semantic consistency between object and background, affects the word-object mapping process during comprehension. Two visual world paradigm experiments were conducted to investigate how the scene consistency dynamically influenced the language-driven eye movements in a speech comprehension and a scene comprehension task. In each trial, participants listened to a spoken sentence while viewing a picture with two critical objects: one is the mentioned target object (e.g., *tiger*), which was embedded in either a consistent (e.g., *field*), inconsistent (e.g., *sky*) or blank background; the other is an unmentioned non-target object (e.g., *eagle*), which was always consistent with its background. The results showed that the fixation proportion of the inconsistent target was higher than the consistent target, and the task demand can affect the strength and the direction of the inconsistency effect before and after the target had been mentioned. In summary, the spoken language, scene-based knowledge and task demand were intertwined to determine eye movements during audio-visual integration for comprehension.

1. Introduction

People tend to perceive and understand the complex world from multisensory information. For example, when watching TV news, the audience receives visual and audio information concurrently: one is a video clip or a static scene portraying the news event, while the other is the sound of utterances from the reporter narrating the news event. In order to understand the news, people need to simultaneously integrate the audiovisual information about the whole story.

This cross-modal integration phenomenon has been a topic of particular interest among researchers in psycholinguistics and visual attention. Huettig, Olivers, and Hartsuiker (2011) proposed a working memory model to explain language-vision interaction during audio-visual integration. In this model, all sensory inputs (e.g., visual display, spoken utterance) can be mapped onto stored attributes (i.e., visual shape, phonological codes, semantic representations) from long-term memory. Then, the activated information can be integrated into the working memory to guide both covert and overt attention to serve the cognitive system so as to understand the world. Although this model provides a comprehensive framework for explaining the underlying mechanism, many questions regarding the phenomenon have not yet been fully examined. The aim of this study was therefore to investigate the influence of scene-based knowledge on spoken language processing through combining the research traditions of scene viewing and speech comprehension.

1.1. Scene consistency in unimodal scene viewing tasks

When viewing a real-world scene, since objects tend to co-occur within contexts, the human cognitive system can quickly grasp the global meaning and utilize this contextual information to facilitate scene perception. Oliva (2005) defined this kind of global knowledge (or *gist*) as a general spatial representation of the real-world scene that the visual system could extract from a single glance. The gist contains both perceptual and conceptual knowledge that viewers can access rapidly and which helps them establish the rough category of the scene (i.e., manmade versus natural, animal versus non-animal) in a parallel fashion (Joubert, Rousselet, Fize, and Fabre-Thorpe, 2007; Li, VanRullen, Koch, and Perona, 2002). After the initial preattentive stage of extracting the scene's gist, objects which match the background can be identified more accurately and faster than those with an inconsistent background (Biederman, Mezzanotte, and Rabinowitz, 1982; Boyce, Pollatsek, and Rayner, 1989; Davenport and Potter, 2004). For example, Biederman et al. (1982) found that both physical violations (e.g., a floating fire hydrant) and semantic violations (e.g., the hydrant in a kitchen) caused disruptions to object perception with brief exposure (150 ms),

* Corresponding author at: Department of Psychology, National Chengchi University, No. 64, Sec. 2, ZhiNan Rd., Wenshan, Taipei 11605, Taiwan.
E-mail address: jltsai@nccu.edu.tw (J.-L. Tsai).

http://dx.doi.org/10.1016/j.actpsy.2016.09.004
0001-6918/

indicating that viewers can access the object-background relations within a single fixation duration. In general, there are two functional roles of global scene gist in local object processing: one is to generate possible candidates of objects in the scene (Bar, 2004; Bar et al., 2006; Wolfe, Võ, Evans, and Greene, 2011), and the other is to constrain the probable spatial location of these candidates (Neider and Zelinsky, 2006; Wolfe et al., 2011). Furthermore, the role of scene gist in determining eye fixations on objects was proposed by the Contextual Guidance Model (Torralba, Oliva, Castelhano, and Henderson, 2006), which assumed that the fixation locations in a scene were computed based on the low-level saliency (i.e., orientation, intensity, contrast) and the distribution of object locations conditioned on the global representation of scene type.

While most investigators agree with the fact that the contextual information has its unique contribution to object recognition, there exists a discrepancy in the temporal change of the scene consistency effect. Several studies have reported that inconsistent objects preferentially attract covert attention within an initial glance, causing shorter onset latency and longer first fixation duration and total viewing time in visual search and memorization tasks (Bonitz and Gordon, 2008; Loftus and Mackworth, 1978; Underwood and Foulsham, 2006; Underwood, Humphreys, and Cross, 2007; Underwood, Templeman, Lamming, and Foulsham, 2008). This evidence converged to reveal that viewers could access the semantic relation between object and background at the early stage of scene viewing. However, some studies failed to replicate the potency of the early disruption of anomalous objects (De Graef, Christiaens, and D'Ydewalle, 1990; Henderson, Weeks, and Hollingworth, 1999; Võ and Henderson, 2011). For example, De Graef et al. (1990) found that inconsistent objects attracted more fixations only when multiple fixations had been made. Similarly, Võ and Henderson (2011) reported that the fixation advantage of inconsistent objects could only be observed once the 8th or 9th fixation had occurred. Therefore, these studies suggested that viewers prefer to fixate on the inconsistent objects in a scene, but this effect occurs relatively late in time. However, Bornstein, Mash, and Arterberry (2011) reported that participants had more fixations for consistent than for inconsistent scenes, but there was no congruency effect for the object in the scene. In general, previous studies have suggested that eye movements toward objects during scene viewing are affected by the object-background relation, whereas the temporal effect of scene consistency remains uncertain.

1.2. Language-driven eye movements in situated comprehension

Over the last two decades, many psycholinguistics researchers have used visual display as a reference, and the listeners' eye movements were analyzed to reveal the dynamic activation of mental representation in situated comprehension. Several studies which used the visual world paradigm have shown the efficient linking of speech with related objects in visual displays (Altmann and Kamide, 1999, 2007; Cooper, 1974; Huettig and Altmann, 2005; Huettig and McQueen, 2007; Michael K. Tanenhaus, Magnuson, Dahan, and Chambers, 2000; Michael K Tanenhaus, Spivey-Knowlton, Eberhard, and Sedivy, 1995). In a pioneering study, Cooper (1974) asked participants to listen to a short passage (e.g., a safari story) while looking at a visual display containing objects (e.g., a tree, snake, lion, camera) related to the story. The most important finding was that the participants spontaneously directed their gaze toward the visual object (e.g., a lion) upon hearing the identical word (e.g., lion) or a semantically related word (e.g., Africa) in a time-locked period. Later, Huettig and Altmann (2005) further demonstrated the semantic relatedness effect. They had participants hear the word 'piano' while viewing a four-object array: a piano and three semantically unrelated objects, a trumpet and three unrelated objects, or both a piano and a trumpet with two unrelated objects. Their results revealed that the fixation proportion on the piano and trumpet rose in the first two conditions, while only the fixations on the piano rose in the final condition, but the trumpet still received more fixations than the unrelated objects. Three conclusions can be drawn from these studies: (a) listeners can access the semantic representation from concurrent spoken words; (b) participants will map the semantic representation with the visual entities and choose the most closely related item to fixate on; and (c) this word-object mapping process can be monitored from eye movements on a millisecond time scale.

1.3. The influence of real-world scenes on audio-visual integration

Considering the audio-visual integration phenomenon, there are two limitations in the current psycholinguistic studies. First, the human vision system often encounters a more complex environment than the simplified visual display used in typical visual world paradigm experiments. Analogous to the 'array-size' concept in the visual search studies, the efficiency of searching for a target among distractors is determined by the number of items in a display. Thus, a large number of objects or the visual 'clutter' in real-world scenes may cause some difficulties in object identification, and impede the efficiency of linking a word with its visual referent. Sorensen and Bailey (2007) manipulated the set size of a visual display and found that increases in array size (i.e., 3×3, 4×4) led to a delay in the initiation of language-driven eye movements. Also, Ferreira, Foucart, and Engelhardt (2013) had participants listen to a garden-path sentence (e.g., *Put the book on the chair in the bucket*) while viewing object arrays with either low (4 objects) or high (12 objects) visual complexity. Compared with the low visual complexity condition, participants had difficulty inferring the correct object upon hearing the target noun (e.g., *book*), and took fewer looks at the correct location when the noun of the location (e.g., *bucket*) unfolded for the high visual complexity condition. In the study of Hintz and Huettig (2015), the critical trial included one unrelated object and three related objects while asking participants to judge if the target word (e.g., *beaker*) was absent or not in the speech. For these trials, they manipulated the visual complexity of the image by adding four items of human-like characters or meaningless visual shapes in contrast with the simple four-object array. The results showed that the increase in visual complexity enhanced the activation level of semantic representation and visual shape, while inhibiting the phonological activation of the visual objects. Therefore, viewing images with high visual complexity during comprehension caused three consequences: (a) delaying the overall efficiency of language-driven eye movements; (b) posing difficulties in language inference; and (c) altering the activation level of different representations from speech.

Second, using arbitrary object arrays as visual stimuli did not offer an opportunity to examine the role of scene-based information in spoken language processing. Since the human cognitive system has established a set of world knowledge and statistical regularities among objects and contexts through past life experiences, it is reasonable to assume that the scene consistency between object and background makes its individual contribution to eye guidance in situated comprehension. For this reason, Henderson (2005) raised a doubt that the tight linkage between spoken language and visual attention may be exaggerated in the VWP experiments—since the high-level contextual information was minimized in object arrays, the linguistic input became the only information source for the cognitive system to plan eye movements (Henderson and Ferreira, 2004). In addition, rather than causing difficulties in the word-object mapping process, real-world scenes may facilitate the visual search behavior upon linking a spoken word with its visual object through the accessibility of scene-based contextual information.

Recently, some researchers have noticed this issue and substituted traditional object arrays with real-world scenes to investigate language-vision interaction in situated comprehension (Andersson, Ferreira, and Henderson, 2011; Coco, Keller, and Malcolm, 2015; Staub, Abbott, and Bogartz, 2012). Andersson et al. (2011) used full-color photographs that depicted highly cluttered scenes (i.e., a storage