Consciousness and Cognition 21 (2012) 775–787

Contents lists available at SciVerse ScienceDirect

Consciousness and Cognition

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

Assessing the effects of audiovisual semantic congruency on the perception of a bistable figure

Jhih-Yun Hsiao[a], Yi-Chuan Chen[b,c], Charles Spence[b], Su-Ling Yeh[a,*]

[a] Department of Psychology, Graduate Institute of Brain and Mind Sciences, Neurobiology and Cognitive Science Center, National Taiwan University, Taiwan
[b] Department of Experimental Psychology, University of Oxford, UK
[c] Department of Psychology, Lancaster University, UK

ARTICLE INFO

Article history:
Received 13 May 2011
Available online 3 March 2012

Keywords:
Bistable perception
Crossmodal semantic congruency
Fixation
Attention
Visual consciousness

ABSTRACT

Bistable figures provide a fascinating window through which to explore human visual awareness. Here we demonstrate for the first time that the semantic context provided by a background auditory soundtrack (the voice of a young or old female) can modulate an observer's predominant percept while watching the bistable "my wife or my mother-in-law" figure (Experiment 1). The possibility of a response-bias account—that participants simply reported the percept that happened to be congruent with the soundtrack that they were listening to—was excluded in Experiment 2. We further demonstrate that this crossmodal semantic effect was additive with the manipulation of participants' visual fixation (Experiment 3), while it interacted with participants' voluntary attention (Experiment 4). These results indicate that audiovisual semantic congruency constrains the visual processing that gives rise to the conscious perception of bistable visual figures. Crossmodal semantic context therefore provides an important mechanism contributing to the emergence of visual awareness.

1. Introduction

When an observer looks at an ambiguous visual figure that affords two possible interpretations (such as the Necker cube or the Rubin face/vase figure), he/she may perceive either one of the interpretations initially, and then perceive the other one a short while thereafter. Subsequently, the observer's perception often spontaneously alternates between the two possible interpretations. This phenomenon, known as *bistable figure* perception, is critical in that the presentation of a static ambiguous figure can elicit a dynamically-alternating percept. In addition, observers are exposed to a coherent view in both eyes that is similar to their everyday vision. This fact discriminates the perception of bistable figures from the situation with binocular rivalry which occurs when different pictures are presented to each eye at the corresponding visual field position (see Kim & Blake, 2005). Hence, bistable figures provide an excellent means with which to investigate human visual consciousness in order to dissociate it from specific visual sensory inputs (Kim & Blake, 2005; Long & Toppino, 2004).

The processes involved in the perception of bistable figures in humans have been studied for more than a century now (see Vicholkovska, 1906, for a review of early research). Two well-known factors modulating the emergence of conscious perception of bistable figures have been proposed (see Long & Toppino, 2004, for a review). The first concerns the location in the figure where the observer fixates (Garcia-Perez, 1989; Necker, 1932). When observers fixate at the critical features that constitute one of the possible percepts, those features happen to fall in their fovea and hence benefit the recognition of that percept (Gale & Findlay, 1983). As a result, this percept is more likely to be dominant. Fixation position is considered

[*] Corresponding author. Address: Department of Psychology, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan. Fax: +886 2 23629909.

E-mail address: suling@ntu.edu.tw (S.-L. Yeh).

1053-8100/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.concog.2012.02.001

as a bottom-up factor modulating the perception of bistable figures (e.g., Meng & Tong, 2004). On the other hand, researchers have also attempted to investigate top-down, volitional effects when viewing bistable figures (e.g., Washburn & Gillette, 1933). For example, participants have been instructed to attend to a particular percept, while keeping fixation constant, in order to test the modulation of voluntary attention on people's conscious perception of bistable figures (Meng & Tong, 2004; van Ee, van Dam, & Brouwer, 2005).

In addition to visual fixation and attention, it has also been reported that semantic information can influence the perception of bistable figures. So, for example, Chastain and Burnham (1975) demonstrated that the participants' initial percept of the rat-man bistable figure could be primed by the prior presentation of an unambiguous figure which contained the features of a rat (the tail of the rat) or a man (the glasses worn by the man). In a more recent study, Daelli, van Rijsbergen, and Treves (2010) also demonstrated that participants' initial interpretation of an intermediate morph, such as between a tree and an umbrella, can be modulated by the prior presentation of either a tree or an umbrella. These can be considered as *priming* effects that influence the initial interpretation of a bistable figure. However, the priming effects in these two studies may be elicited by *perceptual priming* (such as by grouping the fragments in the same way as the prime), but need not necessarily involve *semantic priming* by the preceding unambiguous figures. The semantic modulation of the perception of a bistable figure reported by Balcetis and Dale (2007) plausibly avoided this form of visual-feature priming. Their participants first read a short story, and their subsequent interpretation of a bistable figure was biased so as to be consistent with the semantic content that had been embedded in the story (see also Feist & Gentner, 2007; Goolkasian & Woodberry, 2010). Note that, in this case, it is possible that the semantic context provided by the story may have directed participants' attention to certain clues that led to a particular perceptual outcome (see Balcetis & Dale, 2007, p. 593).

In the present study, a novel method was used to elicit a genuine semantic priming effect while bypassing the influences of visual feature analysis. That is, the semantic context was presented in another sensory modality, namely audition. Human perception normally involves the stimulation of more than one sensory modality (e.g., see Calvert, Spence, & Stein, 2004), and crossmodal semantic congruency is special because it serves as a heuristic with which to associate the information from different sensory modalities and bind them into unitary representations (*the unity assumption*, see Spence, 2007; Welch & Warren, 1980). These heuristics are likely established according to our everyday experience in the real world (see Ernst, 2007), and, as a result, modulate the perceptual outcome of visual processing in humans (e.g., Chen & Spence, 2010; Iordanescu, Grabowecky, Franconeri, Theeuwes, & Suzuki, 2010; Schneider, Engel, & Debener, 2008). Yet, to date, only a few studies have attempted to investigate whether the semantic information available in one sensory modality can help people to interpret ambiguous stimuli presented in another sensory modality (e.g., see Hupé, Joffo, & Pressnitzer, 2008, for a demonstration of the independence of the perception of bistable stimuli in the visual and auditory modalities).

Using crossmodal stimulus presentation in order to provide a semantic context for the perception of bistable figures ensures, first of all, that the modulatory effect can be genuinely attributed to the semantic cues rather than to any form of visual-feature priming. In addition, one can manipulate the semantic context by presenting different auditory soundtracks while keeping the visual stimuli (i.e., the bistable figure) constant for a long period of time. This allowed us to further compare the modulations induced by crossmodal semantic congruency under different biased states caused by manipulating the unimodal factors such as fixation and attention (Meng & Tong, 2004; van Ee et al., 2005). That is, when participants are listening to a soundtrack that provides a particular semantic context, they can also fixate a specific position or continuously attend to a particular percept during the test period.

In the present study, we examined the crossmodal semantic effect of sound on the perception of a visually-presented bistable figure. Specifically, we presented the "my wife or my mother-in-law" bistable figure (Boring, 1930) as the visual stimulus, whilst playing a monologue spoken by either an old woman or by a young lady. Note that the semantic congruency was conveyed by the voice rather than by the content of the speech (e.g., Smith, Grabowecky, & Suzuki, 2007), given that we presented the speech in French, a language with which our participants were completely unfamiliar. After having demonstrated that people's reports concerning their predominance perception of this bistable figure could be modulated by the soundtrack that they happened to hear, we then went onto test whether this crossmodal modulatory effect interacted with the fixation or attention manipulations. Note that we were not only interested in participants' initial percept when encountering the bistable figure (as in several previous studies; see Balcetis & Dale, 2007; Chastain & Burnham, 1975; Daelli et al., 2010; Goolkasian & Woodberry, 2010), but also in the proportion of time for which each percept was dominant within a given exposure period (Meng & Tong, 2004; van Ee et al., 2005). Presumably, these two measures reveal the initial and ongoing competition between two possible percepts, respectively.

2. Experiment 1

In this experiment, the participants viewed the "my wife or my mother-in-law" bistable figure whilst listening to the voice of an old woman, the voice of a young lady, a series of beeps, or else to no sound. The voice, rather than the content of the monologue per se, provides a crossmodal context to the old women or the young lady percept. The participants had to continuously report either the old woman or young lady being dominant, or else that they experienced a mixed percept (this occurs when a participant cannot determine which percept is dominant). We provided the final alternative (i.e., mixed percept) in order to reduce the possibility that the participants simply reported the percept that was associated with the sound they were hearing when they were experiencing an ambiguous (or undetermined) perception.

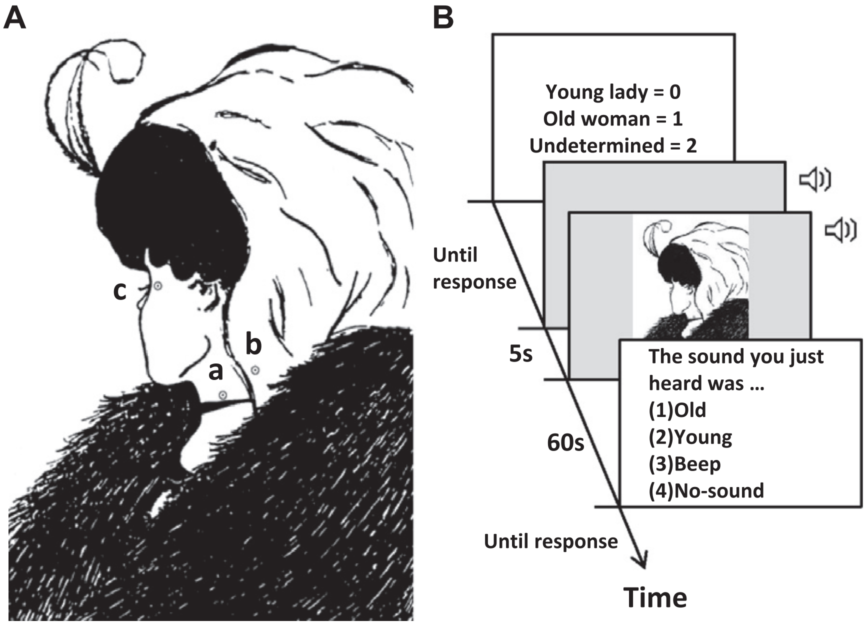

Fig. 1. Stimuli and procedure utilized in the present study. (A) The participants had to fixate the (a) unbiased fixation point in Experiments 1–4. The other two fixation points, (b) fixation favoring old woman percept and (c) fixation favoring young lady percept, were used in Experiment 3. The fixation positions were based on the results of Gale and Findlay (1983). All fixations were presented in red. (B) The instructions concerning the task were presented on the screen until the participant responded. After a 5-s blank period, the figure was displayed for 1 min. A multiple-choice question regarding the soundtrack was shown after the figure disappeared. There were four choices (old, young, beep, or nosound) in Experiment 1, three choices (human voice, wave, or nosound) in Experiment 2, and two choices (old or young) in Experiments 3 and 4.

The hypothesis underlying the present study was that if participants combine the auditory and visual information when perceiving the bistable figure, they should report their initial percept to be the one that was congruent with the soundtrack that they happened to listen to more frequently. In addition, the semantically-congruent percept (with the soundtrack) should be dominant for a larger proportion of time than the other during the test period.

2.1. Methods

2.1.1. Participants

Fourteen undergraduates from the National Taiwan University (NTU) took part in this study in exchange for course credit. All of the participants were naïve as to the purpose of the experiment and all had normal or corrected-to-normal vision and normal hearing. Critically, none of the participants had experience in learning French previously by self-report. The protocol was approved by the academic and ethical committee in the Department of Psychology, NTU.

2.1.2. Apparatus and stimuli

The experiment was conducted in a dimly-lit experimental chamber. The "my wife or my mother-in-law" figure (see Fig. 1A) was presented on a 20-in color monitor controlled by a personal computer with a refresh rate of 85 Hz. The figure was black-white on a gray background (RGB = [127, 127, 127]), extending 19° × 24° at a viewing distance of 60 cm. A small red bull's-eye (0.6° × 0.6°) was presented near the mouth of the old woman and served as the fixation point (i.e., point (a) in Fig. 1A). This fixation point was chosen because this is presumably the position at which the observers should perceive either the old woman or young lady with a similar probability (see Gale & Findlay, 1983).

The auditory monologues of the old woman and young lady were taken from French movies. The soundtracks were rated on a 7-point scale by a group of 12 participants on the extent of the congruency between the voice and the corresponding visual percept of the bistable figure. A score of seven represents "highly congruent", while one represents "highly incongruent". The rated degree of congruency was 6.3 for both the voice of old woman and young lady (SD = 0.94 and 0.75, respectively). A series of 300 Hz pure tone beeps was presented in an intermittent manner, designed to mimic the tempo of the speech soundtracks. The tempo of the first half of the beep soundtrack was identical to the rhythm of the speech of the old woman, and the second half period was identical to the rhythm of the speech of the young lady.[1] The auditory stimuli were presented over closed-ear headphones. All of the soundtracks were presented at approximately 52 dB SPL.

[1] A further ANOVA revealed that the different tempo of the beep in the first and second half of the soundtrack did not influence the proportion of predominance perceptual measure for the bistable figure ($F(1,10) = 0.001$, $MSE = 0.01$, $p = 1.00$, $\eta_p^2 < 0.01$).

2.1.3. Design

The factor of sound (old woman, young lady, beeps, or no-sound) was manipulated on a trial-by-trial basis. The experimental session comprised of three blocks. In each block, there were eight trials, two for each sound condition. The order of presentation of these eight trials in each block was randomized. Six no-sound practice trials were conducted before the experimental session in order to familiarize the participants with the experimental procedure.

At the beginning of each trial, the instruction was presented on the screen to inform the participants that they should press the "0" key whenever they perceived the young lady, the "1" key whenever they perceived the old woman, and the "2" key whenever their percept was intermixed using the numerical keypad situated in front of them. The first percept was defined by the key, "0" or "1", that the participants pressed first. Note that the participants were unaware of this measure. The duration of each percept was calculated by taking the duration between the two key presses.

2.1.4. Procedure

The participants initiated a trial by pressing the space bar. Subsequently, there was a 5-s blank at the start of the trial, followed by the presentation of the bistable figure for 1 min. The soundtrack was presented from the onset of the blank screen until the offset of the visual figure (i.e., for a duration of 65 s in total, see Fig. 1B). The participant were instructed to monitor the bistable figure continuously during the 1-min test period and to report their perceptual state by pressing one of the three corresponding keys. They were also instructed to fixate the red bull's-eye and to establish a constant criterion for reporting their dominant percept of the bistable figure. At the end of each trial, a multiple-choice question was presented on the screen: *"The sound you just heard was (1) old woman; (2) young lady; (3) beeps; or (4) no-sound?"* The participants responded by pressing one of four corresponding number keys (1, 2, 3, or 4) on the keyboard. The instruction of the next trial was then presented on the screen waiting for the participants to initiate.

2.2. Results and discussion

The data from three participants were removed: One participant did not experience bistability during the 1-min test period in the no-sound condition, and the other two did not keep reporting their percept until the end of the test period for more than 10% of total trials. Hence, there were data from 11 participants left for further analysis.

Two performance indices were used. The first performance index consisted of the number of trials in which the *first percept* was the old woman out of six trials in each sound condition. This revealed the outcome of the initial competition between the two percepts (see Balcetis & Dale, 2007; Chastain & Burnham, 1975; Daelli et al., 2010). In addition, we provide the *proportion of predominance* of the old woman percept during the test period as an index of the on-going competition between the two percepts. This index was calculated by dividing the total duration of the old woman percept by the total duration of both the old woman and the young lady percepts within the test period (thus the proportion of predominance of the old woman percept and that of young lady percept are *reciprocally-related*). Separate one-way analyses of variance (ANOVAs) were conducted on each index.

The results of the first percept data (see Fig. 2A) revealed a significant main effect of sound ($F(3,30) = 5.17$, $MSE = 2.13$, $p < .01$, $\eta_p^2 = 0.34$). Critically, a *post hoc* Tukey's HSD test revealed that the participants perceived the old woman as the first percept for more of the trials when the figure was presented with the voice of the old woman than when it was presented with the voice of the young lady ($p < .01$) or with the beep sound ($p < .05$). Similarly, the results of the proportion of predominance data (see Fig. 2B) revealed a significant main effect of sound ($F(3,30) = 4.33$, $MSE = 0.02$, $p < .01$, $\eta_p^2 = 0.30$). A *post hoc* Tukey's HSD test revealed that participants perceived the old woman percept for more of the time when they heard the voice of the old woman than when they heard the voice of the young lady ($p < .01$).

We further calculated the proportion of unambiguous percepts by dividing the total duration of both the old woman and the young lady percepts by the 1-min test period; that is, the period of time in which a mixed percept was reported was excluded (see Table 1). The ANOVA revealed a significant main effect of sound ($F(3,30) = 3.12$, $MSE = 0.004$, $p < .05$, $\eta_p^2 = 0.24$). The *post hoc* Tukey's HSD test indicated that participants were more likely to perceive an unambiguous percept (i.e., either the old woman or young lady percept) when they listened to the voice of old woman than when they listened to the beep sound ($p < .05$). The mean accuracy of the forced-choice sound identification task (see Table 2) was not significantly different in the four sound conditions ($F(3,30) = 0.38$, $MSE = 0.01$, $p = .77$, $\eta_p^2 = 0.04$).

The results of Experiment 1 therefore demonstrate that the participants perceived the old woman as their first percept for more of the trials when they were listening to the voice of the old woman than when they were listening to the voice of the young lady. In terms of the proportion of predominance index, the results demonstrated that the participants reported seeing the old woman percept for more of the time when they were listening to the voice of old woman than when they were listening to the voice of the young lady. Besides, for both indices, the values were intermediate between those of the old woman and young lady conditions when the participants were tested in silence (i.e., in the no-sound condition), or else when they listened to a series of beeps. These results highlight a significant novel crossmodal modulation on the perception of a bistable figure, which can be attributed to the semantic context provided by the auditory soundtrack that the observers were concurrently listening to. Analysis of the data concerning the proportion of unambiguous percepts indicated that the participants were less likely to perceive a mixed percept while listening to the voice of the old woman as compared to when they were listening to meaningless beeps instead. That is, the presentation of the auditory semantic context also reduced the visual ambiguity when viewing the bistable figure. Nevertheless, we need to exclude a possible alternative explanation for the

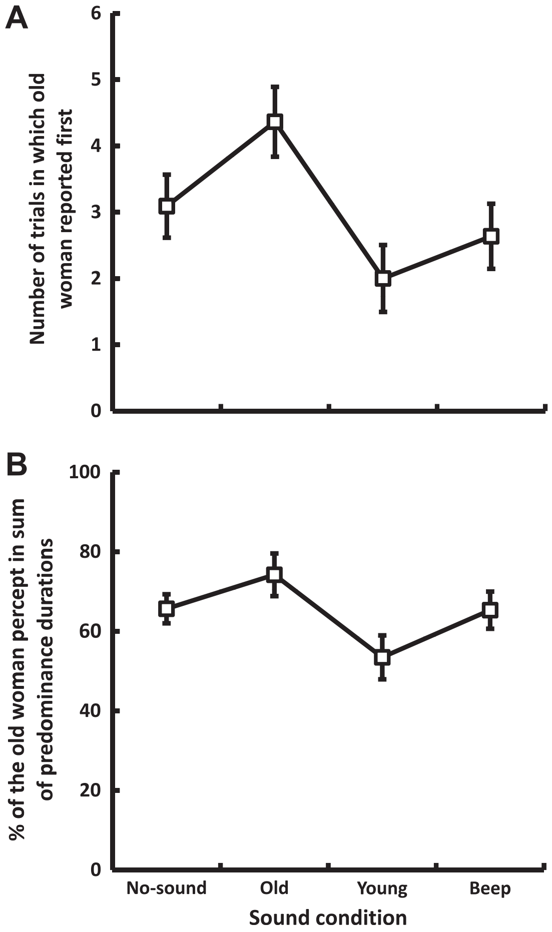

Fig. 2. Results of (A) the number of trials that participants reported the old woman as the first percept (out of 6 trials), (B) the predominance duration of the old woman percept (proportion to sum of predominance durations of old woman and young lady percept) in Experiment 1. Error bars represent ± 1 standard error of the mean.

Table 1
Mean percentages of the unambiguous percept, and their standard errors (in parentheses) during the test period in Experiments 1, 3 and 4.

Experiment 1: Sound condition			
No-sound	Old	Young	Beep
82.01 (3.62)	86.94 (2.92)	83.03 (3.71)	78.28 (5.67)
Experiment 3: Fixation position			
Sound	Old	Unbiased	Young
Old	90.56 (3.51)	88.36 (3.65)	92.00 (4.75)
Young	87.00 (4.32)	89.10 (3.23)	93.37 (3.64)
Experiment 4: Selective attention			
Sound	Old	Passive	Young
Old	96.34 (1.37)	97.21 (1.00)	95.85 (1.38)
Young	97.08 (1.00)	96.28 (1.34)	96.87 (1.08)

present effects in terms of response bias. That is, the participants might simply have reported the percept which was congruent with the soundtrack that they were listening to for more of the time. We therefore designed Experiment 2 in order to rule out this possibility.

3. Experiment 2

In Experiment 2, the participants were instructed to report the predominance of only one of the competing percepts. The soundtrack presented in a trial in this experiment was *never* congruent with the visual target that they should report.

Table 2
Mean accuracy (%) of reporting the identities of the soundtracks, and their standard errors (in parentheses) during the test period through Experiments 1–4.

Experiment 1: Sound condition			
No-sound	Old	Young	Beep
96.97 (2.03)	98.48 (1.52)	95.45 (4.55)	93.94 (3.39)
Experiment 2: Control for response bias			
Visual target	Incongruent	Irrelevant	No-sound
Old	100.00 (0.00)	100.00 (0.00)	98.96 (1.04)
Young	100.00 (0.00)	100.00 (0.00)	100.00 (0.00)
Experiment 3: Fixation position			
Sound	Old	Unbiased	Young
Old	100.00 (0.00)	100.00 (0.00)	98.15 (1.75)
Young	100.00 (0.00)	98.15 (1.75)	100.00 (0.00)
Experiment 4: Selective attention			
Sound	Old	Passive	Young
Old	95.83 (2.33)	100.00 (0.00)	92.71 (2.93)
Young	98.96 (1.01)	94.79 (1.93)	100.00 (0.00)

Specifically, the participants were randomly assigned to one of two groups: In Group 1, they were instructed to report the predominance of the old woman percept; in contrast, in Group 2, the participants were instructed to report the predominance of the young lady. In both groups, they were listening to a soundtrack which was incongruent to their visual target (the voice of the young lady for Group 1 and the voice of the old woman for Group 2), a soundtrack that was irrelevant to the bistable figure (ocean waves), or else without sound. Note that the participants in either group had never heard the soundtrack that was congruent with the percept of their pre-designated target. The task was designed to minimize the possibility that participants might strategically report the visual percept which was congruent with the auditory stimulus that they were concurrently hearing during the test period.

3.1. Methods

3.1.1. Participants

Thirty-five undergraduate students at NTU took part in Experiment 2. None of the participants had taken part in Experiment 1, and all were naïve as to the purpose of the study. They were randomly assigned to either Group 1 ($N = 17$) or Group 2 ($N = 18$).

3.1.2. Stimuli, design and procedure

Two factors, the sound (incongruent, irrelevant, or no-sound) and the visual target (old woman or young lady), were manipulated. The visual target factor was manipulated between-participants. In the incongruent condition, the old woman and young lady soundtracks were identical to those used in Experiment 1. The irrelevant soundtrack consisted of the sound of waves crashing on the beach. This ocean wave sound was downloaded from www.soundsnap.com on 06/11/2008. Each group of participants completed three blocks of six trials. Within a block, there were two trials for each sound condition.

The procedure was the same as in Experiment 1, with the exception of the instructions given to participants. At the beginning of each trial, the instruction for the participants in Group 1 was to press the “z” key whenever they perceived the *old woman* as dominant, and the “/” key whenever they *did not*. The participants in Group 2 were instructed to respond to the *young lady* percept instead, by pressing the “z” and “/” keys to indicate its predominance. As soon as the figure was presented, the participants started to monitor and report the predominance of their visual target by pressing the two corresponding keys. At the end of each trial, a multiple-choice question was presented on the screen: *“The sound you just heard was (1) human voice; (2) waves; or (3) no-sound?”* The participants had to press one of three corresponding number keys in order to respond. Due to the fact that the participants only responded to one of the possible percepts (i.e., their visual target), it is hard to know whether the target percept is the initial one that they perceived. Hence, only the proportion of predominance index was used in Experiment 2.

3.2. Results and discussion

One participant in Group 1 and four participants in Group 2 were excluded from the data analysis because they did not respond until the end of the test period on more than 10% of the trials. Three more participants in Group 2 pressed only one key when responding, which prevented us from calculating the predominance duration of the target percept during the test period, so their data were also excluded. The data from 16 participants in Group 1 and 11 in Group 2 remained for further

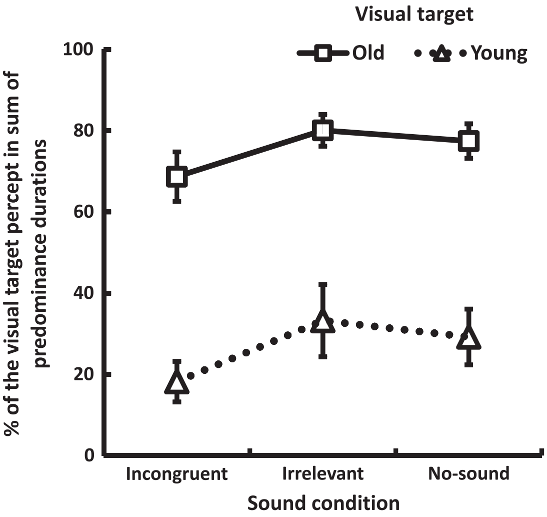

Fig. 3. Results of the predominance duration of the visual target percept (proportion to the 1-min test period) in Experiment 2. Error bars represent ±1 standard error of the mean.

analysis. The proportion of predominance of the target percept was calculated by dividing the total duration of each target percept by the 1-min test period. The fact that the participants were instructed simply to respond to the visual target also made the calculation of the percentage of unambiguous percepts impossible.

The proportion of predominance data (see Fig. 3) were submitted to a two-way ANOVA. The results of this analysis revealed a significant main effect of sound ($F(2,50) = 7.67$, $MSE = 0.02$, $p < .001$, $\eta_p^2 = 0.23$). The *post hoc Bonferroni t*-test revealed that participants perceived the visual target for less of the time when they heard the incongruent soundtrack as compared to when they heard the irrelevant soundtrack or else when no sound was presented ($ps < .05$). The main effect of the visual target was also significant ($F(1,25) = 40.65$, $MSE = 0.11$, $p < .001$, $\eta_p^2 = 0.62$). Nevertheless, the interaction between the sound and target percept was not significant ($F(2,50) = 0.14$, $MSE = 0.02$, $p = .87$, $\eta_p^2 = 0.005$).[2] The mean accuracy of performance on the forced-choice sound identification task (Table 2) was submitted to a two-way ANOVA. None of the main effects, nor their interaction, were significant ($Fs < 0.68$, $ps > .42$, $\eta_p^2 s =< 0.03$).

In Experiment 2, the participants' task was modified in order to prevent them from strategically reporting the percept that happened to be congruent with the sound as being dominant. The participants reported that they perceived the old woman percept for less of the time when they heard the voice of a young lady, and perceived the young lady percept for less of the time when they heard the voice of an old woman, as compared to when they heard an irrelevant sound (the ocean waves here) in terms of the proportion of predominance measure. In other words, the crossmodal modulation in this experiment should be attributed to the fact that the presentation of a semantically-incongruent soundtrack shortened the predominance duration of the visual target. Hence, the robustness of the crossmodal modulation by sound observed in this experiment suggests that it does not require a direct congruency between the sound and the visual target in order for the crossmodal semantic effect to occur. In sum, these results suggest that the crossmodal modulation of the conscious perception of the bistable figure likely results from the perceptual consequences of listening to the soundtrack, rather than from some forms of response bias attributable to participants' strategically reporting the visual percept that happened to be congruent with the sound.

The results of Experiment 2 also revealed that the proportion of predominance duration reported by the participants whose target was the old woman was significantly larger than those whose target was the young lady. This result indicates that there is an intrinsic bias toward the old woman percept when viewing the "my wife or my mother-in-law" bistable figure. This bias may be attributable to the fact that the face of the old woman extends over a larger area and exhibits more front angle than the face of young lady. Hence, even when we tried to use an unbiased fixation position, this intrinsic bias was unavoidable. Note that this bias did not interact with the modulation of crossmodal semantic congruency in this experiment. The results of Experiment 1 were influenced by this intrinsic bias as well: The participants' predominance duration of the old woman percept reached 66% of the time when no sound was presented, which was significantly above chance levels ($t10 = 4.11$, $p < .05$; see Fig. 2B). In fact, one of the percepts in a bistable figure is often more dominant than the other, such as

[2] Even though the participants were instructed to respond when they perceived the "target" versus "non-target" percept, one may worry that, after a few trials, they may have changed the task into reporting "old woman" versus "young lady" percept. We therefore analyzed the participants' proportion of predominance data of the first trial only. The results revealed exactly the same pattern: The main effect of sound ($F(2,50) = 4.84$, $MSE = 0.05$, $p < .01$, $\eta_p^2 = 0.16$), as well as the visual target ($F(1,25) = 39.67$, $MSE = 0.12$, $p < .001$, $\eta_p^2 = 0.61$), was significant. Nevertheless, their interaction was not ($F(2,50) = 0.21$, $MSE = 0.05$, $p = .81$, $\eta_p^2 = 0.008$). The *post hoc Bonferroni* test on the sound factor manifested that it was less likely for participants to perceive the visual target when they were concurrently listening to a soundtrack that was incongruent with the visual target, as compared to an irrelevant soundtrack ($p < .05$).

the top view in the Necker cube (Meng & Tong, 2004), even though there is lacking of any obvious explanation for this intrinsic tendency.

4. Experiment 3

When participants fixate at a particular position either favoring a particular percept, they are more likely to perceive that percept initially and for more of the time thereafter (Gale & Findlay, 1983; Meng & Tong, 2004). If the crossmodal semantic effect on the perception of a bistable figure were to have resulted from participants' focusing on certain critical features that constitute the semantically-congruent percept with the sound, the crossmodal effect should have been reduced or eliminated when the participants' fixation was now manipulated as favoring for a particular percept. Experiment 3 was designed to test whether the crossmodal semantic effect is robust when the participants' fixation position is manipulated.

4.1. Methods

4.1.1. Participants

A new group of nine undergraduates at NTU was tested. None of them had participated in the former experiments.

4.1.2. Stimuli, design and procedure

Two factors, the sound (old and young) and fixation position (fixation for old, unbiased fixation, or fixation for young) were manipulated. The soundtrack of either the old woman or the young lady voice (those used in Experiment 1) was presented. One of three possible fixation points was presented on the bistable figure in each trial: Point (a) in Fig. 1A, which has been used in our previous experiments, was again used in the unbiased fixation condition. The fixation point favoring the old woman percept was the point on the head cloak (see point (b) in Fig. 1A), and the fixation point favoring the young lady percept was the point near the eye of the percept of young lady (see point (c) in Fig. 1A; Gale & Findlay, 1983).

During the test period, the participants' fixation position was monitored by the EyeLink 2000 (SR Research, Mississauga, Ontario, Canada) with a sampling rate of 1000 Hz. Once the participants' fixation diverted from the center of bull's eye for more than 1.5°, the trial terminated. In addition, an instruction to remind the participants that they had to fixate at the predesignated position during the test period was presented. Only the data from the completed trials went on for further analysis. There were 25.2% (SD = 14.5%) trials removed because of the failure to maintain fixation. Due to the fact that it is hard for the participants to strictly fixate at a particular location for a long period, a trial in this experiment was 30 s. The fixation appeared on the screen concurrently with the onset of soundtrack (which was 5 s earlier than the onset of the bistable figure) to help the participants fixated steadily as soon as the test period began. At the end of each trial, the following multiple-choice question was presented on the screen: "The voice you just heard was (1) old woman or (2) young lady?" The participants pressed one of the two corresponding number keys in order to respond. Each sound condition was paired with each of three fixation positions, giving rise to a 2 × 3 factorial design. There were six blocks containing six trials, one trial for each condition. The other details were exactly as in Experiment 1.

The participants completed six no-sound practice trials, which were composed of two trials for each fixation position. During practice, the experimenter instructed the participants to fixate on the different fixation positions across trials as soon as possible once the fixation was presented on the screen.

4.2. Results and discussion

Separate two-way ANOVAs were conducted on the first percept and proportion of predominance measure, respectively. The results of the first percept data (see Fig. 4A) revealed a significant main effect of sound ($F(1,8) = 6.25$, $MSE = 0.67$, $p < .05$, $\eta_p^2 = 0.44$), indicating that the participants perceived the old woman as the first percept for more of the trials when they heard the voice of the old woman than when they heard the voice of the young lady. The main effect of fixation position was also significant ($F(2,16) = 24.59$, $MSE = 3.54$, $p < .001$, $\eta_p^2 = 0.75$). A *post hoc* Tukey's HSD test revealed that the participants perceived the old woman as the first percept more often in both the fixation for old and the unbiased fixation conditions than in the fixation for young condition ($ps < .01$). The interaction between sound and fixation position was not significant ($F(2,16) = 1.35$, $MSE = 0.54$, $p = .29$, $\eta_p^2 = 0.14$).

Analysis of the proportion of predominance data (see Fig. 4B) revealed a significant main effect of sound: The participants perceived the old woman for more of the time when they listened to the voice of old woman than when they listened to the voice of the young lady ($F(1,8) = 8.22$, $MSE = 0.01$, $p < .05$, $\eta_p^2 = 0.50$). The main effect of fixation position was also significant ($F(2,16) = 28.56$, $MSE = 0.07$, $p < .001$, $\eta_p^2 = 0.78$). A *post hoc* Tukey's HSD test revealed that participants perceived the percept of the old woman for more of the time in both the fixation for old and the unbiased fixation conditions than in the fixation for young condition ($ps < .01$). The interaction between the effect of sound and fixation position was not significant ($F(2,16) = 0.19$, $MSE = 0.01$, $p = .83$, $\eta_p^2 = 0.02$). The mean percentage of time for which the participants reported an unambiguous percept was 90% (see Table 1), which was not significantly different in any of the sound × fixation position conditions (all $Fs < 1.32$, $ps > .30$, $\eta_p^2 s =< 0.14$). The mean accuracy of the forced-choice sound identification task (see Table 2) was submitted to a two-way ANOVA. None of the main effect, nor their interaction, was significant ($Fs < 1.00$, $ps > .39$, $\eta_p^2 s =< 0.11$).

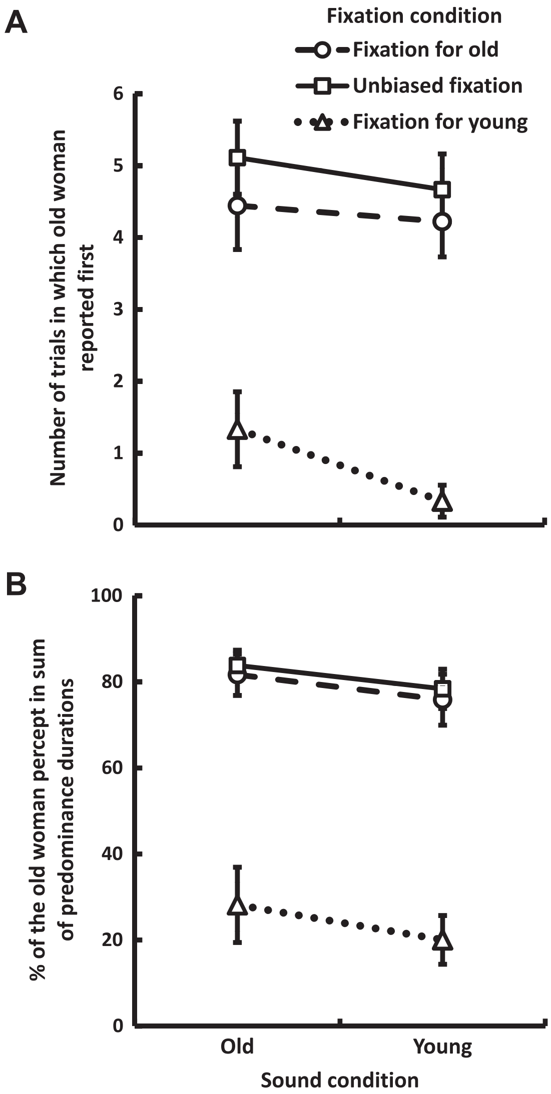

Fig. 4. Results of (A) the number of trials in which participants reported the old woman as the first percept (out of 6 trials) (B) the predominance duration of the old woman percept (proportion to sum of predominance durations of old woman and young lady percept) in Experiment 3. Error bars represent ±1 standard error of the mean.

In Experiment 3, the modulation of crossmodal semantic congruency was robustly observed when the participants were focusing at different locations in terms of both the first percept and the proportion of predominance measures. That is, the participants' perception of a bistable figure was modulated by the sound that they were hearing, regardless of their fixation position either favoring the percept congruent with the sound or not. In conclusion, the robust crossmodal semantic modulatory effect in both indices suggests that the crossmodal modulation by the auditory semantic context could be dissociated from that elicited by a bottom-up factor, namely, fixation position, when viewing a bistable figure.

5. Experiment 4

Meng and Tong (2004) and van Ee et al. (2005) have both demonstrated that participants' selective attention to a particular percept in a bistable figure can enhance the dominance duration of that percept. If the crossmodal semantic effect on the perception of a bistable figure had resulted from participants' selectively attending to the percept that happened to be congruent with the sound, the crossmodal effect should be reduced or eliminated when the participants' attention was now devoted to a particular target. Experiment 4 was therefore designed to investigate whether the crossmodal semantic modulation of bistable figure perception could still be observed when it was manipulated simultaneously with selective attention over the bistable figure.

5.1. Methods

5.1.1. Participants

A new group of 16 undergraduates in NTU completed this experiment. None of them had taken part in the former experiments.

5.1.2. Stimuli, design and procedure

Two factors, sound (old and young) and selective attention (maintain old, passive, or maintain young), were manipulated. In each trial, the soundtrack of either the old woman or the young lady was presented. One of three selective attention instructions (maintain the old woman percept, maintain the young lady percept, or watch the figure passively) was randomly assigned to each trial. There were six blocks, and each block consisted of six trials, one trial for each sound × selective attention condition. The order of presentation of the six trials in each block was randomized.

The 5-s blank (see Fig. 1B) was now replaced by an attention instruction presented on the screen, which informed the participants that they should try to actively maintain the old woman percept or the young lady percept, or else no instruction was presented (presumably resulting in participants viewing the figure passively). Meanwhile, the participants had to focus on the fixation point (a) in Fig. 1A during the course of the 30-s test period. The participants' fixation position was also monitored with an EyeLink 2000 eyetracker as in Experiment 3. There were 20.4% (SD = 8.6%) trials removed because of the failure to maintain fixation. The soundtrack was presented from the onset of the attention instruction until the offset of the bistable figure. At the end of each trial, the following multiple-choice question was presented on the screen: "The sound you just heard was (1) old woman or (2) young lady?" The other details were the same as in Experiment 3. The participants pressed one of the two corresponding number keys in order to respond.

The participants completed six no-sound practice trials, two for each selective attention condition, in order to familiarize themselves with the task before the experiment. The experimenter ensured that the participants could successfully focus their attention on the particular percept during the test period in line with the attention instruction, or else view the figure passively when no specific attention instruction was provided.

5.2. Results and discussion

Separate two-way ANOVAs were conducted on the first percept and proportion of predominance measure, respectively. The analysis of the first percept (see Fig. 5A) revealed a significant main effect of the sound ($F(1,15) = 14.88$, $MSE = 2.04$, $p < .01$, $\eta_p^2 = 0.50$): The participants perceived the old woman as the first percept in more of the trials when they heard the old woman's voice than when they heard the young lady's voice. The main effect of selective attention was also significant ($F(2,30) = 18.65$, $MSE = 1.19$, $p < .001$, $\eta_p^2 = 0.55$). A *post hoc* Tukey's HSD test on the main effect of selective attention revealed that the participants perceived the old woman as the first percept for more of the trials when they tried to maintain the percept of old woman and when they passively watched the bistable figure, as compared to when they tried to maintain the percept of young lady ($ps < .01$). However, the interaction between sound and selective attention was not significant ($F(2,30) = 1.95$, $MSE = 0.92$, $p = .16$, $\eta_p^2 = 0.11$).

The analysis of the proportion of predominance data (see Fig. 5B) revealed that both the main effects of sound ($F(1,15) = 78.95$, $MSE = 0.01$, $p < .001$, $\eta_p^2 = 0.84$) and selective attention ($F(2,30) = 37.89$, $p < .001$, $MSE = 0.04$, $\eta_p^2 = 0.72$) were significant. A *post hoc* Tukey's HSD test revealed that the values in the three selective attention conditions were different from each other (all $ps < .01$). The interaction between these two factors was also significant ($F(2,30) = 14.37$, $MSE = 0.02$, $p < .001$, $\eta_p^2 = 0.49$). The analyses of the simple main effects of sound were both significant in all three attentional conditions ($F(1,45) = 5.37$, $MSE = 0.01$, $p < .05$, $\eta_p^2 = 0.11$ for the maintain old condition, $F(1,45) = 88.81$, $MSE = 0.01$, $p < .001$, $\eta_p^2 = 0.66$ for the passive condition, and $F(1,45) = 12.20$, $MSE = 0.01$, $p < .001$, $\eta_p^2 = 0.21$ for the maintain young condition). The magnitude of the crossmodal semantic effect (i.e., the difference between the old and young sound conditions) was significantly larger in the passive condition than in the maintain young condition ($t15 = 3.59$, $p < .05$), and the former was also larger than in the maintain old condition ($t15 = 4.47$, $p < .05$).

The mean percentage of unambiguous percepts was 97% (see Table 1), which was not significantly different between any pair of sound × selective attention conditions (all $Fs < 1.69$, $ps > .22$, $\eta_p^2 s = < 0.11$). The mean accuracy of the forced-choice sound identification task (Table 2) was submitted to a two-way ANOVA. The two main effects of selective attention and sound were not significant ($Fs < 1.21$, $ps > .29$, $\eta_p^2 s = < 0.07$). However, their interaction was significant ($F(2,30) = 6.18$, $MSE = 0.01$, $p < .01$, $\eta_p^2 = 0.29$). The simple main effect of selective attention was significant when hearing the old woman sound ($F(2,60) = 4.24$, $MSE = 0.01$, $p < .05$, $\eta_p^2 = 0.12$), but not when hearing the young lady sound ($F(2,60) = 2.41$, $MSE = 0.01$, $p = .10$, $\eta_p^2 = 0.07$). A *post hoc* Tukey's HSD test revealed that, when hearing the old woman sound, the accuracy was lower in the maintain young condition than in the passive condition ($p < .05$). That is, it is sometimes hard for the participants to report the sound that was incongruent with their maintaining visual target.

In Experiment 4, the sound that the participants heard (either the voice of the old woman or of the young lady) was manipulated as was the participants' selective attention over the bistable figure. Still, the participants were more likely to perceive the old woman percept as the first percept when they heard the voice of the old woman than when they heard

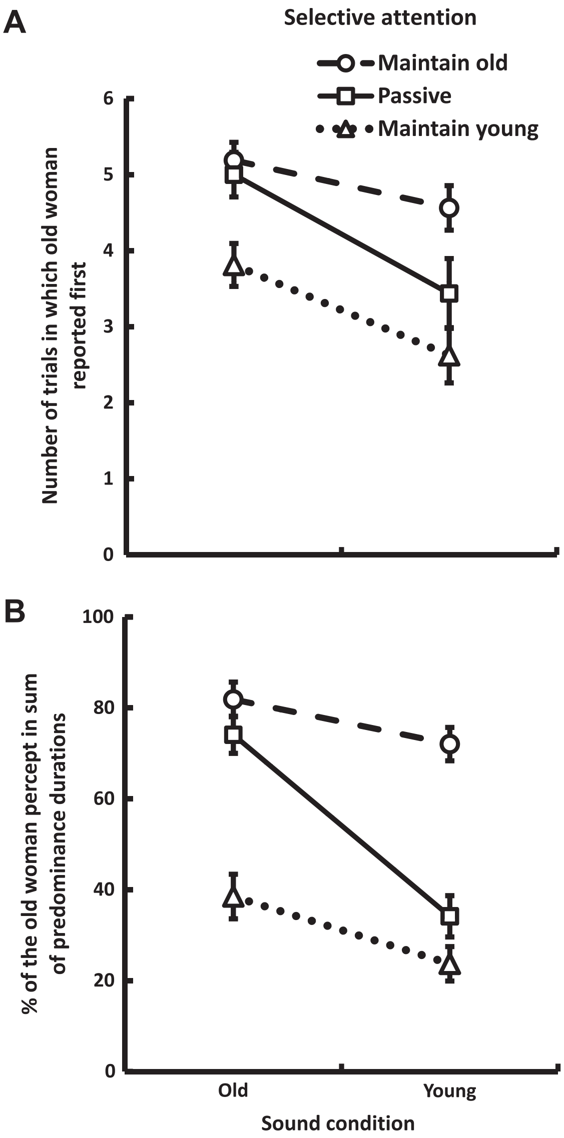

Fig. 5. Results of (A) the number of trials that participants reported the old woman as the first percept (out of 6 trials) (B) the predominance duration of the old woman percept (proportion to sum of predominance durations of old woman and young lady percept) in Experiment 4. Error bars represent ±1 standard error of the mean.

the voice of the young lady, regardless of whether they actively tried to maintain a particular percept or else viewed the figure passively (i.e., no interaction was observed between sound and selective attention). That is, in terms of the first percept measure, the modulation by audiovisual semantic congruency and selective attention on the initial percept of the bistable figure could be dissociated. We therefore suggest that audiovisual semantic congruency is, in-and-of-itself, sufficient to modulate the first percept of a bistable figure, rather than necessarily being mediated by the guiding of participants' attention (cf. Balcetis & Dale, 2007).

On the other hand, the crossmodal semantic effect of sound on the proportion of predominance measure for the bistable figure was reduced when participants tried to actively maintain a particular percept as revealed by the significant interaction between the soundtrack and selective attention factors. This result implies that the modulation by selective attention is more dominant than that by auditory semantic context in the on-going competition between two percepts during the continuous viewing. Note that the crossmodal modulation by the soundtrack was still significant when the participants tried to either maintain the percept of the old woman or the young lady. That said, the crossmodal semantic effect cannot be completely explained by the possibility that the sound serves as an explicit cue as attentional instruction that was designed to bias their perception of the bistable figure.

6. General discussion

In the present study, we investigated whether the presentation of auditory semantic information regarding an old woman or a young lady would influence participants' conscious perception of the "my wife or my mother-in law" bistable figure. We measured the participants' first percept and the predominance duration of each percept during the test period. The results of Experiment 1 demonstrated that the semantic context conveyed by the soundtrack modulated both measures. In Experiment 2, the possibility that the crossmodal semantic effect on the proportion of predominance measure simply resulted from some sort of response bias was minimized by presenting a soundtrack that was never congruent with the visual target, and the crossmodal semantic congruency effect was robustly observed. In Experiment 3, we further demonstrated that the modulation of the crossmodal semantic context and the fixation factors additively influenced both the first percept and the proportion of predominance measures. In Experiment 4, the results revealed that the modulation of the crossmodal semantic context and the attention factors additively influenced the participants' first percept. Nevertheless, the effect of crossmodal semantic modulation was reduced, but still significant, when the participants were trying to actively maintain either the old woman or young lady percept as compared to the passive condition in the proportion of predominance measure.

To our knowledge, the present study is the first to use crossmodal stimulation in order to provide unequivocal evidence concerning the existence of a semantic congruency effect on the conscious perception of a visually-presented bistable figure. The results further demonstrate that the crossmodal semantic modulation on the outcome of the initial competition between the two possible percepts (i.e., the first percept measure) was independent of both fixation and attention factors (cf. Balcetis & Dale, 2007). However, the crossmodal semantic modulation of the on-going competition (in terms of the proportion of predominance measure) was significantly reduced when the participants voluntarily and covertly (i.e., without moving their eyes) attended to a specific percept. This result implies that attention, a volitional effect of maintaining one particular view, was more dominant than the modulatory effect of crossmodal semantic congruency in terms of determining the predominance of the attended/semantically-congruent percept when continuously viewing the bistable figure. The discrepancy between the first percept and the proportion of predominance measures may be attributed to the different underlying mechanisms: The first percept is the result of the initial competition, which can be accounted for by low-level neural mechanisms that are less susceptible to the high-level modulation such as attention (see Noest, van Ee, Nijs, & van Wezel, 2007). Predominance duration is the results of on-going competition instead, which is quite susceptible to the modulation of attention (Meng & Tong, 2004; van Ee et al., 2005).

In Gale and Findlay's (1983) account of the percept that emerges when viewing a bistable figure, both fixation and attention play a role in modulating the conscious percept. According to their idea, each small group of features can be considered an element belonging to different figures; for example, the center element in the "my wife or my mother-in-law" can be considered either as the "eye" of the old woman percept or as the "ear" of the young lady percept. The likelihood that each element is represented in either way is assigned a weighting, and the weighting of each element is collectively associated rather than isolated in terms of determining the final conscious perception of the bistable figure. For example, observers can modulate the weighting by fixating at the position in favor of that percept, or by trying to maintain a particular percept without moving their eyes. These two mechanisms are categorized as bottom-up and top-down modulations on the perception of bistable figures, respectively (see Long & Toppino, 2004; Meng & Tong, 2004).

We thereby suggest that the crossmodal semantic information may serve as a factor modulating the weighting of elements in a bistable figure as well, which is presumably based on the learned associations from an observer's daily life (e.g., Smith et al., 2007). That is, such crossmodal semantic congruency serves as a form of heuristics, which can be understood in terms of the prior term of the causal inference model in crossmodal perception (see Shams & Beierholm, 2010). For example, when the observer hears an old woman's voice, the weighting of elements may be biased toward the old woman percept. Note also that the crossmodal semantic congruency factor interacted with attention rather than with visual fixation in the on-going competition in terms of the proportion of predominance measure. That said, crossmodal semantic congruency and attention may both modulate the perception in the bistable figure in a top-down fashion, distinct from the effect of fixation working in a bottom-up fashion. This hypothesis can be accounted for by the recent model consisting of multiple levels of interplay between attention and multisensory integration as proposed by Talsma, Senkowski, Soto-Faraco, and Woldorff (2010). According to their model, both crossmodal interaction in terms of semantic congruency and attention should rely on a form of feedback gain adjustment. The interaction of crossmodal semantic information and attention may therefore be weak in the first round of visual processing that gives rise to the outcome of first percept, while their interaction then keeps accumulating to modulate the on-going competition during the continuous viewing period.

Chen, Yeh, and Spence (2011) recently demonstrated that the modulation of people's dominant perception by crossmodal semantic congruency in the binocular rivalry situation can be dissociated from that elicited by attention. It should be noted that, given that the visual stimuli in the case of binocular rivalry consist of two distinct figures presented to each eye, the competition likely occurs between two intraocular pathways or two visual pattern representations (Leopold & Logothetis, 1996; Tong, Meng, & Blake, 2006). By contrast, in the case of bistable figure perception, the visual stimulus cannot be segregated as two distinct visual patterns in early visual processing. That said, the level at which the visual competition occurs should therefore be higher for the case of bistable figure perception than in the binocular rivalry situation. In a similar vein, the level of visual competition which was modulated by the crossmodal semantic congruency should be higher in the bistable figure than binocular rivalry. For example, the modulation of crossmodal semantic congruency can be considered as a form of mid-level crossmodal excitatory mechanism and/or a top-down cognitive modulation in the binocular rivalry situation (see Chen et al., 2011), while the latter possibility provides a more promising way to modulate the bistable figure perception. This may be the reason for that the modulations of crossmodal semantic congruency and attention on the bistable figure perception are hard to dissociate completely (see Experiment 4). In addition, the discrepancy between the effects of crossmodal semantic modulation on the bistable figure and binocular rivalry situations also implies that multisensory interactions in terms of stimulus semantic congruency may not comprise a unitary mechanism in these various situations.

7. Conclusions

The results of the four experiments reported in the present study demonstrate that the semantic information provided by an auditory soundtrack can influence the perceptual interpretation of a bistable visual figure. This effect is independent of the modulation induced by participants' fixation position, and cannot be entirely explained in terms of participants' voluntary attention. The crossmodal semantic interaction, in turn, helps one of the alternative representations to win the visual competition for access to consciousness when viewing a bistable figure. This novel result provides an important alternative mechanism regarding how visual awareness emerges other than simply relying on an observer's attentional focus (e.g., Koch & Tsuchiya, 2007).

Acknowledgments

This research was supported by a joint project funded by the British Academy to Charles Spence (CQ ROKK0) and Taiwan's National Science Council to Su-Ling Yeh (NSC 97-2911-I-002-038 and 98-2410-H-002-023-MY3).

References

Balcetis, E., & Dale, R. (2007). Conceptual set as a top-down constraint on visual object identification. *Perception, 36*, 581–595.
Boring, E. G. (1930). A new ambiguous figure. *American Journal of Psychology, 42*, 444–445.
Calvert, G. A., Spence, C., & Stein, B. E. (Eds.). (2004). *The handbook of multisensory processing*. Cambridge, MA: MIT Press.
Chastain, G., & Burnham, C. A. (1975). The first glimpse determines the perception of an ambiguous figure. *Perception & Psychophysics, 17*, 221–224.
Chen, Y.-C., & Spence, C. (2010). When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. *Cognition, 114*, 389–404.
Chen, Y.-C., Yeh, S.-L., & Spence, C. (2011). Crossmodal constraints on human perceptual awareness: Auditory semantic modulation of binocular rivalry. *Frontiers in Psychology, 2*, 212. doi:10.3389/fpsyg.2011.00212.
Daelli, V., van Rijsbergen, N. J., & Treves, A. (2010). How recent experience affects the perception of ambiguous objects. *Brain Research, 1322*, 81–91.
Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of Vision, 7, 7*, 1–14.
Feist, M., & Gentner, D. (2007). Spatial language influences memory for spatial scenes. *Memory & Cognition, 35*, 283–296.
Gale, A. G., & Findlay, J. M. (1983). Eye movement patterns in viewing ambiguous figures. In R. Groner (Ed.), *Eye movements and psychological functions: International views* (pp. 145–168). Hillsdale, NJ: LEA.
Garcia-Perez, M. A. (1989). Visual inhomogeneity and eye movements in multistable perception. *Perception & Psychophysics, 46*, 397–400.
Goolkasian, P., & Woodberry, C. (2010). Priming effects with ambiguous figures. *Attention, Perception, & Psychophysics, 72*, 168–178.
Hupé, J. M., Joffo, L. M., & Pressnitzer, D. (2008). Bistability for audiovisual stimuli: Perceptual decision is modality specific. *Journal of Vision, 8, 1*, 1–15.
Iordanescu, L., Grabowecky, M., Franconeri, S., Theeuwes, J., & Suzuki, S. (2010). Characteristic sounds make you look at target objects more quickly. *Attention, Perception, & Psychophysics, 72*, 1736–1741.
Kim, C. Y., & Blake, R. (2005). Psychophysical magic: Rendering the visible 'invisible'. *Trends in Cognitive Sciences, 9*, 381–388.
Koch, C., & Tsuchiya, N. (2007). Attention and consciousness: Two distinct brain processes. *Trends in Cognitive Sciences, 11*, 16–22.
Leopold, D. A., & Logothetis, N. K. (1996). Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature, 379*, 549–553.
Long, G. M., & Toppino, T. C. (2004). Enduring interest in perceptual ambiguity: Alternating views of reversible figures. *Psychological Bulletin, 130*, 748–768.
Meng, M., & Tong, F. (2004). Can attention selectively bias bistable perception? Differences between binocular rivalry and ambiguous figures. *Journal of Vision, 4*, 539–551.
Necker, L. A. (1932). Observations on an optical phenomenon which occurs on viewing a figure of a crystal or geometrical solid. *Philosophical Magazine, 1*, 329–337.
Noest, A. J., van Ee, R., Nijs, M. M., & van Wezel, R. J. A. (2007). Percept-choice sequences driven by interrupted ambiguous stimuli: A low-level neural model. *Journal of Vision, 7, 10*, 1–14.
Schneider, T. R., Engel, A. K., & Debener, S. (2008). Multisensory identification of natural objects in a two-way crossmodal priming paradigm. *Experimental Psychology, 55*, 121–132.
Shams, L., & Beierholm, U. R. (2010). Causal inference in perception. *Trends in Cognitive Sciences, 14*, 425–432.
Smith, E. L., Grabowecky, M., & Suzuki, S. (2007). Auditory-visual crossmodal integration in perception of face gender. *Current Biology, 17*, 1680–1685.
Spence, C. (2007). Audiovisual multisensory integration. *Acoustical Science and Technology, 28*, 61–70.
Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences, 14*, 400–410.
Tong, F., Meng, M., & Blake, R. (2006). Neural bases of binocular rivalry. *Trends in Cognitive Sciences, 10*, 502–511.
van Ee, R., van Dam, L. C. J., & Brouwer, G. J. (2005). Voluntary control and the dynamics of perceptual bi-stability. *Vision Research, 45*, 41–55.
Vicholkovska, A. (1906). Illusions of reversible perspective. *Psychological Review, 13*, 276–290.
Washburn, M. F., & Gillette, A. (1933). Motor factors in voluntary control of cube perspective fluctuations and retinal rivalry fluctuations. *American Journal of Psychology, 45*, 315–319.
Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88*, 638–667.