Heliyon 10 (2024) e33828

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

The temporal dynamics of conscious and unconscious audio-visual semantic integration

Mingjie Gao [a], Weina Zhu [a,*], Jan Drewes [b,**]

[a] School of Information Science, Yunnan University, Kunming, China
[b] Institute of Brain and Psychological Sciences, Sichuan Normal University, Chengdu, China

A B S T R A C T

We compared the time course of cross-modal semantic effects induced by both naturalistic sounds and spoken words on the processing of visual stimuli, whether visible or suppressed form awareness through continuous flash suppression. We found that, under visible conditions, spoken words elicited audio-visual semantic effects over longer time (−1000, −500, −250 ms SOAs) than naturalistic sounds (−500, −250 ms SOAs). Performance was generally better with auditory primes, but more so with congruent stimuli. Spoken words presented in advance (−1000, −500 ms) outperformed naturalistic sounds; the opposite was true for (near-)simultaneous presentations. Congruent spoken words demonstrated superior categorization performance compared to congruent naturalistic sounds. The audio-visual semantic congruency effect still occurred with suppressed visual stimuli, although without significant variations in the temporal patterns between auditory types. These findings indicate that: 1. Semantically congruent auditory input can enhance visual processing performance, even when the visual stimulus is imperceptible to conscious awareness. 2. The temporal dynamics is contingent on the auditory types only when the visual stimulus is visible. 3. Audiovisual semantic integration requires sufficient time for processing auditory information.

1. Introduction

Audiovisual integration refers to the brain combining information from both visual and auditory channels to form a comprehensive perceptual experience. This process is pervasive in daily life, for instance, when watching a movie, we merge visual images with audio sounds, enhancing our understanding and immersion in the content. Compared to unimodal sensory input, such multisensory integration significantly enhances task performance and enriches our perception [1,2], for example by boosting detection sensitivity [3,4], shortening reaction times (RTs) [5,6], or enhancing attention and memory [7–9]. In the process of audiovisual integration, visual and auditory information significantly influence each other. A well-known phenomenon demonstrating this interplay is the McGurk illusion, where conflicting visual speech cues, such as lip movements, dramatically alter the perception and understanding of heard speech [10,11]. Conversely, auditory information also has a profound effect on visual processing, as evidenced by phenomena such as the Audiovisual Time-Flow Illusion in spoken language [12], and the double flash illusion in the non-speech domain [13,14]. These examples underscore the complex and bidirectional nature of audiovisual integration, revealing how each sensory modality can alter the perception and processing of the other.

Semantic congruency plays a significant role as a high-level factor in the field of audiovisual integration [15,16]. It refers to the association of visual and auditory stimuli that represent different aspects of the same stimulus [17]. This association is typically assessed by presenting auditory stimuli that match the visual stimuli (congruent) or do not match the visual stimuli (incongruent). On

* Corresponding author. School of Information Science, Yunnan University, 650091 Kunming, China.
** Corresponding author.
E-mail addresses: zhuweina.cn@gmail.com (W. Zhu), mail@jandrewes.de (J. Drewes).

https://doi.org/10.1016/j.heliyon.2024.e33828
Received 27 November 2023; Received in revised form 11 June 2024; Accepted 27 June 2024
Available online 27 June 2024

the one hand, a recent study found that the auditory modality sets the pace of the visual modality during audiovisual speech processing rather than reverse [12]. On the other hand, visual processing provides faster access to semantic meaning compared to auditory processing [18,19]. Therefore, as a foundation of experimental studies, visual categorization tasks are more frequently selected than auditory categorization tasks in order to better control the priming effect [20–22]. Additionally, vision plays a dominant role in multi-sensory object recognition, further supporting the choice of visual categorization tasks [19,23]. Several studies have consistently demonstrated that congruent auditory stimuli significantly expedite the recognition of visual stimuli, while incongruent auditory stimuli notably prolong the recognition time of visual stimuli [24–26].

Naturalistic sounds and spoken words represent two different types of auditory semantic stimuli commonly used in visual perception tasks [8,27,27–29]. These stimuli are reflective of two fundamental types of auditory experience in everyday life [15]. Some recent studies found that naturalistic sounds and spoken words exhibit different time courses in facilitating visual performance. For instance, Chen and Spence measured the Reaction Time (RT) in a picture categorization tasks [21,30], with either naturalistic sounds or spoken words presented simultaneously or non-simultaneously. The results demonstrate that auditory stimuli presented prior to visual stimuli induced a stronger audiovisual semantic congruency effect compared to when they were presented simultaneously. Additionally, the semantic congruency effect primed by naturalistic sounds occurred at approximately 500 ms SOA (stimulus onset asynchrony), whereas for spoken words, it occurred at around 1000 ms SOA. Naturalistic sounds directly access associated semantic representations, whereas spoken words are assumed to involve an additional level of lexical processing representations [31–33]. The authors argued that, in order to give rise to similar cross-modal semantic effects, spoken words therefore need to be presented earlier in time as compared to naturalistic sounds.

A further question is whether an audio-visual semantic effect also exist when visual stimuli are not consciously perceived. In a previous study that examined the effects of verbal labels on detecting unseen visual stimuli, participants were presented with auditory labels prior to the briefly presented backward-masked letters, hearing the letter name in advance improved the detection of the corresponding letters compared to hearing noise or irrelevant words [34]. Another study using backward masking have found that when visual stimuli are rendered invisible, congruent auditory stimuli presented beforehand, i.e., semantic priming, can facilitate audiovisual integration, furthermore, natural sounds and spoken words showed different priming advantages times [35,36]. It is worth noting that while backward masking may be effective in interfering with visual processing to prevent stimuli from reaching conscious awareness [37,38]. backward-masked visual stimuli, even when rendered “invisible”, may still be processed semantically (post--perceptual) [39–41].

Continuous flash suppression (CFS) has emerged as a more effective method for studying subliminal visual processes. In a typical CFS paradigm, a series of high contrast dynamic masks are continuously flashed to one eye, causing the target presented in the other eye to be suppressed from awareness for a relatively long time [42]. Several studies have investigated the audiovisual semantic congruency effect using CFS. For instance, one study paired suppressed video events with either semantically congruent or incongruent auditory soundtracks, presented either simultaneously or prior to the visual events [43]. It was found that congruent audiovisual conditions facilitated faster breakthrough of visual stimuli into consciousness, but this facilitatory influence was observed only when audiovisual stimuli were presented simultaneously. Similarly, when complex scenes were suppressed during visual stimulation, the presence of a semantic congruency effect was contingent on simultaneous presentation of auditory stimuli [44]. Whereas in another study, verbal labels (valid, invalid or no-label) were presented prior to the suppressed visual stimulus, with the result indicating that valid labels significantly improved visual performance [40].

Previous research has demonstrated that semantically congruent auditory stimuli can facilitate the processing of visual images, even in situations where visual stimuli are rendered invisible. However, there remains a debate regarding the time course of unconscious audio-visual integration, specifically whether the facilitation effect arises from the presentation of auditory stimuli preceding visual ones, resulting in a crossmodal priming effect, or if it occurs concurrently as a result of simultaneous presentation [43,44], further raising questions about whether the processing patterns for audio-visual semantic congruency remain consistent when visual stimuli are invisible compared to when they are visible. Moreover, the majority of previous studies investigating audio-visual semantic effects have utilized line drawings as visual stimuli [8,21,24,36]. In our daily life, visual objects are often situated within rich contextual backgrounds, and the underlying processing mechanisms may diverge [1,45], particularly in situations where visual stimuli are rendered invisible [44].

Therefore, the purpose of this study is to investigate the audiovisual semantic congruency effects of natural sounds and spoken words on conscious and unconscious processing of naturalistic images, with this goal encompassing two separate experiments. 2AFC paradigms are commonly used in studies investigating audiovisual congruency. We recorded Reaction Times (RTs) both in a conventional 2AFC task (Supraliminal) and in a 2AFC breaking continuous flash suppression task, (b-CFS, Subliminal) [46]. B-CFS is a variation of CFS that allows for a better exploration of the dynamic processes involved in unconscious visual processing [47]. In a b-CFS paradigm, high contrast dynamic pattern masks are presented to one eye, thereby effectively suppressing a stimulus of increasing intensity presented to the other eye. Eventually, the flash suppression will be overcome, such that the previously suppressed stimulus becomes visible.

2. Ethics statement

All experiments were conducted in compliance with the principles in the Declaration of Helsinki (2004) and were approved by the local Ethics Committee of Yunnan University (CHSRE2021015). All participants signed written informed consent.

3. Experiment 1

In this experiment, we replicated a previous study on audiovisual semantic interactions [21], for two primary reasons. Firstly, we aimed to validate the effectiveness of our experimental materials, which involved the use of naturalistic images instead of outline drawings. Secondly, we sought to investigate both conscious and unconscious visual processing using the same set of materials, with Experiment 1 serving as a baseline measure of conscious visual processing. For comparability, the same stereoscopic setup as in the following experiments was used (see below).

We investigated the impact of audiovisual congruency and different stages of information processing (SOA) on visual categorization performance. Participants were required to discern whether the presented reality image was an animal or a non-animal.

3.1. Methods

3.1.1. Participants

20 participants took part in the experiment (7 male, 13 female, aged 18–26), all of them were college students in Yunnan University with normal or corrected-to-normal vision to perform everyday activities. The participants were paid for the participation and were naïve to the research purpose of the study. Ocular dominance was determined by an ABC test [48,49]. Data collection was performend from August 2021 to March 2022.

3.1.2. Apparatus

In all experiments, the visual display was presented on a CRT monitor (19 inch, 1024 pixels $\times$ 768 pixels; refresh rate: 120 Hz). The distance between the participants and the screen was 57 cm. Participants viewed the stimulus through a mirror stereoscope, and the mirrors were adjusted individually to achieve optimal fusion. A chinrest was always used to minimize head movements throughout the experiment. Auditory display was presented through a Sony WH1000-XM4 headset.

Both visual and auditory stimuli were presented in Python using Psychopy [50,51].

3.1.3. Stimulus

20 different kinds of animal reality images and 20 different kinds of non-animal reality images were selected from the internet, 12 images of each category were selected as the experimental materials, and the remaining were used as the pre-experimental training materials. The images were converted to gray-scale in MATLAB using the rgb2gray routine. The brightness and contrast of all images were globally equalized by using the SHINE toolbox to minimize low-level confounds [52]. Samples of the images after equalization are shown in Fig. 1, with 1A representing animal images and 1B non-animal images.

The auditory stimuli (digitized at 48 KHz) were presented using binaural closed-ear headphones, and ranged -16 LUFS $\pm$ 0.5. They consisted of naturalistic sounds and spoken words that corresponded to the type of images presented, for example, the naturalistic sound corresponding to the picture of a cat was a feline meowing, and the spoken word was "猫" ("mao", the Chinese word for cat). The naturalistic sounds were downloaded from the internet, while the spoken words were recorded by a male native Chinese speaker.

Based on previous research [21], the duration of monosyllabic spoken words was controlled at 350 ms, while disyllabic spoken words were controlled at 500 ms. The naturalistic sounds maintained consistent durations with their corresponding spoken words. As a baseline, a white noise auditory cue was used with the same duration as the single-word spoken words (350 ms).

Fig. 1. Sample target images, after equalization. (A) animal images. (B) non-animal images.

3.1.4. Design and procedures

Based on previous research on conscious audio–visual integration [21], we designed a 2-alternative forced choice (2AFC) paradigm in which participants classified the pictures. We used a 2 (Audio type: naturalistic sound, spoken word) × 4 (Semantic Congruency: congruent, relative congruent, incongruent, white noise × 4 (SOA: −1000, −500, −250, 0 ms) within subject design. Additionally, a no sound condition was included as a baseline for comparison, resulting in a total of 33 experimental conditions.

Naturalistic sounds and spoken words are known to have different preferred SOAs. We therefore decided to choose SOAs based on previous research [21].

Examples of the semantic congruency between visual and auditory stimuli are shown in Fig. 2. When a picture depicted a cat and the associated audio was also a cat, the consistency relationship was considered congruent. In the case of the audio being a dog, the consistency relationship was related but not fully consistent since both the cat and dog were biological entities, if not of the same species. If the audio was a piano, the consistency relationship became incongruent because the cat was a living creature while the piano was an inanimate object. When white noise was played as the audio, the consistency relationship was categorized as noise. Lastly, when a picture of a cat was presented without any accompanying audio, the trial was categorized as no sound.

The visual stimuli were presented to the participants through a mirror stereoscope, in which the stimulus images (11.5° × 11.5°) were always presented to the non-dominant eye, and the dominant eye contained only a gray background [53]. Two white frames (13° × 13°) were placed around the images and on the gray background accordingly so each frame would be visible for the corresponding eye only (see Fig. 3). Prior to the formal experiment, individual adjustments were made to the mirror setup and the distance between the two surrounding frames. This was done to ensure optimal fusion between the frames for each participant.

Each trial started with a white central cross shown in a random period between 1400 and 1900 ms. Subsequently, the target image appeared, while the audio stimulus was played according to the SOA in that trial. The target were displayed in 100 % contrast within 1500 ms before disappeared. During this period, participants were required to determine whether the image depicted an animal or a non-animal by pressing the corresponding arrow key (Left key for animal, right key for non-animal). The rear mask appeared 200 ms after the participant pressed a button or after the trial timed out.

The participants were instructed to respond to the pictures rather than the sounds, and to respond as quickly as possible in the correct way. Participants were permitted to rest between trials at any time.

3.1.5. Data analysis

In this study, the main indicators analyzed were reaction time and accuracy. Reaction time was used to assess the speed at which participants acquired visual information when presented with auditory cues. By analyzing the reaction time, we could assess whether the presence of auditory information facilitated or hindered participants' acquisition of visual information, thereby influencing their performance in the target recognition task. Accuracy was primarily used to assess the participants' level of engagement in the experiment and detect any anomalies in the data for each case.

All 20 participants completed the experiment. To ensure proper data processing, this article adheres to two criteria: Firstly, it is necessary to guarantee a correct response rate of over 90 % for each participant, and only data from participants meeting this criterion will be analyzed. Secondly, data with incorrect judgments will be excluded from the dataset of participants meeting the correct response rate, and these incorrect data will not be considered during the analysis of reaction time. After applying the error rate screening, all 20 participants met the accuracy standard. A total of 15,840 data points were collected, resulting in a high collection rate of 96.9 %.

In order to minimize the potential influence of skewed distributions in the data, the median reaction time under each condition for each participant was used for analysis [54].

Repeat measure ANOVA was used for data analysis. The no sound condition, which lacks an SOA, was not included in the ANOVA analysis and served as the baseline. Greenhouse–Geisser adjustments to the degrees of freedom were used when appropriate and are reported in subscripts. Post-hoc multiple comparisons and simple effects analysis employed Bonferroni correction.

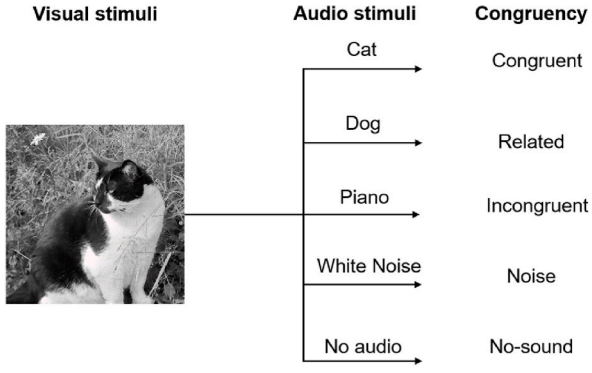

Fig. 2. Semantic congruency illustration.

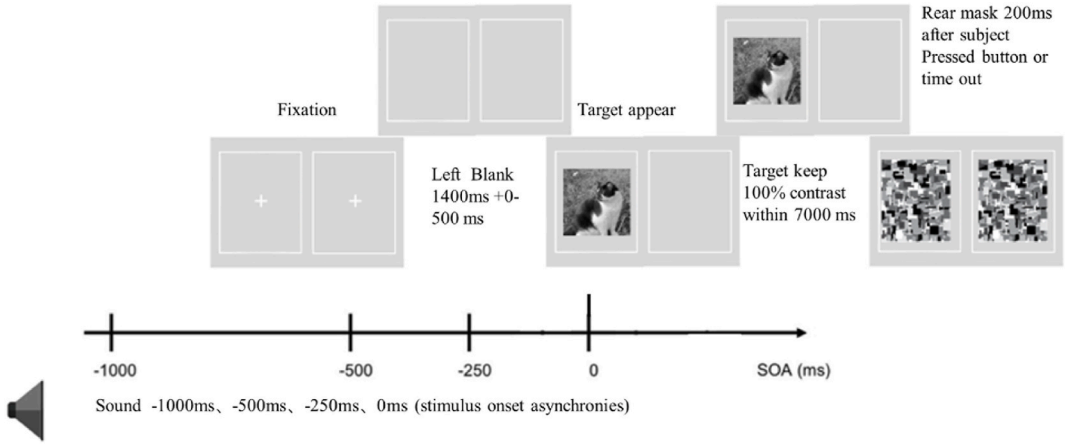

Fig. 3. Schematic representation of the 2AFC paradigm in experiment 1. Naturalistic images were presented to the nondominant, with no target in the other eye. Auditory stimuli were presented either in advance of (−1000, −500, −250 ms SOA) or simultaneously (0 ms SOA) with the target. The 24 images (12 animals and 12 non-animals) were presented once in each of the 33 experimental conditions, in a total of 792 trials. All trials were divided into 4 blocks in a random order. Each participant completed 4 blocks comprised of 198 trials each.

3.2. Results

(1) Accuracy

For both naturalistic sounds (97 % correct) and spoken words (96.9 % correct), the main effects of type, SOAs and congruency, as well as their interactions, were not found to be significant. Therefore, our focus was on analyzing the reaction time.

(2) Reaction time

A three-factor repeated measure ANOVA was performed for the median reaction time, including the factors type (Naturalistic Sounds, Spoken Words), Semantic Congruency (4 congruency types) and SOA (4 levels). Three main effects were significant: Type, $F(1, 19) = 5.461$, $p < 0.031$, $\eta_p^2 = .223$. SOA, $F(3, 57) = 21.559$, $p < 0.031$, $\eta_p^2 = .532$. Congruency Types, $F(3, 2.552) = 24.358$, $p < 0.001$, $\eta_p^2 = .562$. Two two-way interactions were significant: Type and SOA, $F(3, 57) = 9.227$, $p < 0.001$, $\eta_p^2 = .327$. Type and congruency, $F(3, 57) = 5.549$, $p < 0.0013$, $\eta_p^2 = .462$. Three-way interaction was significant, $F(9, 171) = 2.548$, $p < 0.0013$, $\eta_p^2 = .118$.

We conducted separate two-way repeated ANOVAs on congruency and SOA factors for Naturalistic Sounds and Spoken Words. For naturalistic sounds, the main effect of SOA was significant, $F(3, 57) = 7.642$, $p < 0.001$, $\eta_p^2 = .287$ (Fig. 4B). Post-hoc multiple comparisons revealed that the reaction time was shorter at the −500 ms (588.3 ms, SE = 14.2) and −250 ms (588.2 ms, SE = 13.0) SOA than at 0 ms (612.4 ms, SE = 14.2) SOA, $ps < 0.05$. This suggested that presenting auditory stimuli before visual stimuli (−500, −250 ms) speeded up visual stimulus recognition. The main effect of congruency was significant, $F(3, 57) = 9.652$, $p < 0.001$, $\eta_p^2 = .337$ (Fig. 4C). Post-hoc multiple comparisons revealed that the reaction time was shorter in congruent (585.9 ms, SE = 13.7) than incongruent (605.7 ms, SE = 14.0) and noise conditions (597.1 ms, SE = 13.5), $ps < 0.005$. This indicated that visual and auditory congruency speeded up reaction time to the visual stimuli. The interaction between SOA and congruency was not significant (Fig. 4A).

For spoken words, the main effect of both SOA, $F(3, 57) = 29.542$, $p < 0.001$, $\eta_p^2 = .609$ (Fig. 4E), and Congruency, $F(3, 57) = 23.119$, $p < 0.001$, $\eta_p^2 = .549$ (Fig. 4F), were significant. Post-hoc multiple comparisons revealed that the reaction time was shorter at the −1000 (577.9 ms, SE = 13.9), −500 (577.4 ms, SE = 13.3) and −250 ms (588.3 ms, SE = 12.8) SOA than at 0 ms (623.7, SE = 15.1) SOA, $ps < 0.001$. And the reaction time was shorter in the congruent condition (569.5 ms, SE = 13.7) than in the related (593.5 ms, SE = 13.4), incongruent (607.2 ms, SE = 14.1) and noise conditions (607.2 ms, SE = 14.1), $ps < 0.005$. The interaction between SOA and congruency was significant, $F(9, 171) = 5.523$, $p < 0.001$, $\eta_p^2 = .225$ (Fig. 4D). Simple effects analysis revealed that, in congruent, related and incongruent conditions, −1000, −500 and −250 ms SOAs were shorter than 0 ms SOA, $ps < 0.001$, whereas the difference of SOAs were not significant under the noise condition. This suggested that the auditory type "noise" had no SOA-related effect on target categorization, while for the auditory types "congruent," "related," or "incongruent," presenting the auditory stimulus before the visual stimulus speeded up recognition.

Besides, for both naturalistic sounds and spoken words, the no sound condition required the longest reaction time compared to when auditory cues were presented prior to the visual target. This indicated that the auditory stimulus presented prior to the visual stimulus, irrespective of the level of congruency, to some extent enhances the recognition of the visual stimulus.

Furthermore, we directly compared auditory types (natural sounds, spoken words), by conducting separate two-factor repeated measures ANOVAs for type within SOA and type within congruency. For type within SOA (see Fig. 5A), the main effect of both SOA, $F(3, 57) = 21.559$, $p < 0.001$, $\eta_p^2 = .532$, and type, $F(1, 19) = 5.461$, $p = .031$, $\eta_p^2 = .223$, were significant. The interaction between SOA and type was significant, $F(3, 57) = 9.227$, $p < 0.001$, $\eta_p^2 = .327$. Simple effects analysis revealed that, when SOA was −1000 ms, −500

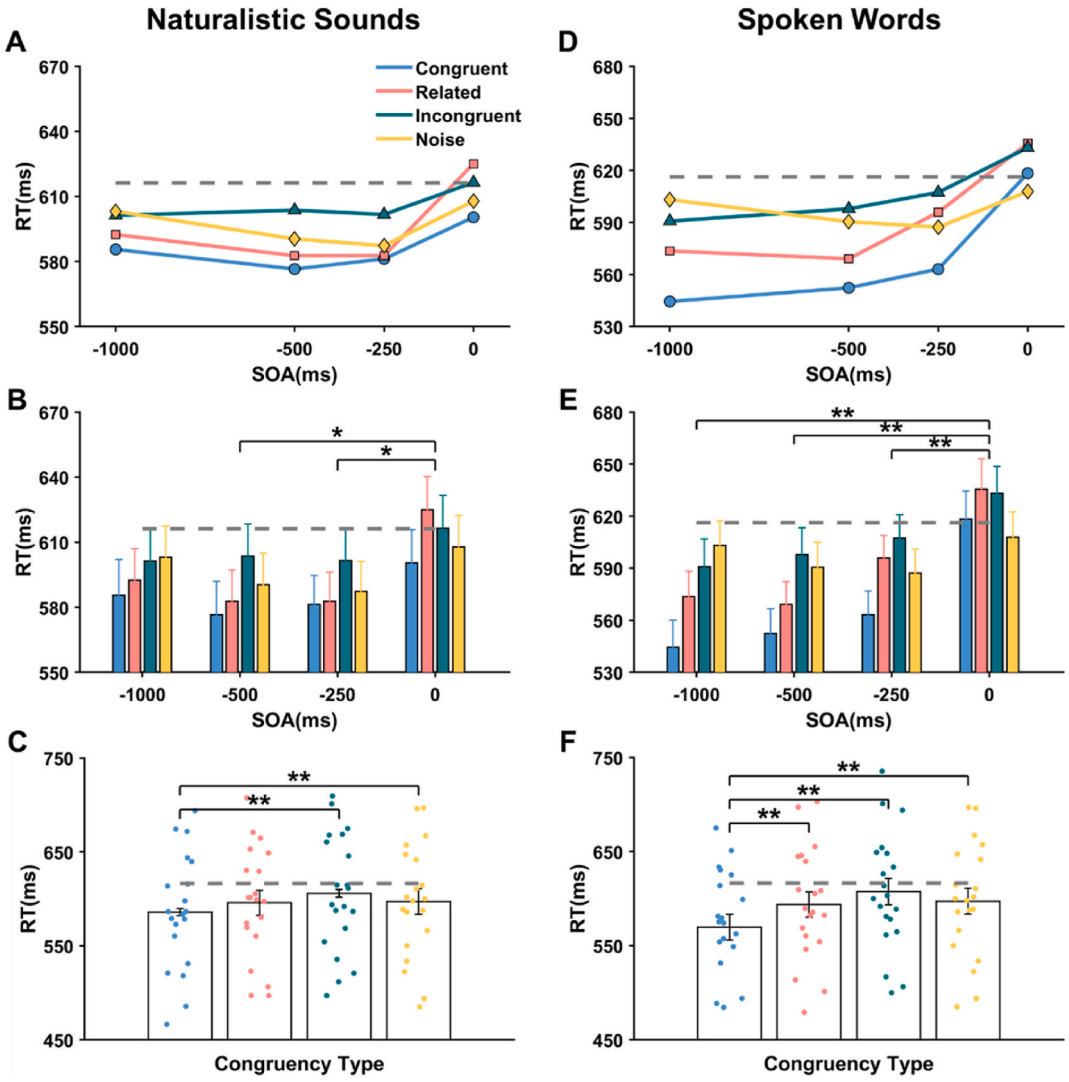

Fig. 4. Median reaction times in Experiment 1 for naturalistic sounds (Left) and spoken words (Right) across different congruency types (congruent, related, incongruent, noise) and Stimulus Onset Asynchrony (SOA) values (−1000, −500, −250, 0 ms). Naturalistic sounds, (A) interaction effect of SOA * congruency type. (B) main effect of SOA. (C) main effect of congruency type. Spoken words, (D) interaction effect of SOA * congruency type. (E) main effect of SOA. (F) main effect of congruency type. No-sound condition is represented with the gray dash line. Error bars indicate ±1 SE; *p < 0.05; **p < 0.01; ***p < 0.001.

ms, spoken words were faster than naturalistic sounds, *ps* < 0.05, whereas naturalistic sounds were faster than spoken words when presented simultaneously (SOA = 0 ms), no difference was observed when SOA was −250 ms.

For type within congruency (see Fig. 5B), the main effect of both congruency, $F(3, 57) = 24.358$, $p < 0.001$, $\eta_p^2 = .562$, and type, $F(1, 19) = 5.461$, $p = .031$, $\eta_p^2 = .223$, were significant. The interaction between SOA and type was significant, $F(3, 57) = 5.549$, $p = 0.002$, $\eta_p^2 = .226$. Simple effects analysis revealed that in congruent conditions, spoken words elicited faster RTs than naturalistic sounds, $p < 0.001$. No difference was observed in related, incongruent and noise conditions.

4. Experiment 2

In this experiment, the visual stimulus was presented unconsciously in a 2AFC-CFS paradigm, with an auditory cue at different SOAs. Throughout the viewing period, a continuous stream (with a specific temporal frequency) of contour-rich, high-contrast masks known as Mondrian patterns was consistently flashed to one eye, effectively suppressing the static low-contrast target image presented to the other eye over a period of time. Participants were asked to make animal vs. non-animal judgments after identifying the type of visual stimulus.

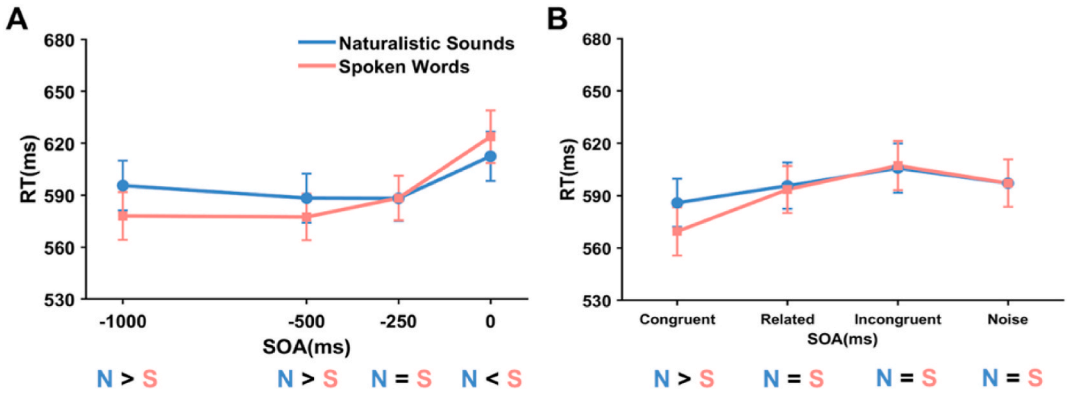

Fig. 5. Median reaction times comparison for Naturalistic Sounds and Spoken Words in Experiment 1. As functions of (A) SOA (−1000, −500, −250, 0 ms). (B) Congruency type (congruent, related, incongruent, noise). The pairwise comparisons were summarized below (N: Naturalistic Sounds, S: Spoken Words).

4.1. Methods

4.1.1. Participants

30 new participants (8 male, 22 female, aged 18–26) from Yunnan University were recruited to the experiment. The participants were paid for the participation and were naïve to the research purpose of the study. Data collection was performend from May to July 2022.

4.1.2. Apparatus and stimulus

The apparatus and target stimuli were the same as Experiment 1.

4.1.3. Design and procedures

The experiment utilized a 2AFC-CFS paradigm, where participants were presented with the visual stimulus under CFS suppression and then made judgments about the image type after breakthrough. To accommodate the expected suppression-delayed responses, we introduced a −750 ms SOA, while also reducing the number of stimulus images to 20 types (10 animals, 10 non-animals) in order to maintain a balanced experiment duration, incorporating the following factors: audio type (naturalistic sound, spoken word), semantic congruency (congruent, relative congruent, incongruent, white noise, no sound), and SOA (−1000, −750, −500, −250, and 0 ms).

Previous b-CFS experiments have revealed individual variations in unconscious visual processing performance [53,55]. To mitigate the influence of these differences, we introduced a composite SOA (Stimulus Onset Asynchrony) approach.

Three days before the main experiment, participants underwent a Baseline SOA Assessment pre-experiment to establish their baseline reaction times (RT). In the experiment (see Fig. 6), the visual stimulus (different congruency types) was presented to the non-dominant eye with contrast gradually increasing from 0 % to 100 % over a 6-s duration, while the dominant eye was presented with only a gray background, and no auditory stimuli were used. We recorded the RTs for each participant under these conditions.

We then introduced a △SOA method based on each participant's response performance in the Baseline SOA Assessment pre-experiment. Firstly, we calculated the median response time for each participant in the Baseline SOA Assessment pre-experiment,

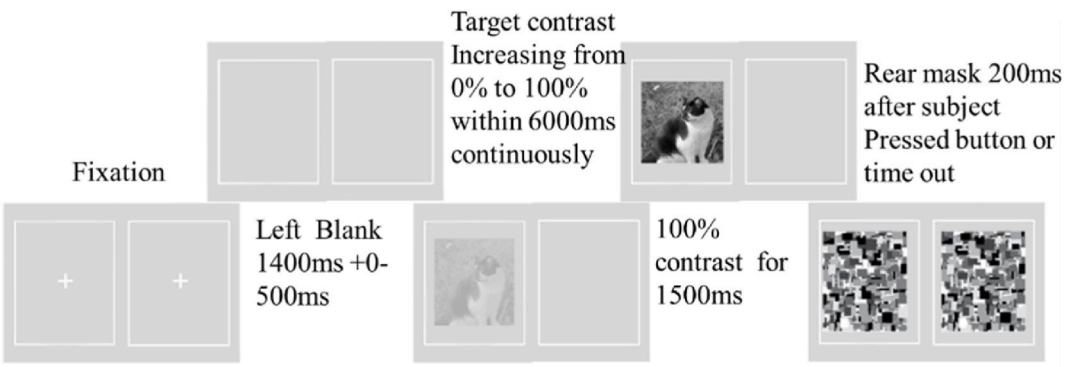

Fig. 6. Schematic representation of the Baseline SOA Assessment paradigm in experiment 2. Naturalistic images were presented to the nondominant, with no target in the other eye. The contrast of the images increased from 0 % to 100 % within 6000 ms continuously.

which served as △SOA. Then, the composite SOA values were derived by shifting the five predetermined SOAs (−1000, −720, −500, −250, 0 ms) by the △SOA. For instance, if a participant's median response time in the Baseline SOA Assessment pre-experiment was 850 ms, their composite SOA values would be −150, 100, 350, 600, and 850 ms, respectively.

In the main experiment (see Fig. 7), the visual stimuli were presented to the participants through a mirror stereoscope, in which high-contrast achromatic Mondrian masks were flashed to the dominant eye at 6 Hz temporal frequency [56], while the animal or non-animal image (11.5° × 11.5°) was presented to the non-dominant eye [46,53,56]. Two white frames (13° × 13°) were positioned around the images and masks to aid fusion of the stereoscopic presentation. Before the formal experiment, adjustments were made to the mirror setup and the distance between the two surrounding frames on an individual basis for each participant, in order to facilitate optimal fusion between the frames.

Previous studies found a "Learning to see" effect in CFS paradigms, with performance increasing towards the end of the experiment [57]. Therefore, to enhance the internal validity and reliability of the experiment, we employed a Latin square design instead of a random sequence as used in previous studies.

The 20 images (10 animals and 10 non-animals) were presented once in each of 40 conditions (2 Auditory types * 5 SOAs * 4 Congruency types). Additionally, there was a 'no-sound' condition that was repeated 4 times, for a total of 44 conditions and 880 trials. To gauge false alarms, we incorporated 110 catch trials, which accounted for 12.5 % of the total number of trials. During catch trials, participants would only hear audio without any visual images presented. In the event of incorrect judgments by participants, a brief, high-frequency tone would serve as a warning signal to indicate the erroneous response.

Each trial began with a white central cross displayed for a random period lasting 400–600 ms. Following this, dynamic Mondrian masks were shown to the dominant eye. After a random delay of 1400–1900 ms from the first mask's appearance, the visual stimulus was presented to the non-dominant eye. The stimulus contrast increased from 0 % to 100 % over 6000 ms and then remained at 100 % for an additional 1500 ms. Simultaneously, the audio stimulus was played based on the trial's specific Stimulus Onset Asynchrony (SOA). Participants were tasked with quickly determining if the image depicted an animal or a non-animal by pressing the left key for animals and the right key for non-animals. Following the participant's response or if the trial timed out, a "rear mask" was presented to participants in both eyes. This "rear mask" served to avoid possible effects of after images.

4.1.4. Data analysis

The analysis steps were analogous to Experiment 1, with the exception of the levels of SOAs.

The data was first submitted to a three-factor repeated measure ANOVA, then analyzed separately based on the type of auditory stimuli (naturalistic sound, spoken word). For each type, a two-factor repeated measure ANOVA was performed for the median reaction time, including the factors Semantic Congruency (4 congruency types) and SOA (5 levels). The no sound condition, which lacks an SOA, was not included in the ANOVA analysis and served as the baseline. Greenhouse–Geisser adjustments to the degree of freedoms were used when appropriate and are reported in subscripts. Post-hoc multiple comparisons and simple effects analysis employ Bonferroni correction.

4.2. Results

A credibility analysis was performed based on the catch trial results, utilizing the proportion of errors in catch trials as the index. The results were as follows: AVG = 1.1 %, STD = 2.4 %, MIN = 0 %, MAX = 10.9 %. These results indicated high data quality from the participants, enabling further analysis to be conducted on this dataset. It is important to note that the reaction times of participants with incorrect judgments on individual trials were not included in the subsequent analysis to eliminate the influence of judgment errors on reaction time analysis.

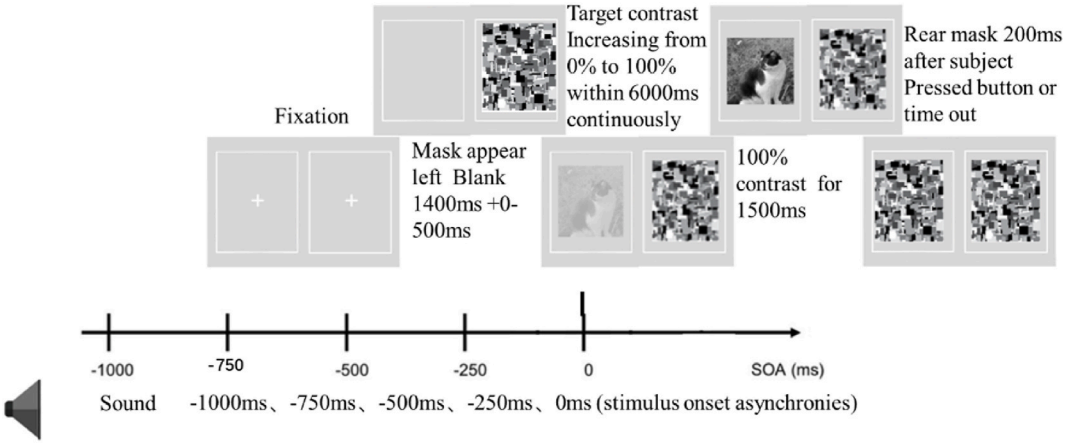

Fig. 7. Schematic representation of the bCFS-2AFC paradigm in experiment 2. Naturalistic images were presented to the nondominant, with dynamic Mondrian masks presented to the other eye. The contrast of the images increased from 0 % to 100 % within 6000 ms continuously.

(1) Accuracy

A three-factor repeated measure ANOVA was performed for the accuracy, including the factors type (Naturalistic Sounds, Spoken Words), Semantic Congruency (4 congruency types) and SOA (5 levels). The main effect of congruency was significant, $F(3, 87) = 5.263$, $p = 0.004$, $\eta_p^2 = .154$. No other effect was found to be significant.

We further conducted separate two-way repeated measures ANOVAs for natural sounds and spoken words. For naturalistic sounds, the main effect of SOA was significant, $F(4_{3.111}, 116_{90.223}) = 3.478$, $p = 0.018$, $\eta^2 = 0.107$, $\varepsilon GG = 0.778$. However, post-hoc multiple comparisons revealed only marginally significant differences for −250 ms SOA (96 % correct) and 0 ms SOA (97.4 correct), $p = 0.075$. This could be attributed to the small effect size.

For spoken words, the main effect of congruency was significant, $F(3, 87) = 5.647$, $p = 0.001$, $\eta_p^2 = .107$. Post-hoc multiple comparisons revealed that the accuracy was higher in the related condition (97.5 % correct) than in the incongruency (96.3 % correct) and noise conditions (95.4 % correct), $ps < 0.001$.

(2) Reaction Time

A three-factor repeated measure ANOVA was performed for the median reaction time, including the factors type (Naturalistic

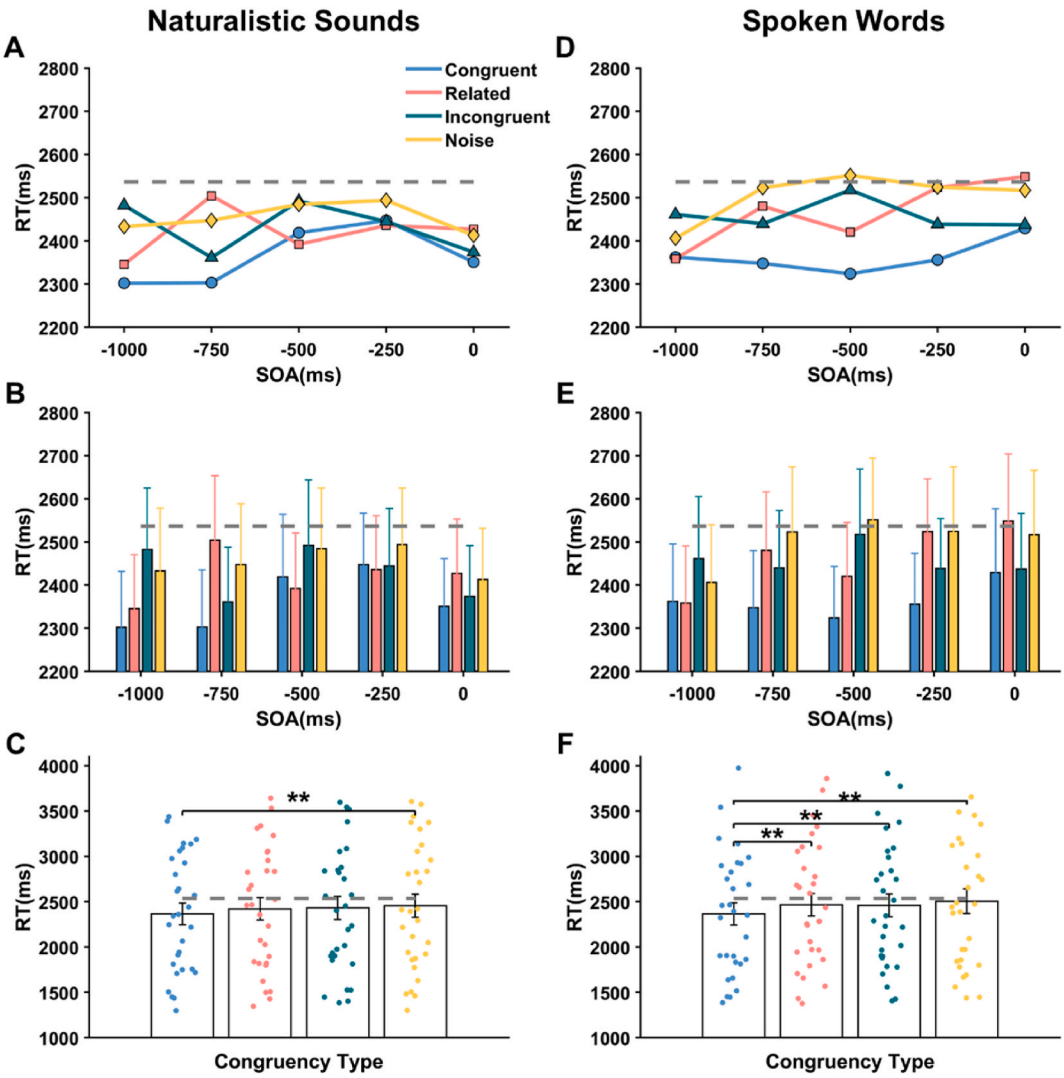

Fig. 8. Median reaction times in Experiment 2 for naturalistic sounds (Left) and spoken words (Right) across different congruency types (congruent, related, incongruent, noise) and Stimulus Onset Asynchrony (SOA) values (−1000, −500, −250, 0 ms). Naturalistic sounds, (A) interaction effect of SOA * congruency type. (B) main effect of SOA. (C) main effect of congruency type. Spoken words, (D) interaction effect of SOA * congruency type. (E) main effect of SOA. (F) main effect of congruency type. No-sound condition is represented with the gray dash line. Error bars indicate ±1 SE; *p < 0.05; **p < 0.01; ***p < 0.001.

Sounds, Spoken Words), Semantic Congruency (4 congruency types) and SOA (5 levels). One main effect was significant: Congruency, $F(3, 87) = 11.726$, $p < 0.001$, $\eta_p^2 = .288$. No other effect was significant. We further conducted separate two-way repeated measures ANOVAs for natural sounds and spoken words.

For naturalistic sounds, the main effect of SOA (Stimulus Onset Asynchrony) did not reach statistical significance, $F(4, 116) = 0.905$, $p = 0.0464$, $\eta_p^2 = .03$ (Fig. 8B). However, the main effect of congruency was significant, $F(3, 87) = 4.425$, $p = 0.006$, $\eta_p^2 = .132$ (Fig. 8C). Post-hoc multiple comparisons revealed that reaction times in the congruent condition (2364.4 ms, SE = 118.2) were significantly shorter than in the noise condition (2454.3 ms, SE = 128.7), with a mean difference of 89.9 ms ($p = 0.016$). The interaction between SOA and congruency was not significant (Fig. 8A). Thus, regardless of the SOAs, auditory stimuli of the same type as visual stimuli demonstrated an advantage in unconscious processing compared to the noise condition.

For spoken words, the main effect of SOA did not reach statistical significance, $F(4, 116) = 1.065$, $p = .394$, $\eta_p^2 = .141$ (Fig. 8E). The main effect of congruency was significant, $F(3, 87) = 12.569$, $p < 0.001$, $\eta_p^2 = .583$ (Fig. 8 F). Post-hoc multiple comparisons revealed that reaction times in the congruent condition (2363.7 ms, SE = 122.5) were significantly shorter than related (2466.2 ms, SE = 125.2), incongruent (2458.9 ms, SE = 127.2) and noise (M = 2504.3 ms, SE = 137.3) conditions ($ps < 0.005$). The interaction between SOA and congruency was not significant (Fig. 8 D). This indicates that congruent auditory information generally outperformed the other three types of auditory information (i.e., related, incongruent, and noise) across various SOA conditions.

5. General discussion

In our current study, we compared the time course of audiovisual semantic effects elicited by naturalistic sound and spoken words on the perception of naturalistic images, either visible or suppressed. When the visual stimulus was visible (Experiment 1), we found that spoken words (−1000, −500, −250 ms) elicited audiovisual semantic integration over a broader time course compared to naturalistic sounds (−500, −250 ms). Meanwhile, congruent auditory stimuli, both naturalistic sounds and spoken words, significantly enhanced the classification performance. However, the effect of SOA appears to occur independent of semantic congruency. Moreover, we found that the performance was consistently better with auditory primes compared to the no-sound condition. Further direct comparisons between spoken words and naturalistic sounds revealed that spoken words presented in advance (−1000, −500 ms) outperformed naturalistic sounds, whereas the opposite was true for simultaneous presentations. Additionally, when audiovisual semantics were congruent, spoken words demonstrated superior categorization performance compared to naturalistic sounds. When visual stimuli were rendered invisible (Experiment 2), semantically congruent auditory stimuli significantly facilitated participants' visual categorization tasks. However, neither naturalistic sounds nor spoken words exhibited a clear time course.

5.1. The audiovisual semantic effect of supraliminal visual processing

In the supraliminal visual categorization tasks, our aim was to replicate previous research within more naturalistic visual settings and to provide a benchmark for sub-threshold classification experiments. Contrary to our expectations, our findings revealed no temporal progression differences in semantic congruency for naturalistic sounds. Spoken words, on the other hand, displayed temporal progression differences across a broader semantic spectrum. Specifically, the effects of semantic congruency (congruent, related, incongruent) concerning naturalistic sounds appeared to be independent of Stimulus Onset Asynchrony (SOA). In contrast, spoken words presented prior to the main stimulus showed a significantly better performance than those presented simultaneously. Further direct comparisons between spoken words and naturalistic sounds indicated that, at earlier presentation times (−1000 ms, −500 ms), the categorization performance for spoken words was faster than for naturalistic sounds. However, when presented simultaneously, spoken words were slower than naturalistic sounds. Moreover, both spoken words and naturalistic sounds outperformed the no-sound condition. These findings suggest interpretations from two perspectives.

First, the activation mechanisms and performances of spoken words and naturalistic sounds may differ. Spoken words demonstrated superior performance starting from −1000 ms (at −1000 ms, −500 ms, and −250 ms) in advance compared to simultaneous presentation, while naturalistic sounds showed this advantage starting from −500 ms (at −500 ms and −250 ms). The occurrence of the audio-visual semantic effect induced by naturalistic sounds at shorter SOAs presented beforehand in the supraliminal image classification task indicates that accessing meaning for the former is indeed faster than suggested by spoken words [36]. Such result can be explained by the potentially different natures of naturalistic sounds and spoken words. For example, several ERP studies observed an asymmetric hemispheric laterality for the two types, with N400s to words evoking larger responses in the right hemisphere and environmental sounds eliciting larger responses in the left hemisphere [58,59]. Besides, brain responses to natural sounds from living things can be distinguished around 100 ms[60], while the semantic component for spoken words (N400) typically emerges about 200 ms after onset [60]. Our findings support the notion that naturalistic sounds directly engage semantic processing, whereas spoken words appear to access their meanings through a semantic network or lexical representations [21,36]. Moreover, spoken words presented in advance (−1000 ms, −500 ms) performed better than naturalistic sounds, while the reverse was true for simultaneous presentations. This finding aligns with prior research, which established that familiar and clear environmental sounds, like a dog's bark, lead to slower recognition than their corresponding verbal cues, such as the word "dog", when viewing images of the same dog [61]. These outcomes highlight the superior efficacy of spoken words in conceptual activation. Specifically, spoken words trigger the concept of a dog with greater abstraction and precision [27], resulting in enhanced performance post-activation compared to that achieved through naturalistic sounds.

Secondly, the effect of SOA appears unrelated to auditory conditions, meanwhile we observed a consistent advantage for auditory conditions over no-sound conditions, regardless of congruency types [62,63]. These findings differ from a recent study by Chen and

Spence [21], which found that the audiovisual semantic congruency effect exhibits different temporal dynamics for spoken words and naturalistic sounds. Furthermore, only specific congruency conditions outperformed the no-sound condition - in the related or the noise conditions, performance was even slower than in the no-sound condition. A plausible explanation for this disparity may stem from the complexity and naturalness of the visual stimuli used. In our study, the utilization of real-world images, which are more complex and ecologically valid, might have served a dual purpose. In addition to facilitating audiovisual integration, these naturalistic images may also have triggered an alerting effect, which has been found to reduce reaction times in response to various stimuli [64,65], and may occur during earlier perceptual processes [66,67]. This potential alerting effect could have affected or even be the cause for the consistent advantage for auditory conditions, overshadowing the variations seen in sound congruency. In contrast, the prior study, which relied on simpler line drawings, might not have triggered this alerting effect to the same degree, leading to different performance outcomes.

5.2. The audiovisual semantic effect of subliminal visual processing

Since the emergence of CFS, there is a growing body of research starting to support unconscious semantic processing. Native Hebrew speakers were found to process invisible Hebrew words faster than Chinese characters while native Chinese speakers reacted more slowly to invisible Hebrew characters [46]. The semantic processing of a single word can be extracted without visual awareness [68,69]. Congruent invisible lip movements facilitated the classification of spoken target words [70]. Further evidence from the cross-modal domain shows dynamic events that are suppressed interocularly to achieve faster access to visual awareness when congruent with naturalistic sounds [43]. Hearing an associated verbal label prior to the visual target can boost an invisible object into awareness [57]. Furthermore, the findings from Shahin et al. (2017) and Bhat et al. (2015) on audiovisual onset asynchrony (AVOA) provide additional context to our results. These studies demonstrate that the perception of audiovisual synchrony for words can vary with the reliability of the acoustic and visual stimuli and is evident as early as the P1 auditory evoked potential. This suggests that the brain begins to process audiovisual synchrony at a very early stage. Such early processing might underlie the semantic analysis of audiovisual information in our experiments.

The present study showed that interocular suppressed naturalistic images can be classified faster when semantically congruent with sound, for both naturalistic sounds and spoken words. This finding suggests that early-stage visual processing can incorporate audio-visual integration, wherein semantically congruent auditory stimuli establish a connection with subliminal visual stimuli, facilitating faster breakthroughs into visual consciousness [71]. However, this result alone does not provide a conclusive answer to whether subliminal visual stimuli undergo semantic analysis during audio-visual integration.

5.3. The time course in audiovisual semantic integration

The simultaneous presentation of auditory stimuli exhibits varying effects when visual stimuli are visible versus when they are not. Specifically, during Experiment 1, when a congruent auditory stimulus was concurrently presented with a visual stimulus at a constant 100 % contrast, we did not find any significant audiovisual semantic congruency effect. This finding was consistent with prior research by Chen and Spence [21,30,35], where no cross-modal facilitation was detected following the simultaneous presentation of a congruent cue. These earlier studies employed tasks involving picture detection and picture categorization. In contrast, in experiment 2, we observed an audiovisual semantic congruency effect during simultaneous audiovisual presentation. Notably, the contrast of the visual stimulus in Experiment 2 gradually increased from 0 % to 100 %. One possible interpretation of these findings is the role of task demands in shaping audiovisual interactions [21]. In scenarios where the response threshold for the visual target is rapidly attained, such as in the case of the supraliminal picture detection task, it appears that the simultaneously presented sound may lack the temporal window required to access its semantic meaning before the task is concluded. However, CFS-like subliminal visual classification tasks require more time, affording the concurrently presented sound ample opportunity to access its semantic meaning and subsequently influence visual processing.

While we balanced the reaction time differences among participants, our study did not reveal significant differences in the time course of audiovisual integration for both naturalistic sounds and spoken words. Both congruent auditory stimuli, whether presented in advance or simultaneously, demonstrated the ability to enhance the perceptual emergence of subliminal visual stimuli. Previous research has indicated that introducing auditory stimuli in advance does not enhance subliminal visual processing [43,44]. One plausible explanation is their utilization of a soundtrack presented 3 s before the visual stimulus, possibly occurring too early to establish effective temporal proximity [72].

Taken together, CFS-like experiments, due to their extended processing time, may still exhibit audiovisual priming effects even when presented simultaneously. Given the audiovisual priming effect observed in the current study, it appears plausible that the integration of unconscious visual stimuli necessitates semantic analysis of visual information, although it may not reach the level observed in supraliminal visual processing.

5.4. Conclusion

This study explored the temporal dynamics of cross-modal semantic effects induced by naturalistic sounds and spoken words on the processing of visual stimuli, including conditions where visual stimuli were visible or suppressed from consciousness through continuous flash suppression. The findings revealed two key insights.

1. Under conditions where visual stimuli were visible, spoken words, compared to naturalistic sounds, were capable of eliciting audio-visual semantic effects over a longer duration (−1000, 500, −250 ms SOA) and demonstrated superior performance, especially when the stimuli were congruent.
2. The audio-visual semantic congruency effect persisted even when the visual stimuli were suppressed, thereby becoming imperceptible to consciousness, but there were no significant differences in the temporal patterns between types of auditory stimuli.

6. Limitations of the study

Increasing evidence suggests that the meaning of complex visual objects, faces, and symbols can be extracted even when suppressed at the unconscious level using the Continuous Flash Suppression (CFS) paradigm [73]. Compared to backward masking, CFS extends the processing time for subliminal stimuli, allowing for more thorough processing of subliminal semantic visual stimuli. This is one of the reasons we chose CFS for our study on subliminal audiovisual semantic integration. However, in CFS experiments, the contrast of the suppressed stimuli gradually increases, allowing for the gradual formation of stimulus representation even when not consciously perceived. In this way, CFS prolongs the time for subliminal processing, enhancing the effectiveness of unconscious research. This leads to variability in breakthrough times for CFS both across participants and within trials, making it challenging to study the temporal dynamics of subliminal audiovisual integration. Since performance in suprathreshold picture categorization tasks is relatively stable, our study employs an exploratory calibration method using subtractive reasoning: 2AFC-CFS minus 2AFC-CFS without mask ≈ unconscious processing. This approach aims to balance these temporal differences as much as possible. However, there is no direct evidence that this method effectively reduces the variability associated with the Continuous Flash Suppression (CFS) paradigm. Perhaps CFS needs further enhancements for studying the temporal processes of subliminal audiovisual integration, or it may require more sophisticated controls. Future research on the methodology of CFS could focus on how to reduce trial-to-trial and participant-to-participant variability.

Data availability statement

Data will be made available upon request.

CRediT authorship contribution statement

Mingjie Gao: Writing – original draft, Methodology, Formal analysis. **Weina Zhu:** Writing – review & editing, Supervision, Conceptualization. **Jan Drewes:** Writing – review & editing, Supervision, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

WZ was supported by the National Natural Science Foundation of China (61263042, 61563056). We would like to thank Kun Zhou for helping with data collection.

References

[1] R. Epstein, The cortical basis of visual scene processing, Vis. Cognit. 12 (6) (2005) 954–978.
[2] K.K. Evans, A. Treisman, Natural cross-modal mappings between visual and auditory features, J. Vis. 10 (1) (2010) 6, 6.
[3] T.S. Andersen, P. Mamassian, Audiovisual integration of stimulus transients, Vis. Res. 48 (25) (2008) 2537–2544.
[4] R. Eramudugolla, R. Henderson, J.B. Mattingley, Effects of audio–visual integration on the detection of masked speech and non-speech sounds, Brain Cognit. 75 (1) (2011) 60–66.
[5] E. Ben-Artzi, L.E. Marks, Visual-auditory interaction in speeded classification: role of stimulus difference, Percept. Psychophys. 57 (8) (1995) 1151–1162.
[6] O. Collignon, S. Girard, F. Gosselin, S. Roy, D. Saint-Amour, M. Lassonde, F. Lepore, Audio-visual integration of emotion expression, Brain Res. 1242 (2008) 126–135.
[7] R. Adam, U. Noppeney, A phonologically congruent sound boosts a visual target into perceptual awareness, Front. Integr. Neurosci. 8 (2014) 70.
[8] J. Heikkilä, K. Alho, H. Hyvönen, K. Tiippana, Audiovisual semantic congruency during encoding enhances memory performance, Exp. Psychol. 62 (2) (2015) 123–130.
[9] C.N. Olivers, E. Van der Burg, Bleeping you out of the blink: sound saves vision from oblivion, Brain Res. 1242 (2008) 191–199.
[10] H. McGurk, J. MacDonald, Hearing lips and seeing voices, Nature 264 (5588) (1976) 746–748.
[11] L.D. Rosenblum, Speech perception as a multimodal phenomenon, Curr. Dir. Psychol. Sci. 17 (6) (2008) 405–409, https://doi.org/10.1111/j.1467-8721.2008.00615.x.
[12] M.G. Gonzales, K.C. Backer, Y. Yan, L.M. Miller, H. Bortfeld, A.J. Shahin, Audition controls the flow of visual time during multisensory perception, iScience 25 (7) (2022).
[13] J. Keil, Double flash illusions: current findings and future directions, Front. Neurosci. 14 (2020) 298.
[14] L. Shams, Y. Kamitani, S. Shimojo, What you see is what you hear, Nature 408 (6814) (2000) 788, 788.
[15] P.J. Laurienti, R.A. Kraft, J.A. Maldjian, J.H. Burdette, M.T. Wallace, Semantic congruence is a critical factor in multisensory behavioral performance, Exp. Brain Res. 158 (2004) 405–414.
[16] C. Spence, Crossmodal correspondences: a tutorial review, Atten. Percept. Psychophys. 73 (2011) 971–995.

[17] O. Doehrmann, M.J. Naumer, Semantics and the multisensory brain: how meaning modulates processes of audio-visual integration, Brain Res. 1242 (2008) 136–150.
[18] Y. Kim, A.M. Porter, P. Goolkasian, Conceptual priming with pictures and environmental sounds, Acta Psychol. 146 (2014) 73–83.
[19] K. Weatherford, M. Mills, A.M. Porter, P. Goolkasian, Target categorization with primes that vary in both congruency and sense modality, Front. Psychol. 6 (2015) 20.
[20] A. Alsius, K.G. Munhall, Detection of audiovisual speech correspondences without visual awareness, Psychol. Sci. 24 (4) (2013) 423–431.
[21] Y.-C. Chen, C. Spence, Audiovisual semantic interactions between linguistic and nonlinguistic stimuli: the time-courses and categorical specificity, J. Exp. Psychol. Hum. Percept. Perform. 44 (10) (2018) 1488.
[22] J.T. Enns, S. Soto-Faraco, Cross-modal prediction in speech perception, PLoS One 6 (10) (2011).
[23] A.A. Ghazanfar, C.E. Schroeder, Is neocortex essentially multisensory? Trends Cognit. Sci. 10 (6) (2006) 278–285.
[24] Y.-C. Chen, C. Spence, When hearing the bark helps to identify the dog: semantically-congruent sounds modulate the identification of masked pictures, Cognition 114 (3) (2010) 389–404.
[25] P.J. Laurienti, M.T. Wallace, J.A. Maldjian, C.M. Susi, B.E. Stein, J.H. Burdette, Cross-modal sensory processing in the anterior cingulate and medial prefrontal cortices, Hum. Brain Mapp. 19 (4) (2003) 213–223.
[26] T.R. Schneider, A.K. Engel, S. Debener, Multisensory identification of natural objects in a two-way crossmodal priming paradigm, Exp. Psychol. 55 (2) (2008) 121–132.
[27] P. Edmiston, G. Lupyan, What makes words special? Words as unmotivated cues, Cognition 143 (2015) 93–100, https://doi.org/10.1016/j.cognition.2015.06.008.
[28] D. Kvasova, L. Garcia-Vernet, S. Soto-Faraco, Characteristic sounds facilitate object search in real-life scenes, Front. Psychol. 10 (2019) 2511.
[29] J.R. Williams, Y.A. Markov, N.A. Tiurina, V.S. Störmer, What you see is what you hear: sounds alter the contents of visual perception, Psychol. Sci. 33 (12) (2022) 2109–2122.
[30] Y.-C. Chen, C. Spence, The time-course of the cross-modal semantic modulation of visual picture processing by naturalistic sounds and spoken words, Multisensory Res. 26 (4) (2013) 371–386.
[31] M. Coltheart, Dual routes from print to speech and dual routes from print to meaning: some theoretical issues, Read. Percept. Process (2000) 475–490.
[32] W.R. Glaser, M.O. Glaser, Context effects in stroop-like word and picture processing, J. Exp. Psychol. Gen. 118 (1) (1989) 13.
[33] A. Roelofs, The visual-auditory color-word stroop asymmetry and its time course, Mem. Cognit. 33 (8) (2005) 1325–1336.
[34] G. Lupyan, M.J. Spivey, Making the invisible visible: auditory cues facilitate visual object detection, PLoS One 5 (7) (2010) e11452.
[35] Y.-C. Chen, C. Spence, Crossmodal semantic priming by naturalistic sounds and spoken words enhances visual sensitivity, J. Exp. Psychol. Hum. Percept. Perform. 37 (5) (2011) 1554.
[36] Y.-C. Chen, C. Spence, Dissociating the time courses of the cross-modal semantic priming effects elicited by naturalistic sounds and spoken words, Psychon. Bull. Rev. 25 (2018) 1138–1146.
[37] A.G. Greenwald, S.C. Draine, R.L. Abrams, Three cognitive markers of unconscious semantic activation, Science 273 (5282) (1996) 1699–1702.
[38] J.D. Knotts, H. Lau, M.A. Peters, Continuous flash suppression and monocular pattern masking impact subjective awareness similarly, Atten. Percept. Psychophys. 80 (2018) 1974–1987.
[39] S. Kouider, S. Dehaene, Levels of processing during non-conscious perception: a critical review of visual masking, Philos. Trans. R. Soc. B Biol. Sci. 362 (1481) (2007) 857–875.
[40] G. Lupyan, E.J. Ward, Language can boost otherwise unseen objects into visual awareness, Proc. Natl. Acad. Sci. USA 110 (35) (2013) 14196–14201.
[41] Y.-H. Yang, J. Zhou, K.-A. Li, T. Hung, A.J. Pegna, S.-L. Yeh, Opposite ERP effects for conscious and unconscious semantic processing under continuous flash suppression, Conscious. Cognit. 54 (2017) 114–128.
[42] N. Tsuchiya, C. Koch, Continuous flash suppression reduces negative afterimages, Nat. Neurosci. 8 (8) (2005) 1096–1101.
[43] D. Cox, S.W. Hong, Semantic-based crossmodal processing during visual suppression, Front. Psychol. 6 (2015) 722.
[44] J.-S. Tan, S.-L. Yeh, Audiovisual integration facilitates unconscious visual scene processing, J. Exp. Psychol. Hum. Percept. Perform. 41 (5) (2015) 1325.
[45] J.L. Davenport, M.C. Potter, Scene consistency in object and background perception, Psychol. Sci. 15 (8) (2004) 559–564.
[46] Y. Jiang, P. Costello, S. He, Processing of invisible stimuli: advantage of upright faces and recognizable words in overcoming interocular suppression, Psychol. Sci. 18 (4) (2007) 349–355.
[47] S. Gayet, S. Van der Stigchel, C.L. Paffen, Breaking continuous flash suppression: competing for consciousness on the pre-semantic battlefield, Front. Psychol. 5 (2014) 460.
[48] W.R. Miles, Ocular dominance demonstrated by unconscious sighting, J. Exp. Psychol. 12 (2) (1929) 113.
[49] W.R. Miles, Ocular dominance in human adults, J. Gen. Psychol. 3 (3) (1930) 412–430.
[50] J.W. Peirce, PsychoPy—psychophysics software in Python, J. Neurosci. Methods 162 (1–2) (2007) 8–13.
[51] M.F. Sanner, Python: a programming language for software integration and development, J. Mol. Graph. Model. 17 (1) (1999) 57–61.
[52] V. Willenbockel, J. Sadr, D. Fiset, G.O. Horne, F. Gosselin, J.W. Tanaka, Controlling low-level image properties: the SHINE toolbox, Behav. Res. Methods 42 (2010) 671–684.
[53] W. Zhu, J. Drewes, N.A. Peatfield, D. Melcher, Differential visual processing of animal images, with and without conscious awareness, Front. Hum. Neurosci. 10 (2016) 513.
[54] R. Whelan, Effective analysis of reaction time data, Psychol. Rec. 58 (2008) 475–482.
[55] R. Blake, R. Goodman, A. Tomarken, H.-W. Kim, Individual differences in continuous flash suppression: potency and linkages to binocular rivalry dynamics, Vis. Res. 160 (2019) 10–23.
[56] W. Zhu, J. Drewes, D. Melcher, Time for awareness: the influence of temporal properties of the mask on continuous flash suppression effectiveness, PLoS One 11 (7) (2016) e0159206.
[57] G. Lupyan, E.J. Ward, Language can boost otherwise unseen objects into visual awareness, Proc. Natl. Acad. Sci. USA 110 (35) (2013) 14196–14201.
[58] E. Plante, C. Van Petten, A.J. Senkfor, Electrophysiological dissociation between verbal and nonverbal semantic processing in learning disabled adults, Neuropsychologia 38 (13) (2000) 1669–1684.
[59] C. Van Petten, H. Rheinfelder, Conceptual relationships between spoken words and environmental sounds: event-related brain potential measures, Neuropsychologia 33 (4) (1995) 485–508.
[60] C. Van Petten, S. Coulson, S. Rubin, E. Plante, M. Parks, Time course of word identification and semantic integration in spoken language, J. Exp. Psychol. Learn. Mem. Cogn. 25 (2) (1999) 394.
[61] G. Lupyan, S.L. Thompson-Schill, The evocative power of words: activation of concepts by verbal and nonverbal means, J. Exp. Psychol. Gen. 141 (1) (2012) 170.
[62] C.J. Marsolek, What antipriming reveals about priming, Trends Cognit. Sci. 12 (5) (2008) 176–181.
[63] H.A. Simon, Information processing models of cognition, Annu. Rev. Psychol. 30 (1) (1979) 363–396.
[64] J. Fan, B.D. McCandliss, T. Sommer, A. Raz, M.I. Posner, Testing the efficiency and independence of attentional networks, J. Cognit. Neurosci. 14 (3) (2002) 340–347.
[65] M.I. Posner, S.J. Boies, Components of attention, Psychol. Rev. 78 (5) (1971) 391.
[66] F. Kusnir, A.B. Chica, M.A. Mitsumasu, P. Bartolomeo, Phasic auditory alerting improves visual conscious perception, Conscious. Cognit. 20 (4) (2011) 1201–1210.
[67] E. Matthias, P. Bublak, H.J. Müller, W.X. Schneider, J. Krummenacher, K. Finke, The influence of alertness on spatial and nonspatial components of visual attention, J. Exp. Psychol. Hum. Percept. Perform. 36 (1) (2010) 38.

[68] J. Zhou, C.-L. Lee, K.-A. Li, Y.-H. Tien, S.-L. Yeh, Does temporal integration occur for unrecognizable words in visual crowding? PLoS One 11 (2) (2016) e0149355.
[69] J. Zhou, C.-L. Lee, S.-L. Yeh, Word meanings survive visual crowding: evidence from ERPs, Lang. Cogn. Neurosci. 31 (9) (2016) 1167–1177.
[70] J. Plass, E. Guzman-Martinez, L. Ortega, M. Grabowecky, S. Suzuki, Lip reading without awareness, Psychol. Sci. 25 (9) (2014) 1835–1837.
[71] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, G. Rigoll, LSTM-modeling of continuous emotions in an audiovisual affect recognition framework, Image Vis Comput. 31 (2) (2013) 153–163.
[72] G.B. Remijn, H. Ito, Y. Nakajima, Audiovisual integration: an investigation of the 'streaming-Bouncing'Phenomenon, J. Physiol. Anthropol. Appl. Hum. Sci. 23 (6) (2004) 243–247.
[73] A. Pournaghdali, B.L. Schwartz, Continuous flash suppression: known and unknowns, Psychon. Bull. Rev. 27 (6) (2020) 1071–1103, https://doi.org/10.3758/s13423-020-01771-2.