



Article

Audiovisual Emotional Congruency Modulates the Stimulus-Driven Cross-Modal Spread of Attention

Minran Chen ^{1,†} , Song Zhao ^{1,†}, Jiaqi Yu ¹, Xuechen Leng ¹, Mengdie Zhai ¹, Chengzhi Feng ¹
and Wenfeng Feng ^{1,2,*} 

¹ Department of Psychology, School of Education, Soochow University, Suzhou 215123, China

² Research Center for Psychology and Behavioral Sciences, Soochow University, Suzhou 215123, China

* Correspondence: fengwf@suda.edu.cn; Tel.: +86-0512-65880907

† These authors contributed equally to this work.

Abstract: It has been reported that attending to stimuli in visual modality can spread to task-irrelevant but synchronously presented stimuli in auditory modality, a phenomenon termed the cross-modal spread of attention, which could be either stimulus-driven or representation-driven depending on whether the visual constituent of an audiovisual object is further selected based on the object representation. The stimulus-driven spread of attention occurs whenever a task-irrelevant sound synchronizes with an attended visual stimulus, regardless of the cross-modal semantic congruency. The present study recorded event-related potentials (ERPs) to investigate whether the stimulus-driven cross-modal spread of attention could be modulated by audio-visual emotional congruency in a visual oddball task where emotion (positive/negative) was task-irrelevant. The results first demonstrated a prominent stimulus-driven spread of attention regardless of audio-visual emotional congruency by showing that for all audiovisual pairs, the extracted ERPs to the auditory constituents of audiovisual stimuli within the time window of 200–300 ms were significantly larger than ERPs to the same auditory stimuli delivered alone. However, the amplitude of this stimulus-driven auditory Nd component during 200–300 ms was significantly larger for emotionally incongruent than congruent audiovisual stimuli when their visual constituents' emotional valences were negative. Moreover, the Nd was sustained during 300–400 ms only for the incongruent audiovisual stimuli with emotionally negative visual constituents. These findings suggest that although the occurrence of the stimulus-driven cross-modal spread of attention is independent of audio-visual emotional congruency, its magnitude is nevertheless modulated even when emotion is task-irrelevant.

Keywords: emotional congruency; spread of attention; stimulus-driven; cross-modal interaction; ERP



Citation: Chen, M.; Zhao, S.; Yu, J.; Leng, X.; Zhai, M.; Feng, C.; Feng, W. Audiovisual Emotional Congruency Modulates the Stimulus-Driven Cross-Modal Spread of Attention. *Brain Sci.* **2022**, *12*, 1229. <https://doi.org/10.3390/brainsci12091229>

Academic Editors: Aihua Chen and Antoine Shahin

Received: 14 August 2022

Accepted: 7 September 2022

Published: 10 September 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In recent years, researchers have increasingly focused on the neurophysiological mechanisms involved in multisensory processing. It has long been a question how the brain integrates information from different modalities to create the perception of a unified object. Numerous studies have supported the idea that attention plays a crucial role in multisensory integration (for reviews, see [1–4]). A portion of these studies used EEG/ERP techniques to reveal the electrophysiological bases of these effects [5–7]. EEG (electroencephalogram) signals are the electrical activities of human brain that measure field potential in the space around neurons and can be categorized into several types of activities with the characteristic of specific signal frequencies and amplitudes such as: Alpha, Beta, Delta, Gamma, Theta and Mu [8]. However, it is worth mentioning that the neural responses associated with specific sensory, cognitive, and motor events are drowned within the EEG. These specific neural responses are called event-related potentials (ERPs) which can be extracted from the overall EEG by means of a simple averaging technique [9]. One of the most striking neural examples illustrating the crucial role of attention on multisensory

integration is that visual attention effect can spread to the task-irrelevant auditory modality, resulting in the originally unattended auditory features of an audiovisual object being pulled into the attentional spotlight and bestowed with enhanced processing [10]. This cross-modal attentional spreading effect is typically manifested as a sustained ERP difference analogous to the attention-related auditory Nd component [11–14]. The Nd is an ERP difference component manifested as greater negative amplitude over the fronto-central scalp elicited by attended relative to unattended auditory stimuli, beginning around 150–200 ms post-stimulus [15], which is generally thought to reflect an enhancement of auditory processing due to selective attention [16,17].

The cross-modal spread of attention can be elicited either in a stimulus-driven manner [10,11,14,18–20] or/and in a representation-driven manner [12,13,21,22]. Specifically, the *stimulus-driven* spread of attention occurs whenever a task-irrelevant sound is presented simultaneously with an attended visual stimulus, and thus was interpreted as a bottom-up audiovisual binding process on the basis of the temporal co-occurrence of visual and auditory stimuli [13,14]. In contrast, the *representation-driven* spread of attention occurs only when the object representation of a task-irrelevant sound (e.g., a bark of a dog) is semantically congruent with that of the visual stimulus receiving extra representation-based selective attention (e.g., when searching for an image of a dog), hence was considered to be a top-down audiovisual priming process relying on the activation of highly learned associations between features in different modalities [12,13]. In addition, if a semantically congruent sound is delivered synchronously with the visual stimulus receiving representation-based selective attention, both types of the attentional spreading occur in an additive manner [8,16,17].

The stimulus-driven cross-sensory attentional spreading has been investigated more often than the representation-driven one (stimulus-driven: [10,11,14,18–20]; representation-driven: [12]; both: [13,21,22]), partially because the bottom-up nature of the former enables it to occur under much more circumstances. Among studies involving the stimulus-driven attentional spreading, the most intensively explored issue is the influence of higher-level audiovisual congruency (congruent vs. incongruent), but the results are mixed: some studies found a null effect [13,21] while other studies showed that its magnitude was greater in response to incongruent than congruent audiovisual pairs [18,19]. Regardless of the discrepancy, it is noteworthy that all of these studies have only manipulated the *semantic* congruency between non-emotional visual and auditory stimuli (e.g., an image of a dog with a bark of a dog vs. an image of a car with a bark of a dog). However, real-life visual and auditory stimuli may also convey emotional information (such as when seeing a championship trophy or hearing a gloomy melody), and the ability to integrate emotional information appropriately from the visual and auditory modalities is fundamental to emotion recognition and social interaction (for review, see [23]). Therefore, in order to advance our understanding regarding whether and how higher-level audiovisual congruency modulates the stimulus-driven cross-sensory spread of attention, it is necessary to take the *emotional* congruency between visual and auditory stimuli into consideration.

Human ERP studies in recent years have consistently reported that emotionally incongruent audiovisual pairs can evoke smaller auditory P2 and/or greater auditory N2 amplitudes over the fronto-central scalp than emotionally congruent audiovisual pairs [24–28], with the timing and scalp distribution of this P2/N2 modulation resembling the neural correlate of the stimulus-driven spread of attention (i.e., the auditory Nd difference component). Moreover, some of these studies even found an earlier occurring audiovisual emotional congruency effect during the auditory N1 interval [24,25,28]. However, it should be noted that the task paradigms utilized in these studies required participants either to *attend to the auditory modality voluntarily* and/or to *respond based on emotional information conveyed by the stimuli*. In the former case, it is insufficient to determine whether the observed P2/N2 modulation, as well as the earlier N1 modulation, is associated with the stimulus-driven attentional spreading, because the stimulus-driven process is measurable only when the crucial auditory stimuli are initially ignored, but not attended [11]. In the

latter case, the task-relevant nature of high-order representations (i.e., emotions) might have rendered the underlying bottom-up, stimulus-driven attentional spreading impure if measured, which could also explain, at least in part, the aforementioned mixed findings concerning the influence of audiovisual semantic congruency (see the preceding paragraph). Therefore, in order to precisely quantify the stimulus-driven attentional spreading process and determine the effect of audiovisual emotional congruency on it, a novel experimental paradigm is needed.

The current study investigated whether and how audiovisual emotional congruency would modulate the stimulus-driven cross-sensory spread of attention by recording ERPs in a sound-accompanying visual oddball task where emotion information was task-irrelevant and the auditory modality was unattended as well. Specifically, although emotionally positive/negative visual-only and auditory-only stimuli, as well as emotionally congruent and incongruent audiovisual pairs, were presented in the task, participants were only required to detect the rarely presented blurred pictures while ignoring all auditory stimuli if delivered (Figure 1). Our results demonstrated that the isolating auditory Nd difference component within 200–300 ms post-stimulus was greater for emotionally incongruent than congruent audiovisual stimuli when their visual constituents were emotionally negative, thereby demonstrating for the first time that audiovisual emotional congruency can modulate the stimulus-driven visual-to-auditory attentional spreading at its early phase even when emotion is task-irrelevant.

Task paradigm

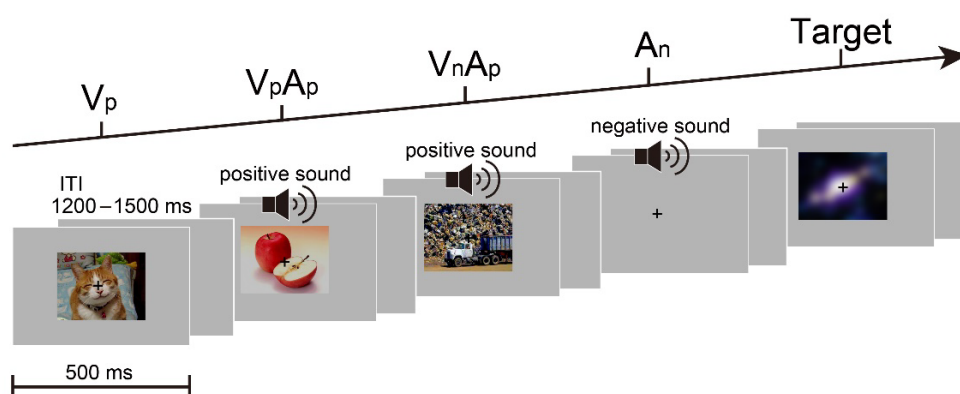


Figure 1. Task paradigm shown for four non-target trials and one target trial. A *non-target* trial could be an emotionally positive or negative visual or auditory stimulus presented alone [e.g., a positive picture (V_p) or a negative sound (A_n)], or the two stimuli presented synchronously (VA) to form an audiovisual pair being either emotionally congruent [e.g., a positive picture paired with a positive sound (V_pA_p)] or emotionally incongruent [e.g., a negative picture paired with a positive sound (V_nA_p)], resulting in eight sub-types of non-target stimuli. A *target* trial could be a blurred picture presented alone or presented synchronously with an emotionally positive or negative sound, resulting in three sub-types of target stimuli. The task for participants was to press a button in response to the target stimuli, while ignoring all sounds if delivered. Each type of trial consisted of a 500 ms stimulus presentation and an inter-trial interval (ITI) of 1200–1500 ms.

2. Materials and Methods

2.1. Subjects

Thirty volunteers participated in the assessment experiment (mean age 21.27 ± 1.29 years; 10 males and 20 females). In addition, we calculated the sample size for the formal experiment using MorePower 6.0.4 software (Saskatoon, Canada) [29]. Based on previous relevant studies, a priori sample size of 18 participants was derived by setting the parameters with $\alpha = 0.05$, $power = 0.80$ and $\eta^2_p = 0.119$ [22]. A total of 30 participants were recruited in the formal experiment. Three participants were excluded because of excessive artifacts in

EEG recordings. Data from the remaining 27 participants (mean age 21.94 ± 1.67 years; 8 males and 19 females) were included in future analysis. All participants reported normal or corrected-to-normal vision as well as normal hearing and no history of psychiatric or neurological disorders. They were naive as to the purpose of the experiment. Prior to the experiment procedures, all participants received informed consent, which was in agreement with the Declaration of Helsinki. After the experiment, participants were paid for their participation.

2.2. Assessment Experiment

The objective of the assessment experiment was to evaluate and standardize the emotional picture and sound materials that would be used in our formal experiment. A total of 60 pictures with 30 of positive valence and 30 of negative valence were selected through the Chinese Affective Picture System (CAPS) [30] adapted from the International Affective Picture System (IAPS) [31]. The brightness, saturation and coloration of each pixel point in the pictures were determined using the function *rgb2hsv* in Matlab. The software Adobe Photoshop CC 2019 was used to adjust the brightness, saturation and coloration of these pictures based on the average of all pixels in each picture, as recommended in the previous literature [32,33], so that each of the three parameters was comparable among pictures. All pictures were re-sized to a uniform size of 400×300 pixels. A total of 60 segments of voices with 30 of positive valence and 30 of negative valence were selected from the International Affective Digitized Sounds (IADS) [34] and then standardized to 500 ms in duration using the software Sound Engine, in order to meet the presentation requirement of the paradigm. The voice sampling rate was 48,000 Hz/s and all voices were tuned to a consistent volume of 65 dB SPL using the software MP3 Gain GUI. As the fundamental frequency of auditory stimuli could influence the ERP responses, a custom script in Matlab was used to measure the acoustic parameters of the auditory stimuli, including f_0 , pitch height, pitch range and pitch variance for each voice. Each acoustic parameter was compared between the emotionally positive and negative voices using independent-samples *t*-tests. The results showed that none of the acoustic parameters differed significantly between positive vs. negative voices [f_0 : $t_{(58)} = 1.21, p = 0.23, d = 0.31$; pitch height: $t_{(58)} = -0.93, p = 0.36, d = -0.24$; pitch range: $t_{(58)} = -1.16, p = 0.25, d = -0.30$; pitch variance: $t_{(58)} = 0.02, p = 0.98, d = 0.01$].

The program of the assessment experiment was scripted by Psychopy 3.0 (Python version 3.7, Nottingham, UK) [35] to present the emotional stimuli and record the participants' responses. On each trial, a fixation was first displayed in the center of the screen for 1000 ms, followed by a picture or sound presented for 2000 ms, and then the valence and arousal of this stimulus needed to be rated on a 9-point scale, respectively, using thumb-nails and the Self-Assessment Manikin (SAM) scale [36,37]. The pictures and voices were assessed separately in two sessions, with the order of precedence being counterbalanced between participants. The rating scores of valence and arousal were compared between the emotionally positive and negative stimuli using paired-samples *t*-tests, separately for pictures and voices. The results showed that there were significant valence differences between positive vs. negative pictures [$t_{(29)} = 34.11, p < 0.0001, d = 6.23$] and between positive vs. negative voices [$t_{(29)} = 19.05, p < 0.0001, d = 3.48$], such that the rated valence scores of positive stimuli [pictures: 6.74 ± 0.09 ($M \pm SE$); voices: 6.50 ± 0.11] were higher than those of negative stimuli (pictures: 2.93 ± 0.08 ; voices: 3.59 ± 0.08). In contrast, neither pictures [$t_{(29)} = -2.02, p = 0.053, d = -0.37$] nor voices [$t_{(29)} = -0.36, p = 0.72, d = -0.07$] showed a significant difference between the arousal scores of positive emotion (pictures: 5.78 ± 0.11 ; voices: 4.94 ± 0.20) and negative emotion (pictures: 5.99 ± 0.09 ; voices: 5.02 ± 0.11).

2.3. Experimental Materials and Procedures

In the formal experiment, participants sat in a dark and sound-attenuated room with a viewing distance of approximately 80 cm away from a 27-inch LCD monitor (ASUS PG279Q, $1920 \times 1080, 120$ Hz) where visual stimuli were presented. Two speakers (HiVi X3) were

located on either side of the monitor at an equal height parallel to the center of the monitor screen for sound presentation, so that the sound played from both speakers simultaneously would be perceived as coming from the center of the monitor [38]. During the experiment, the monitor screen remained gray (RGB: 128, 128, 128) at all times and participants were asked to maintain their eyes fixated on a black cross (RGB: 0, 0, 0; $0.3^\circ \times 0.3^\circ$ in size) at the center of the screen. “Presentation” software (version 18.0, NeuroBehavioral Systems, Inc., Berkeley, CA, USA) was used to display all visual and auditory stimuli.

For the purpose of isolating the stimulus-driven cross-modal spread of attention effect when analyzing EEG data [13,14], two main types of trials, namely, non-target trial and target trial, were designed in the formal experiment. A *non-target* trial could be an emotional visual or auditory stimulus presented alone, or could be the two stimuli presented synchronously. The emotional visual stimulus could be one of the sixty emotional pictures selected from the assessment experiment (thirty positive pictures and thirty negative pictures; each $9.5^\circ \times 7.2^\circ$ in size) equiprobably, which was presented for 500 ms at the center of the monitor (Figure 1). The emotional auditory stimulus that was also centrally presented could be one of the sixty emotional sounds (thirty positive sounds and thirty negative sounds) with equal probability, which was also 500 ms in duration (with 10 ms rise and fall periods) and was approximately 65 dB SPL at participants’ ears. These emotional pictures and sounds were presented either alone or synchronously, resulting in three main stimulus types for non-target trials [i.e., visual alone (labeled as V condition), auditory alone (A condition), and audiovisual (labeled as VA condition)]. Of note, depending on the emotional valence, there were two sub-types for V and A conditions, respectively, namely, positive visual-alone (labeled as V_p), negative visual-alone (labeled as V_n), positive auditory-alone (labeled as A_p), and negative auditory-alone (labeled as A_n). Accordingly, for VA condition, when both the visual and auditory constituents of an audiovisual pair were the same emotional valence (i.e., emotionally congruent), there were two sub-types, namely, a positive picture paired with a positive sound (labeled as V_pA_p) and a negative picture paired with a negative sound (labeled as V_nA_n). Similarly, when the visual and auditory constituents of an audiovisual pair were different in emotional valence (i.e., emotionally incongruent), there were also two sub-types, namely, a positive picture paired with a negative sound (labeled as V_pA_n) and a negative picture paired with a positive sound (labeled as V_nA_p).

Apart from the aforementioned eight sub-types of non-target trials, there were also three sub-types of the *target* trials. Specifically, a target trial could be a blurred picture presented alone (labeled as T condition), or could be a blurred picture presented synchronously with either an emotionally positive sound (labeled as TA_p condition) or an emotionally negative sound (labeled as TA_n condition). On a given target trial, the blurred picture could be one of the above-mentioned 60 emotional pictures with equal probability, but a Gaussian blur with a radius of 22.5 pixel was applied to it in order to render its emotional valence unrecognizable. Taken together, the eight sub-types of non-target stimuli and the three sub-types of target stimuli accounted for a total of 91.67% of the trials, with 8.33% for each sub-type. The remaining 8.33% of the trials were “blank” trials on which neither visual nor auditory stimuli were presented (labeled as B condition), serving as an estimation of anticipatory ERPs elicited by the expectation of upcoming stimulus (for details, see Data analysis section). The aforementioned 12 types of trials were presented in a pseudo-randomized order with an inter-trial interval (ITI) varying from 1200 to 1500 ms randomly (Figure 1). The task for participants was to press the button “J” on a keyboard with their right index finger whenever they detected a blurred picture (i.e., a target trial) while ignoring all auditory stimuli. Therefore, the novel task design here meant *not only* that the auditory stimuli were task-irrelevant (i.e., only the visual stimuli were to-be-attended), *but also* that the emotional valences of both visual and auditory stimuli were task-irrelevant. The whole experiment comprised a total of 1800 trials, which were divided into 25 blocks to complete. The duration of the whole experiment was around 90 min, and a 10-sec rest time was imposed between blocks, after which participants could continue to rest or start the next block.

2.4. Electrophysiological Recording and Preprocessing

Electroencephalographic (EEG) signals were continuously recorded with a NeuroScan SynAmp amplifier (NeuroScan, Inc., El Paso, TX, USA) and a custom-built 64-electrode elastic cap on which the electrodes were positioned in accordance with a modified 10-10 system montage (for details, see [39]). Two additional electrodes, AFz and M1 (left mastoid), served as the ground and reference electrodes, respectively, for online EEG recording. The horizontal electrooculogram (HEOG) was recorded using bipolar electrodes placed on the left and right outer canthi. To monitor blinks and vertical eye movements (vertical electrooculogram, VEOG), bipolar electrodes were placed above and below the participants' left eye. The impedance of all electrodes was maintained below 5 k Ω . The digital sampling rate was 1000 Hz, and a band-pass filtering of 0.05–100 Hz was applied to the online instantaneous EEG data. All EEG and EOG data were recorded via Scan software (version 4.5, NeuroScan, Inc., El Paso, TX, USA).

For the offline preprocessing, the raw continuous EEG data were firstly down-sampled to 500 Hz and then low-pass filtered (half-amplitude cutoff = 33.75 Hz, transition band width = 7.5 Hz) to attenuate high-frequency noise arising from muscle activity and external electrical sources. The arithmetic mean of the bilateral mastoids (M1, M2) served as the reference for the filtered data. The re-referenced data were segmented into 800-ms epochs, which were time-locked to the onset of emotional stimulus with a 200-ms pre-stimulus baseline correction. The epochs contaminated by eye movements, eye blinks and muscle activity were then removed by automatic artifact rejection based on a threshold of $\pm 75 \mu\text{V}$ for both EEG and EOG electrodes. In addition, in order to prevent the interference of motor responses with EEG data, all target trials (i.e., trials on which the blurred pictures were presented) and all false alarm trials were further removed. The remaining artifact-free EEG epochs were averaged separately for each experimental condition (i.e., V_p , V_n , A_p , A_n , V_pA_p , V_pA_n , V_nA_n , V_nA_p and B). EEG preprocessing was performed using the EEGLAB toolbox [40] and a custom script in Matlab, and subsequent ERP analysis was performed in ERPLAB [41].

2.5. Data Analysis

According to the previous literature on the method of isolating the stimulus-driven spread of attention [13,14], firstly, the auditory ERPs in the context of attended visual stimuli were extracted by subtracting ERPs to the unisensory visual stimuli from ERPs to the audiovisual stimuli, separately for each audiovisual emotional combination (i.e., $V_pA_p - V_p$; $V_pA_n - V_p$; $V_nA_n - V_n$; $V_nA_p - V_n$). The resulting difference waves consisted of the contributions from not only the auditory constituents of the audiovisual stimuli but also the potential cross-modal attentional spreading. Secondly, the time-locked ERPs recorded on the blank trials were subtracted from ERPs elicited by the unisensory auditory stimuli alone, separately for emotionally positive and negative sounds (i.e., $A_p - B$; $A_n - B$), in order to cancel out any pre-stimulus anticipatory activities (e.g., CNV) [42] common to all stimuli. Otherwise, these common activities would be balanced out in the extracted auditory ERPs to audiovisual stimuli but left in the ERPs elicited by auditory stimuli alone. In other words, the ERPs recorded on blank trials were used as an estimation of the pre-stimulus anticipatory ERPs [7]. Finally, the extracted auditory ERPs to audiovisual stimuli were compared with the unisensory auditory ERPs based on the auditory emotional valence [i.e., ($V_pA_p - V_p$) vs. ($A_p - B$); ($V_pA_n - V_p$) vs. ($A_n - B$); ($V_nA_n - V_n$) vs. ($A_n - B$); ($V_nA_p - V_n$) vs. ($A_p - B$)], and the differences revealed in these comparisons thus represented the stimulus-driven spread of attention effects under different audiovisual emotional combinations.

The stimulus-driven spread of attention effect was further quantified by the mean amplitude of the auditory negative difference (Nd) component, which was measured with two 100-ms time windows during 200–400 ms after the onset of emotional stimuli over six adjacent fronto-central electrodes (FC1, FCz, FC2, C1, Cz, C2). These time windows and electrodes were selected because the stimulus-driven Nd amplitude is typically maximal

over there [10,13,14,21,22]. In addition, since several prior studies [24,25,28] even found an audiovisual emotional congruency effect during the auditory N1 interval (although the auditory modality was not task-irrelevant), we speculated that the stimulus-driven spread of attention in response to emotional audiovisual stimuli may occur at earlier stages of processing than that to non-emotional audiovisual stimuli. Indeed, based on visual inspection, we found that there may be a difference between the extracted auditory ERPs to audiovisual stimuli and ERPs to the unisensory auditory stimuli during the time window of the auditory N1 component. Accordingly, we also analyzed the auditory N1 component, whose mean amplitude was measured during 90–130 ms over six adjacent fronto-central electrodes (FC1, FCz, FC2, C1, Cz, C2), where its amplitude was greatest when ERP waveforms were collapsed across all conditions to be compared.

For statistical analyses, to explore whether the stimulus-driven spread of attention (i.e., the auditory Nd difference component) occurred significantly under all audiovisual emotional combinations, we conducted paired-samples *t*-tests on the mean amplitudes during the two Nd intervals between the extracted auditory ERPs to audiovisual stimuli vs. the ERPs to unisensory auditory stimuli separately for each audiovisual emotional combination [i.e., for congruent audiovisual pairs with emotionally positive visual and auditory constituents: $(V_pA_p - V_p)$ vs. $(A_p - B)$; for incongruent audiovisual pairs with emotionally positive visual and negative auditory constituents: $(V_pA_n - V_p)$ vs. $(A_n - B)$; for congruent audiovisual pairs with emotionally negative visual and auditory constituents: $(V_nA_n - V_n)$ vs. $(A_n - B)$; for incongruent audiovisual pairs with emotionally negative visual and positive auditory constituents: $(V_nA_p - V_n)$ vs. $(A_p - B)$]. Note that we did not conduct a multi-factor repeated-measures ANOVA prior to these paired *t*-test (the same below), because these *t*-tests alone are sufficient to answer the research question above and are more straightforward, which can reduce the total number of statistical tests conducted, thereby controlling the overall Type I error rate [43]. Moreover, to verify whether the attentional spreading occurred earlier, similar paired-samples *t*-tests were conducted on the auditory N1 amplitude. In order to further examine whether the magnitude of the cross-modal attentional spreading would be modulated by audiovisual emotional congruency, additional paired-samples *t*-tests were performed on the attentional spreading effects (measured as the extracted auditory minus auditory-only ERP differences) between emotionally congruent vs. incongruent audiovisual pairs. These congruent vs. incongruent contrasts were conducted only within the time window wherein the attentional spreading effect was significant under all audiovisual pairs, and were performed separately for: (1) audiovisual pairs with positive visual constituents [congruent: $(V_pA_p - V_p) - (A_p - B)$ vs. incongruent: $(V_pA_n - V_p) - (A_n - B)$] and negative visual constituents [congruent: $(V_nA_n - V_n) - (A_n - B)$ vs. incongruent: $(V_nA_p - V_n) - (A_p - B)$]; and for: (2) audiovisual pairs with positive auditory constituents [congruent: $(V_pA_p - V_p) - (A_p - B)$ vs. incongruent: $(V_nA_p - V_n) - (A_p - B)$] and negative auditory constituents [congruent: $(V_nA_n - V_n) - (A_n - B)$ vs. incongruent: $(V_pA_n - V_p) - (A_n - B)$].

In addition, to verify our conjecture that emotionally positive stimuli would capture more attention than negative stimuli when emotion is task-irrelevant (for details, see Discussion section), we further compared the visual N1 component elicited by emotionally positive vs. negative visual stimuli (i.e., V_p vs. V_n) using a paired-samples *t*-test. The visual N1 component was measured as mean amplitude within the time window of 145–175 ms over two bilaterally occipital electrodes (PO7, PO8), where its negative-going amplitude was greatest when ERP waveforms were collapsed across the two aforementioned unisensory visual conditions.

Based on the results of the traditional statistical method mentioned above, we also expected to explore the additional potential factors (e.g., subject characteristics) to explain more error variance. Hence, the mixed effects models with subjects being entered as a random effect factor were conducted for exploratory analysis (for details, see Supplementary Materials).

3. Results

3.1. Behavior Results

A one-way repeated-measures ANOVA was performed for response times (RTs) and hit rates (HRs), separately, with the factor of target type [TA_p (visual targets accompanied by emotionally positive sounds), TA_n (visual targets accompanied by negative sounds), T (visual targets alone)]. For both RTs and hit rates, there was no significant difference among different target types [RTs: TA_p, 458.25 ± 11.08 ms (*M* ± *SE*); TA_n, 459.71 ± 10.01 ms; T, 458.53 ± 9.64 ms; $F_{(2, 52)} = 0.03$, $p = 0.95$, $\eta^2_p = 0.001$; HRs: TA_p, 99.43 ± 0.16%; TA_n, 97.91 ± 0.85%; T, 99.19 ± 0.31%; $F_{(2, 52)} = 3.57$, $p = 0.063$ (Greenhouse–Geisser corrected due to violation of the sphericity assumption), $\eta^2_p = 0.12$]. These behavioral results indicated that the emotional sounds did not interfere substantially with the detection of the target, suggesting that the auditory stimuli were ignored as required.

3.2. EEG Results

3.2.1. The Stimulus-Driven Spread of Attention Is Modulated by Audiovisual Emotional Congruency

To investigate whether the non-target emotional stimuli elicited the stimulus-driven spread of attention as well as its time course, paired-samples *t*-tests were conducted on the mean amplitudes during each Nd interval (200–300 ms, 300–400 ms) between the extracted auditory ERPs to audiovisual stimuli (VA – V) vs. the ERPs to auditory-only stimuli (A – B) separately for each audiovisual emotional combination. A significant difference would reveal that the auditory Nd component was prominent and the stimulus-driven attentional spreading occurred reliably. The results showed that in the time window of 200–300 ms, the auditory Nd component was prominent in response to all audiovisual emotional combinations. Specifically, for the audiovisual pairs with emotionally positive auditory constituents (Figure 2a), both the extracted auditory ERPs to emotionally congruent audiovisual stimuli [$V_pA_p - V_p$: $-1.29 \pm 0.53 \mu\text{V}$ (*M* ± *SE*)] and the extracted auditory ERPs to incongruent audiovisual stimuli ($V_nA_p - V_n$: $-1.50 \pm 0.51 \mu\text{V}$) were significantly more negative than the ERPs to auditory-only stimuli [$A_p - B$: $1.05 \pm 0.46 \mu\text{V}$; $t_{(26)} = -3.50$, $p = 0.002$, $d = -0.67$; $t_{(26)} = -3.87$, $p < 0.001$, $d = -0.75$]. Similarly, for the audiovisual pairs with emotionally negative auditory constituents (Figure 2b), both the extracted auditory ERPs to congruent audiovisual stimuli ($V_nA_n - V_n$: $0.01 \pm 0.54 \mu\text{V}$) and the extracted auditory ERPs to incongruent audiovisual stimuli ($V_pA_n - V_p$: $-0.25 \pm 0.53 \mu\text{V}$) were significantly more negative than the ERPs to auditory-only stimuli [$A_n - B$: $1.82 \pm 0.41 \mu\text{V}$; $t_{(26)} = -2.78$, $p = 0.01$, $d = -0.53$; $t_{(26)} = -2.92$, $p = 0.007$, $d = -0.55$]. These results indicate that the stimulus-driven spread of attention effect occurred regardless of whether the audiovisual pairs were emotionally congruent or incongruent during the time window of 200–300 ms.

In contrast, in the time window of 300–400 ms, the extracted auditory ERPs to audiovisual stimuli were found to be significantly more negative than the ERPs to auditory-only stimuli only for the incongruent audiovisual pairs with emotionally positive auditory constituents [$V_nA_p - V_n$: $-2.79 \pm 0.44 \mu\text{V}$; $A_p - B$: $-1.90 \pm 0.38 \mu\text{V}$; $t_{(26)} = -2.13$, $p = 0.043$, $d = -0.41$; Figure 2a], but not for the congruent audiovisual pairs with emotionally positive auditory constituents [$V_pA_p - V_p$: $-2.39 \pm 0.52 \mu\text{V}$; $A_p - B$: $-1.90 \pm 0.38 \mu\text{V}$; $t_{(26)} = -1.12$, $p = 0.27$, $d = -0.22$], or the incongruent audiovisual pairs with emotionally negative auditory constituents [$V_pA_n - V_p$: $-1.13 \pm 0.46 \mu\text{V}$; $A_n - B$: $-0.72 \pm 0.41 \mu\text{V}$; $t_{(26)} = -0.78$, $p = 0.44$, $d = -0.15$; Figure 2b], or the congruent audiovisual pairs with emotionally negative auditory constituents [$V_nA_n - V_n$: $-1.05 \pm 0.46 \mu\text{V}$; $A_n - B$: $-0.72 \pm 0.41 \mu\text{V}$; $t_{(26)} = -0.60$, $p = 0.56$, $d = -0.12$]. These results suggest that there might be two prerequisites for the cross-modal attentional spreading to be sustained into the 300–400 ms time window in response to emotional audiovisual stimuli, one being that the auditory constituents of audiovisual stimuli are emotionally positive (i.e., the visual constituents are emotionally negative) and the other being that there is an emotional conflict between the auditory and visual constituents. Therefore, the late phase of the stimulus-driven attentional spreading to emotional sounds can be modulated by audiovisual emotional congruency.

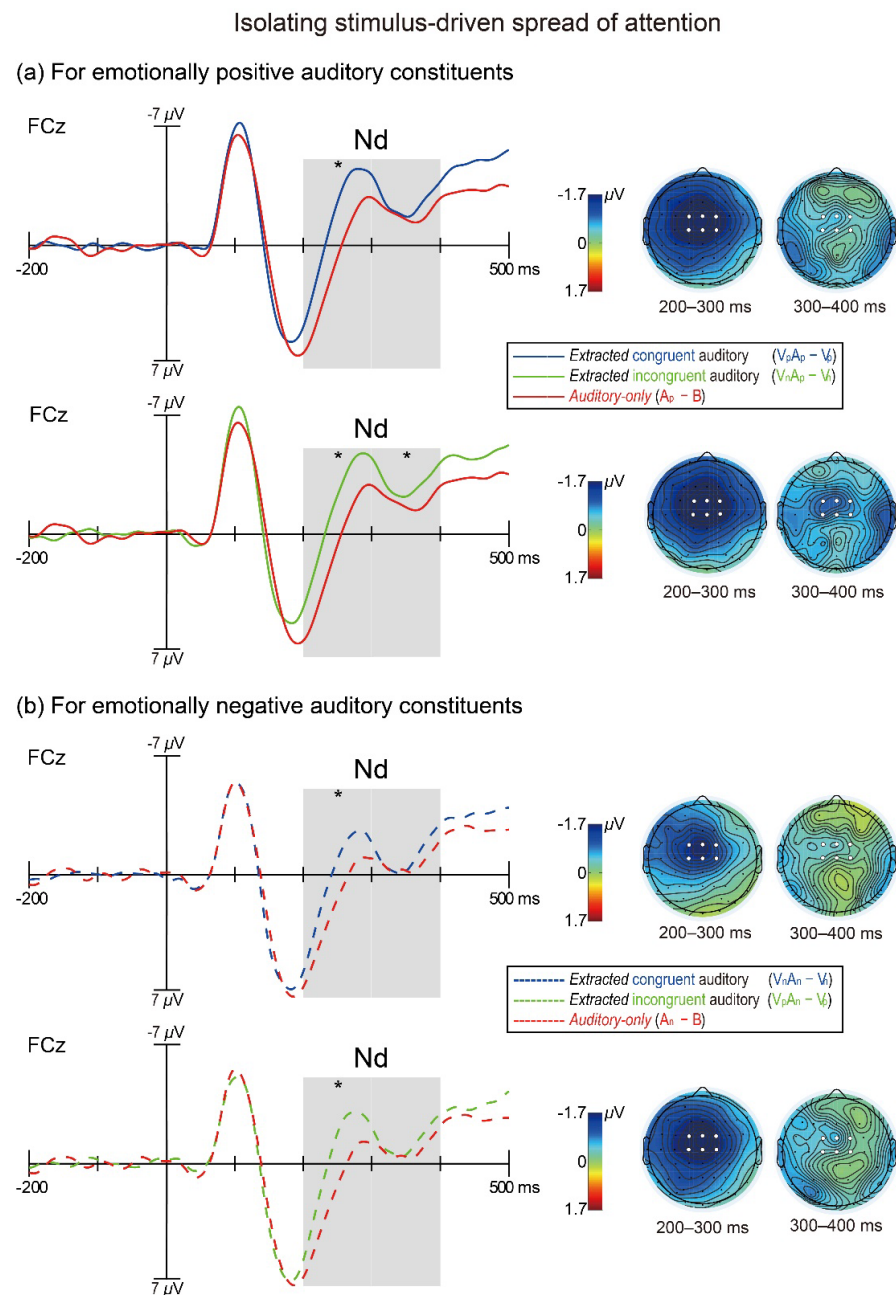


Figure 2. The extracted auditory ERP waveforms to emotionally congruent (blue traces) and incongruent (green traces) nontarget audiovisual stimuli and ERP waveforms evoked by auditory-only nontarget stimuli (red traces), plotted separately for (a) emotionally positive auditory constituents (solid traces) and (b) emotionally negative auditory constituents (dashed traces). These ERP waveforms are from the fronto-central electrode FCz and the shaded areas on waveforms depict the two-time windows (200–300 ms and 300–400 ms) within which the Nd component was measured. Scalp topographies are shown for the extracted-auditory minus auditory-only mean difference amplitudes during each Nd interval. The white dots on topographies depict the fronto-central ROI (FC1, FCz, FC2, C1, Cz, C2) over which the Nd component was measured. The stimulus-driven Nd component, indexed by significantly greater negative amplitude in the extracted auditory than auditory-only waveform, was prominent for all audiovisual emotional combinations within the time window of 200–300 ms. In contrast, only the incongruent audiovisual pairs with emotionally positive auditory and negative visual constituents (V_nA_p) produced a sustained Nd within the time window of 300–400 ms. *: $p < 0.05$ for the extracted-auditory vs. auditory-only contrast.

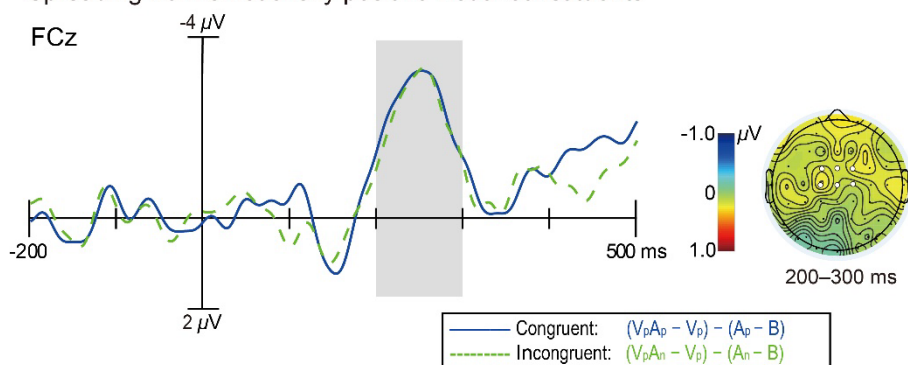
Notably, although the early phase of the stimulus-driven attentional spreading (i.e., 200–300 ms) could occur regardless of audiovisual emotional congruency, it was still unclear whether the magnitude of the early-phase attentional spreading would be modulated by audiovisual emotional congruency. To examine this question in detail, additional paired-samples *t*-tests were performed on the attentional spreading effects (measured as the extracted auditory minus auditory-only ERP differences) during 200–300 ms between emotionally congruent vs. incongruent audiovisual pairs in the following two ways. Firstly, when these comparisons were anchored to the visual constituents' emotional valence, the attentional spreading effect for emotionally incongruent audiovisual stimuli $[(V_nA_p - V_n) - (A_p - B): -2.13 \pm 0.43 \mu\text{V}]$ was found to be significantly greater than that for congruent audiovisual stimuli $[(V_nA_n - V_n) - (A_n - B): -1.40 \pm 0.51 \mu\text{V}]$ only when the visual constituents were emotionally negative [$t_{(26)} = -2.33, p = 0.028, d = -0.45$; Figure 3, lower half], but not when the visual constituents were emotionally positive [incongruent: $(V_pA_n - V_p) - (A_n - B) = -1.89 \pm 0.51 \mu\text{V}$; congruent: $(V_pA_p - V_p) - (A_p - B) = -2.08 \pm 0.44 \mu\text{V}$; $t_{(26)} = 0.64, p = 0.53, d = 0.12$; Figure 3, upper half]. Secondly, when similar incongruent vs. congruent comparisons were anchored to the auditory constituents' emotional valence, however, we found no significant difference either when the auditory constituents were emotionally positive [incongruent: $(V_nA_p - V_n) - (A_p - B) = -2.13 \pm 0.43 \mu\text{V}$; congruent: $(V_pA_p - V_p) - (A_p - B) = -2.08 \pm 0.44 \mu\text{V}$; $t_{(26)} = -0.20, p = 0.84, d = -0.04$] or when they were emotionally negative [incongruent: $(V_pA_n - V_p) - (A_n - B) = -1.89 \pm 0.51 \mu\text{V}$; congruent: $(V_nA_n - V_n) - (A_n - B) = -1.40 \pm 0.51 \mu\text{V}$; $t_{(26)} = -1.45, p = 0.16, d = -0.28$]. Taken together, these findings demonstrate that audiovisual emotional congruency can modulate the early-phase stimulus-driven attentional spreading if the audiovisual pairs' visual constituents are emotionally negative, which echoes the aforementioned finding that the late-phase (300–400 ms) attentional spreading occurred only for the incongruent pairs with emotionally negative visual constituents (Figure 2a).

3.2.2. Post Hoc Exploratory Analyses

Further visual inspection of Figure 2a implies that when the audiovisual pairs' auditory constituents were emotionally positive, the extracted auditory ERPs to both congruent and incongruent audiovisual stimuli seem larger than the ERPs to unisensory auditory stimuli during the time window of auditory N1 component, which may reflect the cross-modal attentional spreading occurring in advance for emotional audiovisual stimuli. To explore this possibility, we performed paired-samples *t*-tests on the auditory N1 amplitude (measured over 90–130 ms) between the extracted auditory ERPs to audiovisual stimuli vs. the ERPs to auditory-only stimuli separately for each of the four audiovisual emotional combinations. However, none of the four *t*-tests' results reached statistical significance [$V_pA_p - V_p$: $-5.75 \pm 0.57 \mu\text{V}$; $V_nA_p - V_n$: $-5.84 \pm 0.62 \mu\text{V}$; $A_p - B$: $-5.41 \pm 0.62 \mu\text{V}$; $(V_pA_p - V_p)$ vs. $(A_p - B)$: $t_{(26)} = -0.92, p = 0.37, d = -0.18$; $(V_nA_p - V_n)$ vs. $(A_p - B)$: $t_{(26)} = -1.28, p = 0.21, d = -0.25$; c.f., Figure 2a; $V_nA_n - V_n$: $-4.08 \pm 0.53 \mu\text{V}$; $V_pA_n - V_p$: $-3.71 \pm 0.51 \mu\text{V}$; $A_n - B$: $-4.14 \pm 0.54 \mu\text{V}$; $(V_nA_n - V_n)$ vs. $(A_n - B)$: $t_{(26)} = 0.23, p = 0.82, d = 0.04$; $(V_pA_n - V_p)$ vs. $(A_n - B)$: $t_{(26)} = 1.27, p = 0.21, d = 0.25$; c.f., Figure 2b]. Therefore, there is no substantial evidence to propose that the stimulus-driven cross-modal spread of attention occurs earlier for emotional than non-emotional of audiovisual stimuli.

Emotional congruency effect on stimulus-driven spread of attention

Spreading from emotionally positive visual constituents



Spreading from emotionally negative visual constituents

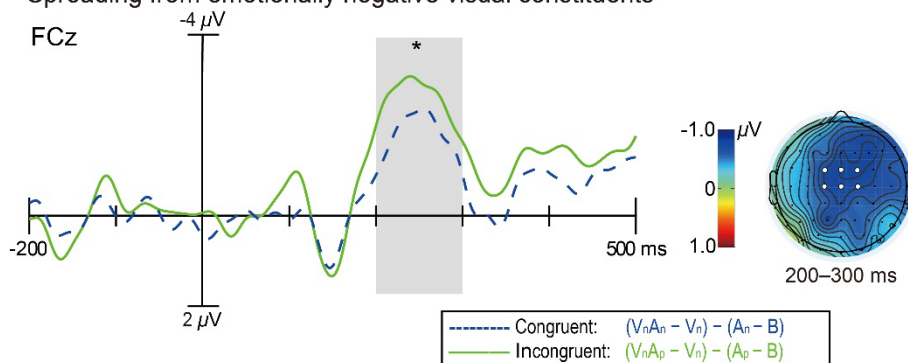


Figure 3. The *magnitude* of the stimulus-driven auditory Nd wave (measured as the extracted-auditory minus auditory-only ERP difference) in response to emotionally congruent (blue traces) and incongruent (green traces) audiovisual stimuli, plotted separately for audiovisual stimuli with emotionally positive (upper half) and negative (lower half) visual constituents. The shaded areas on waveforms depict the time window of 200–300 ms within which the Nd magnitudes were further contrasted. Scalp topographies are shown for incongruent minus congruent Nd magnitude differences during the time window of 200–300 ms. The Nd magnitude was significantly larger for emotionally incongruent than congruent audiovisual stimuli only when their visual constituents' emotional valence was negative. *: $p < 0.05$ for the incongruent vs. congruent contrast.

Finally, to validate our assumption that emotionally positive stimuli would capture more attention than negative stimuli when emotion is task-irrelevant (for details, see Discussion section), we conducted a paired-samples t -test on the visual N1 amplitude (measured within 145–175 ms over electrodes PO7 and PO8) between emotionally positive and negative unisensory visual stimuli that were nontargets but spatially attended. The result showed that the N1 component evoked by positive visual stimuli (V_p : $4.17 \pm 3.91 \mu V$) was more negative-going in amplitude than that evoked by negative visual stimuli [V_n : $4.76 \pm 3.80 \mu V$; $t_{(26)} = -2.23$, $p = 0.035$, $d = -0.43$; see Figure 4]. This finding implies that when the emotions of attended visual stimuli are irrelevant to the current task, the emotionally positive ones would capture more attention than the negative ones, thereby providing evidence for the assumption mentioned above.

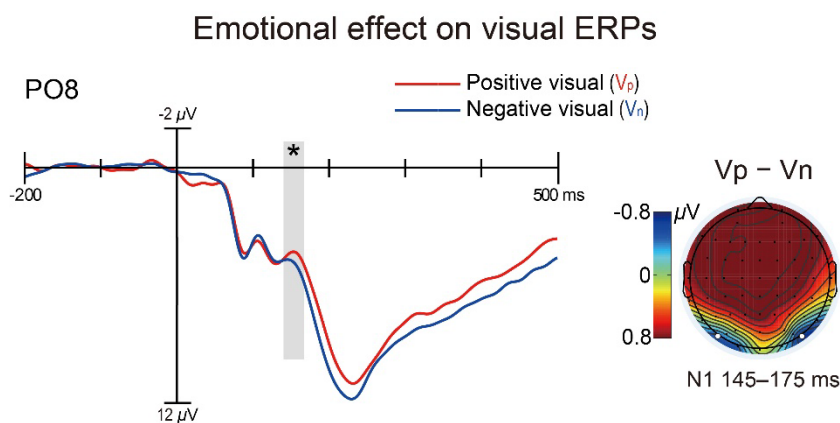


Figure 4. ERP waveforms elicited by emotionally positive visual-only (blue lines: V_p) and negative visual-only (green lines: V_n) stimuli, exemplified from the right occipital electrode PO8. The shaded area on waveforms depicts the time window of 145–175 ms within which the visual N1 component was measured. The scalp topography is shown for positive visual-only minus negative visual-only mean difference amplitudes during the visual N1 interval. The amplitude of N1 was larger for positive visual-only than negative visual-only stimuli. *: $p < 0.05$ for the contrast.

4. Discussion

The current study utilized the high time-resolution ERP technique to explore whether the stimulus-driven attentional spreading from an audiovisual pair's attended visual constituent to its unattended auditory constituent [10] would be affected by high-level emotional congruency between the visual and auditory constituents. In order to render the visual constituents attended and the auditory constituents unattended, while keeping emotion information carried by all stimuli task-irrelevant, we required participants to only detect the rarely presented blurred pictures (targets) under the premise of ignoring all sounds, although emotionally positive/negative visual-only and auditory-only stimuli, as well as emotionally congruent and incongruent audiovisual pairs, were presented (as nontargets) in the task. Our behavioral data showed that the target detection performance, quantified by both RTs and hit rates, was neither improved nor impaired when the blurred pictures were paired with emotionally positive sounds (TA_p) or negative sounds (TA_n), relative to when the blurred pictures were presented alone (T). The absence of previously reported behavioral modulations of emotional sounds [44] suggest that our participants ignored the task-irrelevant auditory inputs as well as emotion information to a high degree as required.

Our electrophysiological data first yielded that the auditory Nd component, indexed by significantly greater negative amplitude in the extracted auditory ERPs to emotional audiovisual stimuli (e.g., $V_nA_p - V_n$) than in the ERPs to emotional auditory-only stimuli (e.g., $A_p - B$), was prominent for all audiovisual emotional combinations (i.e., V_pA_p , V_nA_n , V_pA_n and V_nA_p) within the time window of 200–300 ms. These findings indicate that the occurrence of stimulus-driven visual-to-auditory attentional spreading is independent of audiovisual emotional congruency, confirming its bottom-up nature as proposed in previous studies [13,14,18,20]. However, the more important finding is that the amplitude of Nd component within the 200–300 ms interval, measured as the extracted-auditory minus auditory-only ERP difference, was significantly greater in response to emotionally incongruent than congruent audiovisual pairs when their visual constituents were emotionally negative. Furthermore, the Nd component was found to extend into 300–400 ms only in response to the incongruent audiovisual stimuli with emotionally negative visual constituents. Given that the task-irrelevance of high-level emotional representations in the current paradigm avoided the potential top-down contamination when quantifying the bottom-up, stimulus-driven Nd component, the findings above provide strong and convergent evidence that audiovisual emotional congruency does have a substantial influence

on the *magnitude* of stimulus-driven cross-modal spread of attention, beginning in parallel with its occurrence.

It is noteworthy that the auditory Nd amplitude was larger for emotionally incongruent than congruent audiovisual pairs only when their visual constituents were emotionally negative (i.e., V_nA_p vs. V_nA_n) but not when their visual constituents were emotionally positive (i.e., V_pA_n vs. V_pA_p), indicating that the audiovisual emotional congruency effect is specific to the attentional spreading from visual constituents conveying negative emotions. One might interpret these findings in terms of the well-known “negativity bias” that emotionally negative stimuli, given its superior biological significance, can attract more attentional resources than emotionally positive and neutral stimuli at early stages of processing, which typically leads to greater P1 and/or N1 components in the visual domain [45–49]. In the current visual oddball task, this interpretation would propose that the visual elements of audiovisual pairs captured more attention when these visual elements were emotionally negative than positive, leading to the task-irrelevant auditory elements of the former being bestowed with even more enhanced processing as attention spread across modality. Consequently, the emotional conflicts in incongruent audiovisual pairs with negative visual elements might be detected more sufficiently, hence the observed “negative-visual-specific” emotional congruency effect. However, it should be noted that the basic assumption of this plausible interpretation does not hold in the current study, because our post hoc analysis has shown that the visual N1 amplitude was actually *smaller*, instead of larger, in response to negative than positive visual-only stimuli (see Figure 4), which suggests that positive pictures captured more attention than negative pictures in the current study, not the other way around.

In fact, several prior studies have also observed larger N1 amplitudes elicited by emotionally positive than negative visual stimuli [50,51], and a common feature in these studies is that the emotional characteristics of stimuli were *irrelevant* to their participants’ tasks, in contrast to those studies showing the negativity bias wherein emotions were typically *task-relevant* [45,46,48,49] (but see [47]). Since the emotional characteristics of visual stimuli were also task-irrelevant in the current study, it is possible that early attentional resources were allocated more to the current positive than negative visual stimuli (and visual constituents of audiovisual pairs). If that is the case, we should further predict that the attentional spreading from positive visual constituents was stronger than that from negative visual constituents. Indeed, this prediction is supported, at least in part, by another post hoc test yielding that the isolated auditory Nd amplitude tended to be larger for congruent audiovisual pairs with emotionally positive visual constituents than congruent pairs with negative visual constituents [$t_{(26)} = -1.80, p = 0.083, d = -0.35$; see Figure 3, blue solid and dashed traces]. After substantiating the existence of “positivity bias” in the current study, we proposed that for audiovisual pairs with emotionally negative visual constituents, as visual attention spread across modality to auditory constituents anyway, the incongruent (i.e., V_nA_p) pairs’ positive auditory constituents may attract further attention than the congruent (i.e., V_nA_n) pairs’ negative auditory constituents, thereby intensifying the implicit conflict processing of the incongruent pairs and leading to the significant audiovisual emotional congruency effect in this case. In contrast, for audiovisual pairs with emotionally positive visual constituents, the incongruent (i.e., V_pA_n) pairs’ negative auditory constituents may not capture extra attention relative to the congruent (i.e., V_pA_p) pairs’ positive auditory constituents, which could render the conflict processing of these incongruent pairs to a limited extent, resulting in the observed null effect of audiovisual emotional congruency in that case. Although the interpretation above is relatively tentative, it highlights the subtle interplay between audiovisual emotional congruency and certain emotional combinations in modulating the stimulus-driven attentional spreading, which is consistent with many previous EEG investigations showing that the effects of audiovisual emotional congruency were specific to certain emotional combinations [24–27].

It should also be noted that since we explored the audiovisual emotional congruency effect separately for bimodal pairs with positive (i.e., V_pA_n vs. V_pA_p) and negative (i.e.,

V_nA_p vs. V_nA_n) visual constituents, it is *inevitable* that the emotionally congruent and incongruent pairs differed not only in the degree of congruency, but also in overall valence. For example, for audiovisual pairs with negative visual constituents, the overall valence of the incongruent pairs (V_nA_p) was certainly higher than that of the congruent pairs (V_nA_n). Accordingly, one may argue that the current larger Nd amplitude to the incongruent than congruent pairs with negative visual constituents (Figure 3, lower half) was simply due to the overall valence of the incongruent pairs V_nA_p being higher than that of the congruent pairs V_nA_n , rather than the implicit conflict processing of the incongruent pairs. However, note that for audiovisual pairs with positive visual constituents, the overall valence of the incongruent pairs (V_pA_n) was definitely *lower* than that of the congruent pairs (V_pA_p). Therefore, had this high-valence hypothesis alone held, the Nd amplitude to the incongruent pairs V_pA_n should have been smaller than that to the congruent pairs V_pA_p , but was not (Figure 3, upper half). Based on the ratiocination, we argued that although the role of the audiovisual emotional pairs' overall valence in the congruency effect on stimulus-driven attentional spreading cannot be ruled out in the current study, the implicit conflict processing of the incongruent pairs did contribute substantially to the congruency effect on attentional spreading. Nevertheless, additional research with improvements in the experimental paradigm is strongly required to tease apart the influences of emotional conflict and valence when examining the audiovisual emotional congruency effect.

The current audiovisual *emotional* congruency effect started approximately 200 ms post-stimulus, in parallel with the emergence of stimulus-driven attentional spreading. In contrast, the audiovisual *semantic* congruency effect reported in previous research on stimulus-driven attentional spreading did not begin until around 300 ms post-stimulus [18] (but see [21] for a null result). This discrepancy suggests that the emotional conflict between visual and auditory elements can be processed more rapidly than the semantic conflict between emotionally neutral visual and auditory elements. Indeed, previous electrophysiological studies have revealed that over the fronto-central scalp (i.e., auditory ROI), the timing of audiovisual emotional congruency effect (typically prior to 200 ms [24–28]) was earlier than that of audiovisual semantic congruency effect (starting ~250 ms at the earliest [52]) even when participants needed to actively evaluate the emotional/semantic characteristics of stimuli. However, although some of these studies [24,25,28] even reported a more rapid audiovisual emotional congruency effect during the auditory N1 interval, we did not find its counterpart when measuring the cross-modal attentional spreading, as indexed by the absence of significant difference between the extracted-auditory vs. auditory-only ERPs during the auditory N1 interval for all audiovisual pairs. One reason for this null result could be that the combined task-irrelevance of emotional information and auditory inputs led to the underlying emotional congruency effect being delayed. Additional research with a larger sample size might be needed to further confirm this null result and hence our interpretation of it.

5. Conclusions

In summary, the current ERP data provides clear evidence that the stimulus-driven attentional spreading from an audiovisual pair's attended visual constituent to its unattended auditory constituent could be modulated by emotional congruency between the visual and auditory constituents even when the emotional characteristics of all stimuli were task-irrelevant. This modulation emerged at the same time as the stimulus-driven attentional spreading occurred (~200 ms post-stimulus), and was further contingent on the emotional valence (positive/negative) of the audiovisual pair's visual constituent. These findings not only reveal when and how audiovisual emotional congruency influences the stimulus-driven cross-sensory attentional spreading in particular, adding to the existing studies focusing on the influence of audiovisual semantic congruency [13,18,19,21], but also advance our understanding regarding how high-level stimulus representation affects the low-level, bottom-up audiovisual binding process in general. The limitation of the work is that in order to ensure the task-irrelevance of emotional information (for the purpose of pro-

viding strong evidence for our findings), the *representation*-driven cross-sensory attentional spreading process, whose occurrence would require participants to selectively attend to a particular emotion in the visual modality, did not exist in the current study, hence could not be isolated and investigated in parallel [13]. Further studies with task designs such as the above might be required to concurrently examine the influence of audiovisual emotional congruency on both the stimulus- and representation-driven spreading of attention. Furthermore, recent studies have shown significant gender differences in cross-modal emotion perception [53,54]. It could also be a potential factor influencing the emotional cross-modal attentional spreading reported here. However, the insufficient number of participants recruited for each gender (8 males and 19 females) prevented us from exploring the gender difference with confidence. Further studies with larger sample sizes should examine this potential difference to uncover the role of gender in the emotional cross-modal spread of attention.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/brainsci12091229/s1>, Supplementary Information: Mixed effects models. References [55,56] are cited in the Supplementary Materials.

Author Contributions: M.C.: conceptualization, formal analysis, investigation, software, visualization, writing—original draft, writing—review and editing; S.Z.: conceptualization, formal analysis, funding acquisition, supervision, methodology, writing—original draft, writing—review and editing; J.Y.: conceptualization, formal analysis, investigation, software; X.L.: formal analysis, investigation, validation; M.Z.: formal analysis, investigation, software; C.F.: conceptualization, data curation, methodology, project administration, software, writing—review and editing; W.F.: conceptualization, data curation, funding acquisition, methodology, project administration, resources, supervision, visualization, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: The current study was funded by the National Key Research and Development Program of China (grant number 2021ZD0202600 to W.F.), the National Natural Science Foundation of China (31771200 and 32171048 to W.F.; 32200851 to S.Z.), the China Postdoctoral Science Foundation (2022M710104 to S.Z.), the Jiangsu Funding Program for Excellent Postdoctoral Talent (2022ZB591 to S.Z.) and Postgraduate Research and Practice Innovation Program of Jiangsu Province (KYCX21_2904 to M.C.).

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Soochow University (protocol code SUDA20210808H01, date of approval: 8 August 2021).

Informed Consent Statement: Informed consent was obtained from all participants involved in the study.

Data Availability Statement: The datasets presented in this article are not readily available because the datasets involve unfinished research projects. If necessary, those wishing to access the datasets should contact the corresponding authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Talsma, D.; Senkowski, D.; Soto-Faraco, S.; Woldorff, M.G. The Multifaceted Interplay between Attention and Multisensory Integration. *Trends Cogn. Sci.* **2010**, *14*, 400–410. [[CrossRef](#)]
2. De Meo, R.; Murray, M.M.; Clarke, S.; Matusz, P.J. Top-down Control and Early Multisensory Processes: Chicken vs. Egg. *Front. Integr. Neurosci.* **2015**, *9*, 17. [[CrossRef](#)]
3. Tang, X.; Wu, J.; Shen, Y. The Interactions of Multisensory Integration with Endogenous and Exogenous Attention. *Neurosci. Biobehav. Rev.* **2016**, *61*, 208–224. [[CrossRef](#)]
4. Macaluso, E.; Noppeney, U.; Talsma, D.; Vercillo, T.; Hartcher-O'Brien, J.; Adam, R. The Curious Incident of Attention in Multisensory Integration: Bottom-up vs. Top-Down. *Multisens. Res.* **2016**, *29*, 557–583. [[CrossRef](#)]
5. Senkowski, D.; Talsma, D.; Herrmann, C.S.; Woldorff, M.G. Multisensory Processing and Oscillatory Gamma Responses: Effects of Spatial Selective Attention. *Exp. Brain Res.* **2005**, *166*, 411–426. [[CrossRef](#)]

6. Li, Q.; Wu, J.; Touge, T. Audiovisual Interaction Enhances Auditory Detection in Late Stage: An Event-Related Potential Study. *NeuroReport* **2010**, *21*, 173–178. [[CrossRef](#)]
7. Talsma, D.; Woldorff, M.G. Selective Attention and Multisensory Integration: Multiple Phases of Effects on the Evoked Brain Activity. *J. Cogn. Neurosci.* **2005**, *17*, 1098–1114. [[CrossRef](#)]
8. Paszkiel, S. Analysis and Classification of EEG Signals for Brain–Computer Interfaces. In *Studies in Computational Intelligence*; Springer International Publishing: Cham, Switzerland, 2020; Volume 852, ISBN 978-3-030-30580-2.
9. Luck, S.J. *An Introduction to the Event-Related Potential Technique*, 2nd ed.; The MIT Press: Cambridge, MA, USA, 2014; ISBN 978-0-262-52585-5.
10. Busse, L.; Roberts, K.C.; Crist, R.E.; Weissman, D.H.; Woldorff, M.G. The Spread of Attention across Modalities and Space in a Multisensory Object. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 18751–18756. [[CrossRef](#)]
11. Talsma, D.; Doty, T.J.; Woldorff, M.G. Selective Attention and Audiovisual Integration: Is Attending to Both Modalities a Prerequisite for Early Integration? *Cereb. Cortex* **2007**, *17*, 679–690. [[CrossRef](#)]
12. Molholm, S.; Martinez, A.; Shpaner, M.; Foxe, J.J. Object-Based Attention Is Multisensory: Co-Activation of an Object’s Representations in Ignored Sensory Modalities: Multisensory Transfer of Object-Based Attention. *Eur. J. Neurosci.* **2007**, *26*, 499–509. [[CrossRef](#)]
13. Fiebelkorn, I.C.; Foxe, J.J.; Molholm, S. Dual Mechanisms for the Cross-Sensory Spread of Attention: How Much Do Learned Associations Matter? *Cereb. Cortex* **2010**, *20*, 109–120. [[CrossRef](#)]
14. Fiebelkorn, I.C.; Foxe, J.J.; Schwartz, T.H.; Molholm, S. Staying within the Lines: The Formation of Visuospatial Boundaries Influences Multisensory Feature Integration: Multisensory Feature Integration. *Eur. J. Neurosci.* **2010**, *31*, 1737–1743. [[CrossRef](#)]
15. Hansen, J.C.; Hillyard, S.A. Endogenous Brain Potentials Associated with Selective Auditory Attention. *Electroencephalogr. Clin. Neurophysiol.* **1980**, *49*, 277–290. [[CrossRef](#)]
16. Naatanen, R. Processing Negativity: An Evoked-Potential Reflection of Selective Attention. *Psychol. Bull.* **1982**, *92*, 605–640. [[CrossRef](#)]
17. Singhal, A.; Fowler, B. The Differential Effects of Sternberg Short- and Long-Term Memory Scanning on the Late Nd and P300 in a Dual-Task Paradigm. *Cogn. Brain Res.* **2004**, *21*, 124–132. [[CrossRef](#)]
18. Zimmer, U.; Itthipanyanan, S.; Grent-’t-Jong, T.; Woldorff, M.G. The Electrophysiological Time Course of the Interaction of Stimulus Conflict and the Multisensory Spread of Attention: Attentional Spread Precedes Multisensory Conflict. *Eur. J. Neurosci.* **2010**, *31*, 1744–1754. [[CrossRef](#)]
19. Zimmer, U.; Roberts, K.C.; Harshbarger, T.B.; Woldorff, M.G. Multisensory Conflict Modulates the Spread of Visual Attention across a Multisensory Object. *NeuroImage* **2010**, *52*, 606–616. [[CrossRef](#)]
20. Donohue, S.E.; Roberts, K.C.; Grent-’t-Jong, T.; Woldorff, M.G. The Cross-Modal Spread of Attention Reveals Differential Constraints for the Temporal and Spatial Linking of Visual and Auditory Stimulus Events. *J. Neurosci.* **2011**, *31*, 7982–7990. [[CrossRef](#)]
21. Zhao, S.; Feng, C.; Liao, Y.; Huang, X.; Feng, W. Attentional Blink Suppresses Both Stimulus-driven and Representation-driven Cross-modal Spread of Attention. *Psychophysiology* **2021**, *58*, e13761. [[CrossRef](#)]
22. Zhao, S.; Li, Y.; Wang, C.; Feng, C.; Feng, W. Updating the Dual-mechanism Model for Cross-sensory Attentional Spreading: The Influence of Space-based Visual Selective Attention. *Hum. Brain Mapp.* **2021**, *42*, 6038–6052. [[CrossRef](#)]
23. Klasen, M.; Chen, Y.-H.; Mathiak, K. Multisensory Emotions: Perception, Combination and Underlying Neural Processes. *Rev. Neurosci.* **2012**, *23*, 381–392. [[CrossRef](#)]
24. Kokinous, J.; Kotz, S.A.; Tavano, A.; Schröger, E. The Role of Emotion in Dynamic Audiovisual Integration of Faces and Voices. *Soc. Cogn. Affect. Neurosci.* **2015**, *10*, 713–720. [[CrossRef](#)]
25. Zinchenko, A.; Kanske, P.; Obermeier, C.; Schröger, E.; Kotz, S.A. Emotion and Goal-Directed Behavior: ERP Evidence on Cognitive and Emotional Conflict. *Soc. Cogn. Affect. Neurosci.* **2015**, *10*, 1577–1587. [[CrossRef](#)]
26. Zinchenko, A.; Obermeier, C.; Kanske, P.; Schröger, E.; Kotz, S.A. Positive Emotion Impedes Emotional but Not Cognitive Conflict Processing. *Cogn. Affect. Behav. Neurosci.* **2017**, *17*, 665–677. [[CrossRef](#)]
27. Föcker, J.; Röder, B. Event-Related Potentials Reveal Evidence for Late Integration of Emotional Prosody and Facial Expression in Dynamic Stimuli: An ERP Study. *Multisens. Res.* **2019**, *32*, 473–497. [[CrossRef](#)]
28. Wang, Z.; Chen, M.; Goerlich, K.S.; Aleman, A.; Xu, P.; Luo, Y. Deficient Auditory Emotion Processing but Intact Emotional Multisensory Integration in Alexithymia. *Psychophysiology* **2021**, *58*, e13806. [[CrossRef](#)]
29. Campbell, J.I.D.; Thompson, V.A. MorePower 6.0 for ANOVA with Relational Confidence Intervals and Bayesian Analysis. *Behav. Res. Methods* **2012**, *44*, 1255–1265. [[CrossRef](#)]
30. Lu, B.; Hui, M.; Yu-Xia, H. The Development of Native Chinese Affective Picture System—A Pretest in 46 College Students. *Chin. Ment. Health J.* **2005**, *19*, 712–719.
31. Lang, P.J.; Greenwald, M.K.; Bradley, M.M.; Hamm, A.O. Looking at Pictures: Affective, Facial, Visceral, and Behavioral Reactions. *Psychophysiology* **1993**, *30*, 261–273. [[CrossRef](#)]
32. Baucom, L.B.; Wedell, D.H.; Wang, J.; Blitzer, D.N.; Shinkareva, S.V. Decoding the Neural Representation of Affective States. *NeuroImage* **2012**, *59*, 718–727. [[CrossRef](#)]

33. Shinkareva, S.V.; Wang, J.; Kim, J.; Facciani, M.J.; Baucom, L.B.; Wedell, D.H. Representations of Modality-Specific Affective Processing for Visual and Auditory Stimuli Derived from Functional Magnetic Resonance Imaging Data: Modality-Specific Affective Processing. *Hum. Brain Mapp.* **2014**, *35*, 3558–3568. [[CrossRef](#)]
34. Bradley, M.M.; Lang, P.J. *International Affective Digitized Sounds (IADS): Stimuli, Instruction Manual and Affective Ratings (Tech. Rep. No. B-2)*; University of Florida: Gainesville, FL, USA, 1999.
35. Peirce, J.; Gray, J.R.; Simpson, S.; MacAskill, M.; Höchenberger, R.; Sogo, H.; Kastman, E.; Lindeløv, J.K. PsychoPy2: Experiments in behavior made easy. *Behav. Res.* **2019**, *51*, 195–203. [[CrossRef](#)]
36. Bradley, M.M.; Lang, P.J. Measuring Emotion: The Self-Assessment Manikin and the Semantic Differential. *J. Behav. Ther. Exp. Psychiatry* **1994**, *25*, 49–59. [[CrossRef](#)]
37. Marchewka, A.; Żurawski, Ł.; Jednoróg, K.; Grabowska, A. The Nencki Affective Picture System (NAPS): Introduction to a Novel, Standardized, Wide-Range, High-Quality, Realistic Picture Database. *Behav. Res. Methods* **2014**, *46*, 596–610. [[CrossRef](#)]
38. Bertelson, P.; Aschersleben, G. Automatic Visual Bias of Perceived Auditory Location. *Psychon. Bull. Rev.* **1998**, *5*, 482–489. [[CrossRef](#)]
39. Zhao, S.; Wang, C.; Feng, C.; Wang, Y.; Feng, W. The Interplay between Audiovisual Temporal Synchrony and Semantic Congruency in the Cross-modal Boost of the Visual Target Discrimination during the Attentional Blink. *Hum. Brain Mapp.* **2022**, *43*, 2478–2494. [[CrossRef](#)]
40. Delorme, A.; Makeig, S. EEGLAB: An Open Source Toolbox for Analysis of Single-Trial EEG Dynamics Including Independent Component Analysis. *J. Neurosci. Methods* **2004**, *134*, 9–21. [[CrossRef](#)]
41. Lopez-Calderon, J.; Luck, S.J. ERPLAB: An Open-Source Toolbox for the Analysis of Event-Related Potentials. *Front. Hum. Neurosci.* **2014**, *8*, 213. [[CrossRef](#)]
42. Walter, W.G.; Cooper, R.; Aldridge, V.J.; McCallum, W.C.; Winter, A.L. Contingent Negative Variation: An Electric Sign of Sensori-Motor Association and Expectancy in the Human Brain. *Nature* **1964**, *203*, 380–384. [[CrossRef](#)]
43. Luck, S.J.; Gaspelin, N. How to Get Statistically Significant Effects in Any ERP Experiment (and Why You Shouldn't): How to Get Significant Effects. *Psychophysiology* **2017**, *54*, 146–157. [[CrossRef](#)]
44. Zeelenberg, R.; Bocanegra, B.R. Auditory Emotional Cues Enhance Visual Perception. *Cognition* **2010**, *115*, 202–206. [[CrossRef](#)]
45. Smith, N.K.; Cacioppo, J.T.; Larsen, J.T.; Chartrand, T.L. May I Have Your Attention, Please: Electrocortical Responses to Positive and Negative Stimuli. *Neuropsychologia* **2003**, *41*, 171–183. [[CrossRef](#)]
46. Delplanque, S.; Lavoie, M.E.; Hot, P.; Silvert, L.; Sequeira, H. Modulation of Cognitive Processing by Emotional Valence Studied through Event-Related Potentials in Humans. *Neurosci. Lett.* **2004**, *356*, 1–4. [[CrossRef](#)]
47. Carretié, L.; Hinojosa, J.A.; López-Martín, S.; Tapia, M. An Electrophysiological Study on the Interaction between Emotional Content and Spatial Frequency of Visual Stimuli. *Neuropsychologia* **2007**, *45*, 1187–1195. [[CrossRef](#)]
48. Luo, W.; Feng, W.; He, W.; Wang, N.-Y.; Luo, Y.-J. Three Stages of Facial Expression Processing: ERP Study with Rapid Serial Visual Presentation. *NeuroImage* **2010**, *49*, 1857–1867. [[CrossRef](#)]
49. Zhang, D.; He, W.; Wang, T.; Luo, W.; Zhu, X.; Gu, R.; Li, H.; Luo, Y. Three Stages of Emotional Word Processing: An ERP Study with Rapid Serial Visual Presentation. *Soc. Cogn. Affect. Neurosci.* **2014**, *9*, 1897–1903. [[CrossRef](#)]
50. Keil, A.; Bradley, M.M.; Hauk, O.; Rockstroh, B.; Elbert, T.; Lang, P.J. Large-Scale Neural Correlates of Affective Picture Processing. *Psychophysiology* **2002**, *39*, 641–649. [[CrossRef](#)]
51. De Cesare, A.; Codispoti, M. When Does Size Not Matter? Effects of Stimulus Size on Affective Modulation. *Psychophysiology* **2006**, *43*, 207–215. [[CrossRef](#)]
52. Donohue, S.E.; Todisco, A.E.; Woldorff, M.G. The Rapid Distraction of Attentional Resources toward the Source of Incongruent Stimulus Input during Multisensory Conflict. *J. Cogn. Neurosci.* **2013**, *25*, 623–635. [[CrossRef](#)]
53. Lin, Y.; Ding, H.; Zhang, Y. Unisensory and Multisensory Stroop Effects Modulate Gender Differences in Verbal and Nonverbal Emotion Perception. *J. Speech Lang. Hear. Res.* **2021**, *64*, 4439–4457. [[CrossRef](#)]
54. Lin, Y.; Ding, H.; Zhang, Y. Gender Differences in Identifying Facial, Prosodic, and Semantic Emotions Show Category- and Channel-Specific Effects Mediated by Encoder's Gender. *J. Speech Lang. Hear. Res.* **2021**, *64*, 2941–2955. [[CrossRef](#)] [[PubMed](#)]
55. Kuznetsova, A.; Brockhoff, P.B.; Christensen, R.H. lmerTest package: Tests in linear mixed effects models. *J. Stat. Softw.* **2017**, *82*, 1–26. [[CrossRef](#)]
56. Lenth, R. Emmeans: Estimated Marginal Means, Aka Leastsquares Means. R Package. 2017. Available online: <https://rdrr.io/cran/emmeans/man/emmeans-package.html> (accessed on 1 September 2022).