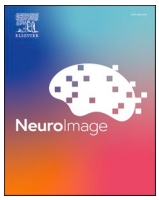
NeuroImage 331 (2026) 121873

Contents lists available at ScienceDirect

NeuroImage

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



Temporally congruent auditory stream modulates visual processing both independently of and interactively with selective attention in a competing scenario

Jieru Chen [a,b], Wenjie Liu [c], Shiqi Tan [a,b], Xiangyong Yuan [a,b,*], Yi Jiang [a,b]

[a] State Key Laboratory of Cognitive Science and Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China
[b] Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049, China
[c] Beijing Huilongguan Hospital, Huilongguan Clinical Medical College, Peking University, Beijing 100096, China

ARTICLE INFO

Keywords:
Audiovisual interaction
Temporal congruency
Selective attention
Rhythmicity
Steady-state visual evoked potentials

ABSTRACT

In competing environments, both selective attention and audiovisual interaction can facilitate visual processing, yet whether their influences operate independently or interactively remains debated. Using electroencephalography (EEG), we addressed this issue by instructing participants to selectively attend to one of two lateralized flickering discs, which also changed their shapes either temporally congruent or incongruent with a pitch-changing sound. We found that reaction times for detecting deviants embedded in the attended visual stream were reduced when a temporally congruent sound was concurrently played. Compared to a temporally incongruent auditory stream, a congruent one selectively enhanced steady-state visual evoked potentials (SSVEPs) in response to flickering of the unattended stream. In contrast, the SSVEP and inter-trial phase coherence in response to the shape-modulation for both attended and unattended streams were enhanced at the harmonic frequencies by the temporally congruent sound. The results indicate that the auditory influence on visual processing orthogonal to audiovisual temporal congruency (flicker) interacts with attention, whereas the auditory influence on visual processing relevant to audiovisual temporal congruency (shape-modulation) is largely independent of attention. However, these congruency effects were observed only under rhythmic audiovisual streams: When audiovisual pitch-shape modulation followed unrhythmic temporal structures, these congruency effects totally disappeared. Together, these findings demonstrate that temporally congruent auditory streams can modulate visual processing both independently of and interactively with selective attention, highlighting a flexible and complex interplay between selective attention and audiovisual interaction.

1. Introduction

At a noisy cocktail party, individuals prioritize relevant conversations while filtering out distracting ones by selectively attending to a partner's voice and ignoring other speakers or background noise. Comprehension of the attended speech can be further facilitated by simultaneously viewing the speaker's face, especially lip movements, through enhanced cortical tracking of the speech envelope in the auditory cortex (Ahmed et al., 2023; Haider et al., 2024; Park et al., 2016; Reisinger et al., 2025; Zion Golumbic et al., 2013). Not only speeches, auditory tracking of selectively attended but meaningless sound streams can also be enhanced by temporally congruent visual streams (Atilgan et al., 2018; Maddox et al., 2015; Peng et al., 2023), and further improved by training of audiovisual temporal precision. Conversely, temporally and semantically congruent auditory stimuli are able to enhance the salience of target among distractors in the visual domain, thereby attracting bottom-up attention and improving visual search (Iordanescu et al., 2008; Kvasova et al., 2019; Shen et al., 2023a; Van der Burg et al., 2008, 2011).

However, in competitive visual environments analogous to the "cocktail party," it remains unclear whether a temporally congruent auditory stream modulates visual processing independently of top-down selective attention, or whether such cross-modal influences interact with selective attention in a manner that differentially impacts attended and unattended visual streams. Existing literature diverges on this issue. One line of studies suggests that the influence of temporally congruent

* Corresponding author at: State Key Laboratory of Cognitive Science and Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China.
E-mail address: yuanxy@psych.ac.cn (X. Yuan).

https://doi.org/10.1016/j.neuroimage.2026.121873
Received 24 September 2025; Received in revised form 10 February 2026; Accepted 23 March 2026
Available online 23 March 2026

auditory stream on visual processing is largely independent of selective attention. For example, a sound whose pitch continuously co-varies with the spatial frequency of the visual stimuli in competitive visual scenes, can equally enhance the steady-state evoked visual responses (SSVEPs) at the harmonic frequencies of the pitch-spatial-frequency co-variation, irrespective of whether the visual stimuli are attended or ignored (Covic et al., 2017; Keitel and Mueller, 2015). Consistently, the independence between audiovisual interaction and selective attention has also been demonstrated in cross-modal illusions, such as the ventriloquism (Bertelson et al., 2000).

In contrast, another line of studies reports that the modulation of visual processing by temporally congruent auditory streams depends on selective attention. An early neuroimaging study showed increased activation in both multisensory integration regions (e.g., superior temporal sulcus and superior colliculus) and visual sensory areas when participants listened to speech congruent with the lip movements they attended, compared to unattended lip movements (Fairhall and Macaluso, 2009). Electroencephalography (EEG) studies also revealed enhanced early auditory-specific high gamma responses (Senkowski et al., 2005) or amplified the audiovisual interaction for attended rather than unattended stimulus sequences (Talsma and Woldorff, 2005). The same conclusion is reached across stimuli and paradigm (Alsius et al., 2014, 2005; Seijdel et al., 2024). On the contrary, others found a greater improvement of visual target detection on the unattended side than the attended side (Zou et al., 2012), see also Van der Stoep et al. (2015) for a similar finding. It has also been demonstrated that the SSVEPs to ignored speakers in multiple speaker scenarios were enhanced by congruent speech sounds, resulting in stronger interference to speech recognition of the attended speakers (Krause et al., 2012; Senkowski et al., 2008). However, the authors did not explicitly compare the SSVEPs between attended and unattended speakers; it is therefore inconclusive which speaker was enhanced to a greater extent.

Taken together, these divergent findings suggest a flexible and complex relationship between audiovisual interaction and selective attention, influenced by various factors such as stimulus complexity (speech or nonspeech), task demands, and competition between sensory streams (Talsma et al., 2010). However, one critical but often overlooked factor in resolving these discrepancies is the relevance of specific stimulus features to audiovisual temporal congruency (Bizley et al., 2016). For continuous audiovisual streams, temporal congruency between the dynamic changes of auditory and visual features (e.g., size-pitch covariation) is a widely recognized prerequisite for audiovisual interaction (Atilgan and Bizley, 2021; Crosse et al., 2015; Maddox et al., 2015; Parise et al., 2013; Shen et al., 2023a; Yuan et al., 2020). However, there are other features orthogonal to audiovisual temporal congruency in defining a unified audiovisual object, such as luminance or timbre (Bizley et al., 2016). In the cocktail party scenario, Peng et al. (2023) has discovered that EEG tracking of an amplitude-modulated sound stream was enhanced by a temporally congruent size-modulated visual streams, independently of attention; while the event-related neural responses to timber deviants, which are irrelevant to temporal congruency, were influenced only in the unattended sound stream. Thus, the first objective of the current study was to explore whether the modulation of temporally congruent auditory streams on visual processing is unaffected, amplified, or attenuated by selective attention depending on specific visual features being processed.

Furthermore, literature that investigated how audiovisual temporal congruency and selective attention collectively influence visual processing mainly used speech or simple rhythmic sequences (Covic et al., 2017; Fairhall and Macaluso, 2009; Keitel and Mueller, 2015; Krause et al., 2012; Seijdel et al., 2024). Speech shares commonality with rhythmic sequences in their intrinsic temporal regularity (Giraud and Poeppel, 2012; Park et al., 2016), which allows content and timing in the future being predicted by the past. These characteristics of rhythmic and quasi-rhythmic stimuli including speech dramatically differ from irregular, unrhythmic stimuli. It has been demonstrated that rhythmic audiovisual stimulation elicits enhanced integration compared to unrhythmic stimulation (Heins et al., 2021; Marchant et al., 2012; ten Oever et al., 2014). For example, auditory detection threshold was lower in rhythmic sequences than in random sequences when slightly preceded by a series of visual cues (ten Oever et al., 2014). Synchronous, rhythmic audiovisual sequences elicited stronger activation in the right inferior parietal lobule than synchronous but unrhythmic audiovisual sequences (Marchant et al., 2012). Accordingly, the second objective of the current study was to explore whether the interplay between audiovisual temporal congruency and selective attention on visual processing is influenced by the temporal structure of the audiovisual streams.

To address the two issues, we conducted an EEG study employing the frequency-tagging technique (Covic et al., 2017; Keitel and Mueller, 2015; Nozaradan et al., 2012; Sciortino and Kayser, 2023) to simultaneously capture the steady-state responses to visual features relevant and orthogonal to audiovisual temporal congruency. It should be emphasized that some studies failed to track an auditory influence on SSVEPs tagging to visual features (a flicker) orthogonal to audiovisual temporal congruency (Covic et al., 2017; Giani et al., 2012; Keitel and Mueller, 2015), while others succeeded (Drijvers et al., 2020; Nozaradan et al., 2012; Sciortino and Kayser, 2023). Having reviewed these studies, we realized that the crucial factor determining whether auditory influences on visual processing orthogonal to audiovisual temporal congruency could be successively tracked by SSVEPs, might be the association between audiovisual stimuli. Among those tested, audiovisual association between shape/motion and pitch/beat changes rather than pitch-hue and pitch-saturation is more likely to evoke robust influences on flicker-induced SSVEPs (Nozaradan et al., 2012; Sciortino and Kayser, 2023).

Thus, the present study employed a pitch-shape co-variation to define audiovisual temporal congruency. Specifically, participants were cued to selectively attend a lateral disc whose shape continuously changed while ignoring another shape-modulated disc located at the opposite side. A tone, whose pitch change was congruent with either the shape change of the attended disc or the unattended disc, was displayed. In the rhythmic context, the pitch-shape modulation was a regular one while in the unrhythmic context it was an irregular one. In addition, the luminance of the two discs changed at a relatively faster but fixed speed. In this design, the steady-state visual responses evoked by shape-modulation and flicker reflect the visual processing relevant or orthogonal to the audiovisual temporal congruency, respectively. We hypothesized that for features relevant to temporal congruency (the shape-modulation), selective attention and temporal congruency would enhance steady-state responses independently, consistent with Covic et al. (2017), Keitel and Mueller (2015). Whereas for orthogonal features (the flicker), we hypothesized an interaction between them as Peng et al. (2023) reported, such that audiovisual temporal congruency would exert a differential influence on attended and unattended visual streams. Finally, we hypothesized that the congruency effect would primarily emerge in the rhythmic context, attenuate or even disappear under unrhythmic stimulation, according to Heins et al. (2021), Marchant et al. (2012), ten Oever et al. (2014).

2. Materials and methods

2.1. Participants

A total of 50 participants took part in the study, with 25 in the rhythmic context (12 males, mean age ± SD = 23.0 ± 2.3 years) and the other 25 in the unrhythmic context (12 males, mean age ± SD = 23.3 ± 2.2 years). All had normal or corrected-to-normal vision and normal hearing. The study was approved by the institutional review board of the Institute of Psychology, Chinese Academy of Sciences, and adhered to the tenets of the Declaration of Helsinki. Two participants in the rhythmic context were excluded from the behavioral analysis due to low behavioral accuracy (<50%), and two participants in the unrhythmic

context were excluded from the EEG analysis due to few valid EEG trials left after preprocessing.

2.2. Apparatus and stimuli

The experiment was conducted in a dim, sound-attenuated and electric-shielded room. Participants sat comfortably at a viewing distance of about 65 cm from a CRT monitor. The monitor had a refresh rate of 60 Hz, and a resolution of 1280 × 1024 (37.5 cm × 30 cm). Visual stimulation consisted of two gray circular discs (49.2 cd/m^2) with a diameter of ~4.1° of visual angle, one serving as the target and the other as the distractor, positioned in either the left or right visual field. Stimuli were presented against a black background (2.36 cd/m^2, Fig. 1A). A gray arrow (0.26° × 0.52°, 49.2 cd/m^2) in the center of the display served as both the fixation and attentional cue. The visual stimulations underwent two independent changes in a trial: (1) One disc followed a cycle of 2 on-frames and 3 off-frames (2/3 on/off-ratio), resulting in a 12 Hz flicker (off-frame luminance = 9.7 cd/m^2). The other disc flickered at a rate of 15 Hz produced by a cycle of 2 on-frames and 2 off-frames (2/2 on/off-ratio). (2) The two discs also changed their shapes from discs to ellipses and then back to discs several times in the sequences. In the rhythmic context, their shapes were modulated at a fixed frequency of 2.61 or 2.07 Hz, respectively. In the unrhythmic context, their shapes are controlled by a modulation function that pseudo-randomly changed its frequency every 0.5 s within a range of 2.31 ± 1 Hz (Fig. 1B). Two sets of unrhythmic sequences were generated in advance and were counterbalanced between participants. The selection of both flicker and modulation frequencies was guided by previous audiovisual frequency-tagging studies (e.g., Keitel and Müller, 2015; Covic et al., 2017; Nozaradan et al., 2012; Sciortino and Kayser, 2023).

The auditory stimulation consisted of tones with a sample rate of 44.1 kHz presented via in-ear headphones. The tones, with a central carrier frequency of 1000 Hz and a deviation of 200 Hz, were frequency modulated by the same sequence as the discs. That is, in the rhythmic context, the tone's pitch was modulated at either 2.07 or 2.61 Hz; in the unrhythmic context, the tone's pitch was pseudo-randomly modulated every 0.5 s ranging within 2.31 ± 1 Hz (Fig. 1B). The tone was presented at a sound intensity of ~39.4 dB(A), while the background noise at 30.9 dB(A). An additional 0.5 s cosine ramp-on was applied to both the visual and auditory streams to prevent cross-modal synchronization based on abrupt onset. All stimuli were generated by MATLAB (The MathWorks, Natick, MA) and presented using Psychtoolbox (Brainard, 1997; Pelli, 1997). To further minimize hardware-related audiovisual latency delay, we employed the ASIO4ALL low-latency audio driver to deliver auditory stimuli.

2.3. Experimental design and procedure

At the beginning of each trial, participants were cued to attend exclusively to the left or the right visual stream by the central arrow. After an interval of 0.5–1 s, the two flickering discs, one on the left side of the arrow, the other on the right, continuously changed their shapes for 9 s. Participants also heard a tone during presentation period of the visual streams and were asked to pay attention to it as well. The tone's pitch was modulated either congruently with the shape modulation of the attended disc or of the unattended disc. At the end of each trial, the screen turned blank and remained for an extra 1 s allowing participants to relax and blink before the next trial started (Fig. 1A). To make sure that participants followed the instruction and selectively attended to the visual target, occasional deviant events were embedded in the discs and tones in 27.3% of trials. In each of these trials, there were 2–3 deviant events with a minimum interval of 1 s between them. These events would not occur in the beginning 1 s of the trial. There were three types of deviant events. The attended disc suddenly decreased its brightness, the unattended disc suddenly decreased its brightness, or another 400-Hz tone suddenly appeared. Each event lasted 0.2 s. Participants were

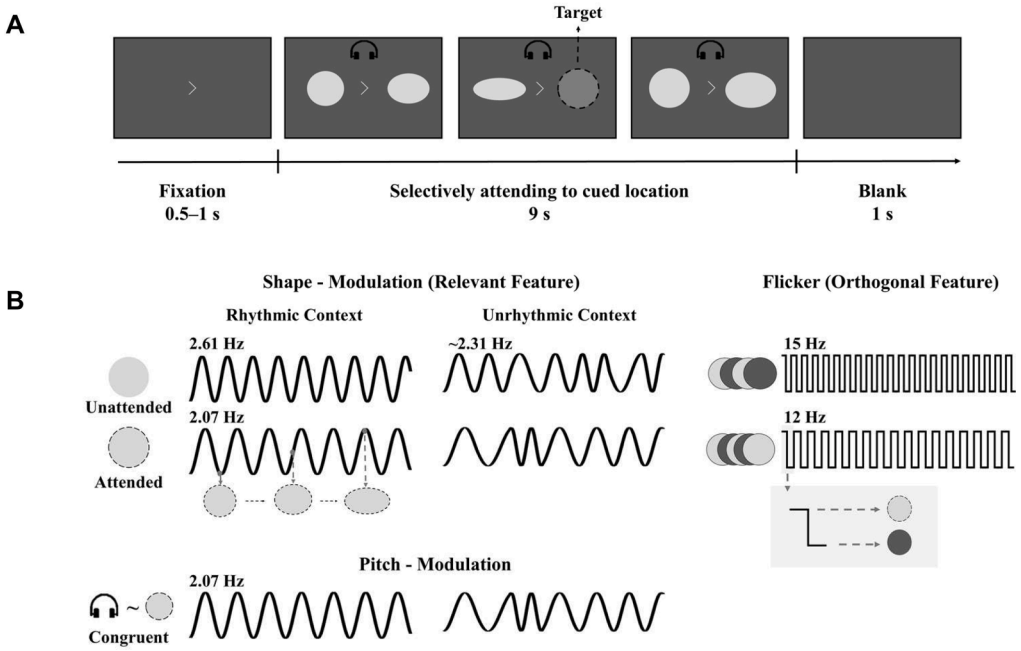


Fig. 1. Experimental procedure and stimulus design. (A) Visual stimulation consisted of two discs continuously flickering and changing their shapes. A central arrow cue instructed participants which disc to attend. Simultaneously, a tone with continuously changing pitch was presented, either temporally congruent or incongruent with the shape modulation of the attended disc. In the example trial shown, the attended disc's luminance suddenly decreased, serving as the visual target. Participants were required to press a button as quickly as possible upon detecting this deviant change. (B) An example sequence of the shape-modulation and flickering functions for the attended and unattended disc, as well as the pitch-modulation function of the tone. In the rhythmic context, both sequences were modulated at frequencies of 2.07 or 2.61 Hz. In the unrhythmic context, the frequency of both sequences varied around 2.31 Hz; In the example shown in (B), the pitch modulation of the tone is temporally congruent with the shape modulation of the attended disc. The shape-modulation represents the feature relevant to audiovisual temporal congruency. In addition, the discs flickered at 12 or 15 Hz, which served as the feature orthogonal to audiovisual temporal congruency.

also instructed to respond as quickly as possible to the deviant events occurring on the attended disc (visual target) and the tone (auditory target) while ignoring them on the unattended disc (visual distractor). Demos of audiovisual stimulation in an example trial from the rhythmic and unrhythmic contexts can be found in the supplementary material.

We manipulated selective attention and audiovisual temporal congruency in both the rhythmic and unrhythmic contexts. Note that the tone was always temporally congruent with one of the two discs; for example, when it was congruent with the attended disc in the trial, it was incongruent with the unattended disc. Likewise, when one disc on one side was attended, the other disc on the other side was unattended. In total, there are 110 trials divided into 5 blocks. In each block, there were 16 trials without and 6 trials with deviants (a total of 15 deviants, 5 in each type). The combination of attended disc, attended side, and congruent disc resulted in 8 treatments in each block, each repeated twice. In the deviant trials, the three variables cannot be fully counterbalanced but randomized. Trials in each block were presented in a randomized order to minimize any potential order effects and systematic biases. Before formal EEG recording, participants were pretested using a 3-down-1-up staircase to individually determine their detection thresholds for the visual and auditory targets (by changing the visual brightness and auditory intensity stepwise). The threshold test and the EEG experiment lasted approximately 45 min.

2.4. Behavioral data analysis

Responses were classified as 'hit' when participants responded within a time window of 0.2–1 s and within three standard deviations of the mean after the target onset. Responses were classified as 'false alarm' if participants responded to distractors within the same time limit. Since our participants made rare false alarms (mean $\pm$ SD = 1.8 $\pm$ 4.1% in the rhythmic context and 1.3 $\pm$ 1.8% in the unrhythmic context), we did not consider 'false alarm' in the following analysis.

The response accuracy (ACC) was calculated as the ratio of hits to the total number of targets for each condition and participant. The mean reaction time (RT) of hits for each participant was also calculated. Afterwards, the RTs and ACCs were subjected to a two-way mixed-design analysis of variance (ANOVA), with rhythmicity (rhythmic and unrhythmic) as a between-subject factor and audiovisual temporal congruency (the tone congruent or incongruent with the attended disc) as a within-subject factor. The ANOVAs were conducted for both visual and auditory targets. We mainly focused on whether audiovisual temporal congruency consistently influenced the processing of selectively attended visual streams in rhythmic and unrhythmic contexts.

2.5. EEG data acquisition

EEG data was acquired using a 64-channel Neuroscan SynAmps RT system (Compumedics, USA) while participants performed the behavioral task. Continuous EEG signals were digitized at a sample rate of 1000 Hz, referenced to an electrode between Cz and CPz. The horizontal and vertical eye movements were measured using four electrodes placed above and below the lower eyelids of the left eye and at the outer canthus of both eyes. Impedances were kept below 10 kΩ for all electrodes.

2.6. EEG data analysis

2.6.1. Preprocessing

Raw EEG data were analyzed using EEGLAB (Delorme and Makeig, 2004) and custom scripts running on MATLAB. The offline data were downsampled to 250 Hz and band-pass filtered to 0.2 and 60 Hz. Bad channels were identified by visual inspection for each participant and spherically interpolated for subsequent analyses (EEGLAB function *eeg_interp*). Data were first segmented into long epochs of 13 secs (−2 to 11 *sec* relative to the stimulation onset of each trial) before independent component analysis (ICA) to remove noisy signals recorded at rest periods while preserving sufficient data length and stationarity for ICA decomposition (Luck, 2022; Ouyang and Li, 2025). The artifacts such as eye blinks and movements were then removed based on the components extracted by ICA. The segmented data were re-referenced to the average mastoid signals, and baseline-corrected using the time window of −500–0 ms. Epochs with amplitude exceeding ±120 μV were automatically rejected. Other residual artifacts were manually checked and rejected through visual inspection. On average, participants had 8.1% trials rejected in the rhythmic context. Two participants were excluded in the unrhythmic context because there were only 13 and 7 clean trials left after preprocessing. For the remaining participants, an average of 6.19% trials were rejected in the unrhythmic context.

2.6.2. Event-related potentials

Artifact-free data were baseline-corrected relative to the interval from −200 ms to 0 ms. In both rhythmic and unrhythmic contexts, three visual event-related potential (ERP) components at the parieto-occipital electrodes were identified: P1 (100–140 ms), N1 (160–200 ms), and P2 (240–280 ms). We then computed the mean amplitude of the three ERP components on electrodes contralateral and ipsilateral to the attended side (P7/P8, P5/P6, P3/P4, PO7/PO8, PO3/PO4) for temporally congruent and incongruent conditions, respectively. The reported ERP components and electrodes of interest were selected based on scalp topography, as well as previous literature which consistently showed that the three early visual ERP components at the lateral parieto-occipital electrodes can capture attentional and multisensory modulations (Natale et al., 2006; Slagter et al., 2016; Störmer et al., 2009). Two-way repeated-measures ANOVAs were conducted for the amplitude of each ERP component in both rhythmic and unrhythmic contexts, with factors of attended side (contralateral and ipsilateral) and audiovisual temporal congruency (the tone congruent or incongruent with the attended disc).

2.6.3. Fast Fourier transform

For each participant, in each trial at each electrode, fast Fourier transform (FFT) was applied to the EEG signal from 1 to 9 s, transforming the data from the time domain to the frequency domain. The first 1 s were discarded to exclude ERPs to stimulus onset from spectral analyses. The nominal frequency resolution was increased to 0.1 Hz by zero-padding. The power was defined as the squared magnitude of the complex signals after FFT and then normalized into decibel scale [$10 \times \log10(\text{power})$]. To remove the 1/f trend of the power spectrum, the power at each frequency was normalized by subtracting out the power at its neighboring frequency bins (two bins on each, 0.1 Hz) (Lenc et al., 2018; Nozaradan et al., 2012; Shen et al., 2023b). The steady-state potentials were then calculated by averaging the normalized power across trials separately for each electrode, participant, and condition.

2.6.4. Steady-state visual evoked potential

The average SSVEPs at the tagging frequencies and their harmonics across the 19 posterior electrodes (Oz, O1, O2, CB1, CB2, POz, PO3, PO4, PO7, PO8, Pz, P1, P2, P3, P4, P5, P6, P7, and P8) were depicted in Fig. 4A (the rhythmic context) and Fig. 6A (the unrhythmic context). Notably, the power spectra display robust harmonic responses at twice the modulate frequencies (4.14 and 5.22 Hz). These harmonics were included in further analyses as fundamental and harmonic responses may reflect different aspects of stimulus processing and modulated by audiovisual temporal congruency distinctively (Keitel and Mueller, 2015; Kim et al., 2011; Porcu et al., 2013)

The grand-average topographical distribution of SSVEPs evoked by shape-modulation and flicker, averaged across conditions and participants, reveals prominent peaks at parieto-occipital electrodes in the rhythmic and unrhythmic contexts (Fig. 4C and Fig. 6C, respectively). As the scalp topography of SSVEPs varied across participants and frequencies, among the 19 posterior electrodes we selected electrodes of

interest in which the power at the tagging frequencies (modulate/flicker frequencies) exceeded the power in their neighboring frequency band, considering that a stable SSVEP should be larger than the noisy power fluctuation at its neighborhood. Specifically, we selected the top 10 electrodes on which the SSVEPs had the largest difference from the maximum power within the neighboring frequency bands for each tagging frequency and participant (±0.3 Hz for modulate 1f, ±0.5 Hz for modulate 2f, and ±1.0 Hz for flicker). In cases where fewer than 10 electrodes met the criteria for a given participant or frequency, all qualified electrodes were included. The electrode selection was conducted for each participant, modulate/flicker frequencies, and attended side, while merging the data across selective attention and audiovisual temporal congruency conditions. The number of electrodes chosen at the modulate and flicker frequencies in the rhythmic context summed across all participants is depicted in Fig. 4D. Similarly, the number of electrodes selected at the flicker frequencies in the unrhythmic context for all participants is shown in Fig. 6D.

The SSVEPs at each condition were averaged across the chosen electrodes. In the rhythmic context, SSVEP averages were calculated at the modulate 1f (2.07 Hz and 2.61 Hz), modulate 2f (4.14 Hz and 5.22 Hz), and flicker frequencies (12 Hz and 15 Hz). To enhance statistical power, the two modulate or flicker frequencies were collapsed to focus on the overall effects. Note that in contrast to the behavioral and ERP analysis that can only measure the audiovisual congruency effects on the attended side, the neural responses from both attended and unattended sides can be measured using frequency-tagging. The SSVEPs at the modulate 1f, modulate 2f, and flicker frequency were analyzed using three-way repeated-measures ANOVAs with factors of attended side (left and right), selective attention (attended and unattended), and audiovisual temporal congruency (the tone congruent or incongruent with the disc). Since neither the main effect of the attended side nor its interactions with other variables were significant, we collapsed the data across the two attended sides for subsequent analyses. Consequently, two-way repeated-measures ANOVAs were conducted for the SSVEPs at the modulate 1f, modulate 2f, and flicker frequency, with selective attention and audiovisual temporal congruency as independent variables. In the unrhythmic context, where no fixed modulate frequencies were present, the same two-way repeated-measures ANOVA was applied only to the SSVEP at the flicker frequency, collapsed across the attended side (no significant effects related to the attended side were observed).

2.6.5. Inter-trial phase coherence

The inter-trial phase coherence (ITC) was computed based on the Fourier transforms of the artifact-free 8-s trial epochs described in the above section. It is calculated as:

$$\mathrm{ITC}(f) = \left| \frac{1}{N} \sum_{n=1}^{N} \frac{c_n(f)}{|c_n(f)|} \right| \quad (1)$$

where $c_n(f)$ is the complex Fourier coefficient of trial n at frequency f and $|.|$ indicates the absolute value. ITC, as another metric of steady-state responses (SSRs), has also shown sensitivities to audiovisual temporal congruency (Covic et al., 2017; Keitel et al., 2019). Resembling the SSVEP, the ITC reached maxima at parieto-occipital electrodes, and had peaks at the flicker frequencies, the modulate frequencies and its second order harmonics in the rhythmic context (Fig. 4B), as well as at the flicker frequencies in the unrhythmic context (Fig. 6B).

For each participant and each condition, the ITCs were first averaged across the chosen electrodes among the 19 posterior ones (refer to the *Steady-state visual evoked potential* section). Similarly, we also collapsed the two specific frequencies at the modulate and flicker frequencies. In the rhythmic context, the collapsed ITCs at the modulate 1f, modulate 2f, and flicker frequency were compared using three-way repeated measures ANOVA, incorporating attended side, selective attention, and audiovisual temporal congruency as independent variables. We also collapsed the ITC across attended sides, since it had no influential effects on the other two variables of interest. Subsequently, a two-way repeated measures ANOVA were performed, considering selective attention and audiovisual temporal congruency.

2.6.6. Correlations between RT and SSR measures

To examine whether the behavioral observed RT facilitation by audiovisual temporal congruency can be predicted by the SSR measures, we calculated their correlations. First, we computed the normalized congruency effect for RT (ΔRT),

$$\Delta\mathrm{RT} = \frac{RT_{asyn} - RT_{syn}}{RT_{asyn} + RT_{syn}} \quad (2)$$

which reflects normalized RT differences when auditory stream temporally congruent with the attended visual stream compared to unattended one. Then, we calculated two SSR congruency effects, one for the SSVEP (ΔPower) and the other one for the ITC (ΔITC), which reflect the differential impact of temporal congruency on the attended and unattended visual streams. The formulas are as follows,

$$\Delta\mathrm{Power} = \left(Power_{attend,con} - Power_{attend,incon}\right) - \left(Power_{unattend,con} - Power_{unattend,incon}\right) \quad (3)$$

$$\Delta\mathrm{ITC} = \left(ITC_{attend,con} - ITC_{attend,incon}\right) - \left(ITC_{unattend,con} - ITC_{unattend,incon}\right) \quad (4)$$

We then correlated ΔPower and ΔITC at the modulate 1f, modulate 2f, and flicker frequency with ΔRT in the rhythmic context, and ΔPower and ΔITC at the flicker frequency with ΔRT in the unrhythmic context. We expected a positive correlation between RTs and SSR measures, which denotes the larger the SSR congruency effects for the attended versus unattended visual streams, the shorter the RTs when the attended visual stream is accompanied by a temporally congruent sound.

2.6.7. Statistics

Post-hoc analysis was performed for significant interactions found in all the ANOVAs on behavioral, SSVEP or ITC data. In addition, Bayesian equivalence tests were conducted to assess statistical evidence using Bayes factors (*BFs*), including BF_{Incl} and BF_{10}. Specifically, BF_{Incl} assesses evidence favoring models including relative to excluding specific effects in the ANOVAs, while BF_{10} quantifies evidence in favor of the alternative hypothesis (*H1*) relative to the null hypothesis (*H0*) in paired-sample comparisons and correlational analyses. Following established conventions, BF_{Incl} (or BF_{10}) values below 1/10 were taken as strong evidence favoring models excluding the effect (or *H0*), values below 1/3 were taken as moderate evidence for exclusion (or *H0*), values between 1/3 and 1 as anecdotal evidence for exclusion (or *H0*), values between 1 and 3 as anecdotal evidence favoring models including the effect (or *H1*), values between 3 and 10 as moderate evidence for inclusion (or *H1*), and values greater than 10 as strong evidence for inclusion (or *H1*) (Lee and Wagenmakers, 2013; Peter Rosenfeld and Olson, 2021; Quintana and Williams, 2018). All the statistical analysis was performed in MATLAB and JASP (version 0.18.1.0, https://jasp-stats.org/). In the Results section, we reported all the *BF* values, but interpreted and concluded only relying on moderate or strong evidence.

3. Results

3.1. A temporally congruent sound modulates the reaction time to attended visual targets in rhythmic but not in unrhythmic contexts

Participants selectively attended lateral visual streams while simultaneously listening to a pitch-modulated sound. They were required to respond as quickly as possible to deviants embedded in either the attended visual or auditory streams while ignoring deviants in the unattended visual stream. The descriptive results of their RT and ACC are listed in Table 1.

Table 1
Average behavioral performance ($M \pm$ SEM).

Rhythm	Rhythmic($N = 23$)		Unrhythmic($N = 25$)	
Temporal congruency	Congruent	Incongruent	Congruent	Incongruent
Visual ACC (%)	82.2 ± 3.1%	77.7 ± 3.4%	81.5 ± 3.4%	85.3 ± 2.9%
Visual RT (ms)	618 ± 15	648 ± 18	610 ± 13	603 ± 11
Auditory ACC (%)	96.4 ± 1.7%	95.9 ± 2.1%	98.4 ± 0.9%	99.1 ± 0.5%
Auditory RT (ms)	559 ± 15	560 ± 13	543 ± 13	537 ± 12

M = mean; SEM = standard error of the mean.

A two-way mixed-design ANOVA on visual RTs revealed a significant interaction between rhythmicity and audiovisual temporal congruency (F (1,46) = 6.704, p = 0.013, η_p^2 = 0.127, BF_{Incl} = 3.917), with no significant main effects of rhythmicity (F (1,46) = 2.004, p = 0.164, η_p^2 = 0.042, BF_{Incl} = 0.843) and audiovisual temporal congruency (F (1,46) = 2.652, p = 0.110, η_p^2 = 0.055 BF_{Incl} = 0.521), illustrated in Fig. 2. Post-hoc analysis showed that visual RTs were significantly shorter when auditory stream was congruent with the attended visual stream, but only in the rhythmic context (t (22) = −3.043, p = 0.006, Cohen's d = −0.634, BF_{10} = 7.582). In contrast, no significant effect of temporal congruency was observed in the unrhythmic context (t (24) = 0.670, p = 0.509, Cohen's d = 0.134, BF_{10} = 0.259). These findings indicate that audiovisual temporal congruency modulates the tracking of continuous, rhythmic rather than unrhythmic visual stream. Additionally, when audiovisual streams were temporally congruent, there was no significant difference in RTs between rhythmic and unrhythmic conditions (t (46) = −0.433, p = 0.667, Cohen's d = 0.125, BF_{10} = 0.311). However, when they were temporally incongruent, RTs in the rhythmic condition were significantly longer than in the unrhythmic condition (t (46) = 2.179, p = 0.035, Cohen's d = 0.629, BF_{10} = 1.899).

Two-way repeated-measures ANOVAs conducted on the visual ACC, auditory ACC and auditory RT, revealed no significant main effects of temporal congruency or rhythmicity, and interaction effects (Fs < 2.960, ps > 0.092, BF_{Incl} < 0.964; Table 1), suggesting that both visual and auditory accuracy, as well as auditory RT, were basically comparable across conditions.

3.2. Early visual processing is modulated by selective attention but unaffected by audiovisual temporal congruency

Previous studies have demonstrated that directing attention to one side induces lateralization of early visual ERPs in the parieto-occipital regions, with greater amplitudes in the hemisphere contralateral to the attended side (Heinze et al., 1994; Hopfinger and West, 2006; Luck et al., 1990; Störmer et al., 2009). To investigate whether temporally congruent auditory streams differentially influence the attended and unattended visual processing at stimulus onset, we compared the contralateral and ipsilateral ERPs at the same region and examined whether these differences varied under temporally congruent versus incongruent conditions in rhythmic and unrhythmic contexts, respectively.

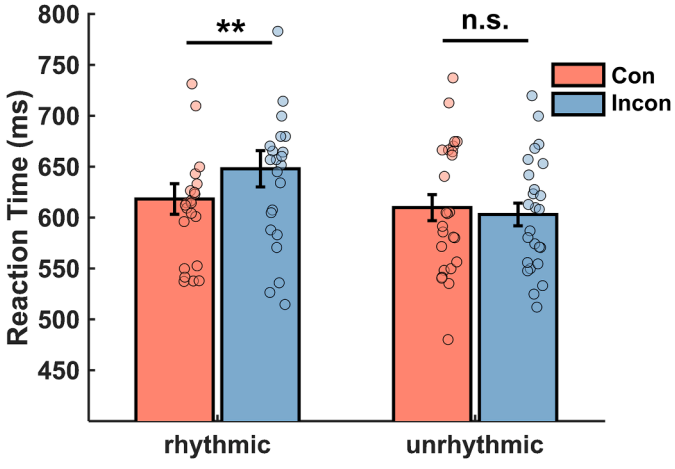


Fig. 2. Behavioral RT results in rhythmic and unrhythmic contexts. Each circle represents the RT from one participant. Error bars showed standard errors of the mean. ** p< 0.01. n.s., non-significant.

Fig. 3 illustrated the three identified visual ERP components, P1 (100–140 ms), N1 (160–200 ms), and P2 (240–280 ms), recorded contralateral and ipsilateral to the attentional cue in both rhythmic (Fig. 3A) and unrhythmic (Fig. 3C) contexts. We averaged the amplitudes of these components in contralateral and ipsilateral electrodes for temporally congruent and incongruent conditions. Two-way repeated-measures ANOVAs revealed a significant main effect of selective attention for P1 and N1 in both rhythmic and unrhythmic contexts (rhythmic P1: F (1,24) = 8.121, p = 0.009, η_p^2 = 0.253, BF_{Incl} = 6.517; rhythmic N1: F (1,24) = 8.645, p = 0.007, η_p^2 = 0.265, BF_{Incl} = 6.697; unrhythmic P1: F (1,22) = 8.084, p = 0.009, η_p^2 = 0.269, BF_{Incl} = 5.660; unrhythmic N1: F (1,22) =7.275, p = 0.013, η_p^2 = 0.248, BF_{Incl} = 4.532, Fig. 3B and Fig. 3D), indicating the P1 and N1 waveform recorded contralateral to the attended side exhibited greater positivity than the ipsilateral waveform. However, no significant main effect of audiovisual temporal congruency and interaction were observed for the two components in either rhythmic or unrhythmic contexts (Fs < 2.412, ps > 0.133, BF_{Incl} < 0.916). No significant effect was found for P2 (Fs < 3.844, ps > 0.063, BF_{Incl} < 1.448). These results demonstrate that selective attention modulates early neural responses in the visual cortex in both rhythmic and unrhythmic contexts. However, the temporally congruent sound may not influence the initial stages of visual processing, suggesting that more time is required for audiovisual interaction to occur in continuous streams.

3.3. Selective attention and audiovisual temporal congruency independently and interactively modulate the SSVEPs in the rhythmic context

Using frequency-tagging methods, we obtained two SSVEPs in the rhythmic context, one at the shape-modulation frequency, the other at the flicker frequency. The average SSVEPs at tagging frequencies across posterior electrodes are shown in Fig. 4A. In addition to the modulate 1f and flicker frequencies, observable SSVEPs were also recorded at the harmonics of the modulate frequency (2f). Consistently, grand-average topographies illustrate maximal modulate and flicker-driven SSVEP amplitudes at parieto-occipital sites (Fig. 4C). The average number of electrodes chosen for statistics at the modulate and flicker frequencies in the rhythmic context across participants were displayed in Fig. 4D.

Two-way repeated measures ANOVAs (selective attention × audiovisual temporal congruency) were conducted at the flicker, modulate 1f and modulate 2f frequencies. First, the results demonstrated that SSVEPs at the flicker frequency (Fig. 5A) were significantly higher for attended sequences compared to unattended sequences (F (1, 22) = 87.905, p < 0.001, η_p^2 = 0.800, BF_{Incl} = 2.017 × 10^6), and significantly higher when there was a temporally congruent auditory stream relative to an incongruent one (F (1, 22) = 5.503, p = 0.028, η_p^2 = 0.200, BF_{Incl} = 0.799). Interestingly, a significant interaction between selective attention and audiovisual temporal congruency was observed (F (1, 22) = 6.763, p = 0.016, η_p^2 = 0.235, BF_{Incl} = 8.920). The temporally congruent auditory stream significantly enhanced the SSVEP at the unattended side compared to an incongruent one (t(22) = 3.659, p = 0.001, Cohen's d = 0.763, BF_{10} = 26.767), whereas it did not influence the SSVEP at the attended side (t(22) = −0.526, p = 0.604, Cohen's d = −0.110, BF_{10} = 0.248), as illustrated in Fig. 5A.

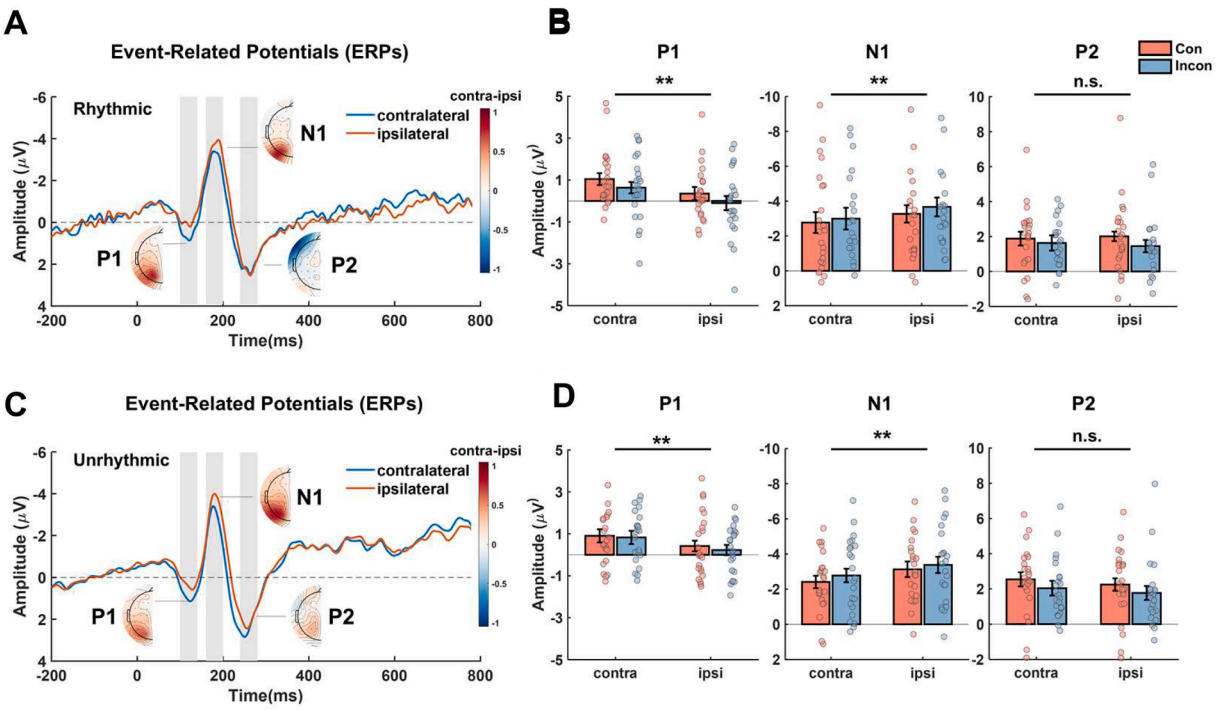


Fig. 3. The results of lateralized ERPs. Average ERP waveforms across electrodes contralateral and ipsilateral to the attended side were plotted for the rhythmic context (A) and unrhythmic context (C). The topography of the P1, N1, and P2 components were drawn alongside the time course. The shaded gray areas mark the time windows corresponding to the P1 (100–140 ms), N1 (160–200 ms), and P2 (240–280 ms) components. The electrodes of interest included P7/8, P5/6, P3/4, PO7/8, and PO3/4. (B) and (D) The average amplitudes of P1, N1, and P2 across the contralateral and ipsilateral electrodes to the attended side were shown for temporally congruent and incongruent conditions in the rhythmic context (B) and unrhythmic context (D), respectively. Error bars showed standard errors of the mean. ** $p < 0.01$. n.s., non-significant. The circles denoted individual data points.

Second, the results showed no significant SSVEP differences across conditions at the modulate 1f frequency ($Fs < 1.343$, $ps > 0.266$, $BF_{Incl} < 0.691$, Fig. 5C). However, at the modulate 2f frequency, SSVEPs were significantly enhanced for attended compared to unattended condition (F (1, 14) = 52.921, $p < 0.001$, $\eta_p^2 = 0.791$, $BF_{Incl} = 395.596$, Fig. 5E), and were significantly higher for temporally congruent than incongruent conditions (F (1, 14) = 4.983, $p = 0.042$, $\eta_p^2 = 0.263$, $BF_{Incl} = 1.485$). No interaction between selective attention and audiovisual temporal congruency was found (F (1, 14) = 0.398, $p = 0.538$, $\eta_p^2 = 0.028$, $BF_{Incl} = 0.955$). The anecdotal Bayesian evidence for the significant congruency effect ($BF_{Incl} < 3$) might be due to the relatively small sample size after strictly excluding the participants without robust SSVEPs at the modulate 2f frequency. However, it was consistent with previous reports (Covic et al., 2017; Keitel and Mueller, 2015).

3.4. Selective attention and audiovisual temporal congruency independently modulate the ITCs in the rhythmic context

Similar two-way repeated measures ANOVAs (selective attention × audiovisual temporal congruency) were conducted at the flicker, modulate 1f, and modulate 2f frequencies. At the flicker frequency, the ITC was significantly higher for attended sequences compared to unattended sequences (F (1, 22) = 136.542, $p < 0.001$, $\eta_p^2 = 0.861$, $BF_{Incl} = 1.166 \times 10^8$, Fig. 5B). However, neither the main effect of audiovisual temporal congruency nor the interaction reached significance ($Fs < 0.009$, $ps > 0.925$, $BF_{Incl} < 0.291$). Similar results were obtained at modulate 1f frequency. There was a significant main effect of selective attention (F (1, 14) = 13.526, $p = 0.002$, $\eta_p^2 = 0.491$, $BF_{Incl} = 18.483$, Fig. 5D). Neither the main effect of audiovisual temporal congruency nor the interaction effect was significant ($Fs < 0.443$, $ps > 0.517$, $BF_{Incl} < 0.364$). At the modulate 2f frequency, in addition to the significant main effect of selective attention (F (1, 14) = 31.298, $p < 0.001$, $\eta_p^2 = 0.691$, $BF_{Incl} = 4177.197$), temporally congruent auditory streams elicited significantly higher ITC relative to incongruent ones (F (1, 14) = 5.445, $p = 0.035$, $\eta_p^2 = 0.280$, $BF_{Incl} = 1.791$, Fig. 5F). No significant interaction was observed either (F (1, 14) = 2.614, $p = 0.128$, $\eta_p^2 = 0.157$, $BF_{Incl} = 0.383$). Similar to the SSVEP congruent effect at the modulate 2f frequency, the ITC congruency effect was consistent with one previous study (Covic et al., 2017), albeit with anecdotal Bayesian evidence.

The SSVEP and ITC results together demonstrated that a temporally congruent auditory stream modulates the visual processing orthogonal to audiovisual temporal congruency (indicated by the flicker-evoked SSVEPs) interactively with selective attention, with larger congruency effect for unattended visual processing. Conversely, it modulates the visual processing relevant to audiovisual temporal congruency (indicated by the shape-modulation-evoked SSVEPs and ITCs) independently of selective attention.

3.5. No enhancement of audiovisual temporal congruency on the SSVEPs and ITCs in the unrhythmic context

Since the pitch-shape modulation was irregular in the unrhythmic context, the frequency tagging analysis could only be performed at the flicker frequency. A summary of the SSR results in the unrhythmic context was shown in Fig. 6, with the average SSVEPs and ITC at the flicker frequencies across posterior electrodes shown in Fig. 6A and 6B, the grand-average topographies in Fig. 6C, and the average number of electrodes chosen at the flicker frequencies across participants in Fig. 6D. As Fig. 6E and 6F illustrated, the main effects of selective attention remained significant for the SSVEP (F (1, 21) = 27.726, $p < 0.001$, $\eta_p^2 = 0.569$, $BF_{Incl} = 356.509$) and ITC (F (1, 21) = 32.271, $p < 0.001$, $\eta_p^2 = 0.640$, $BF_{Incl} = 2774.912$). However, neither the main effects of audiovisual temporal congruency nor their interactions were

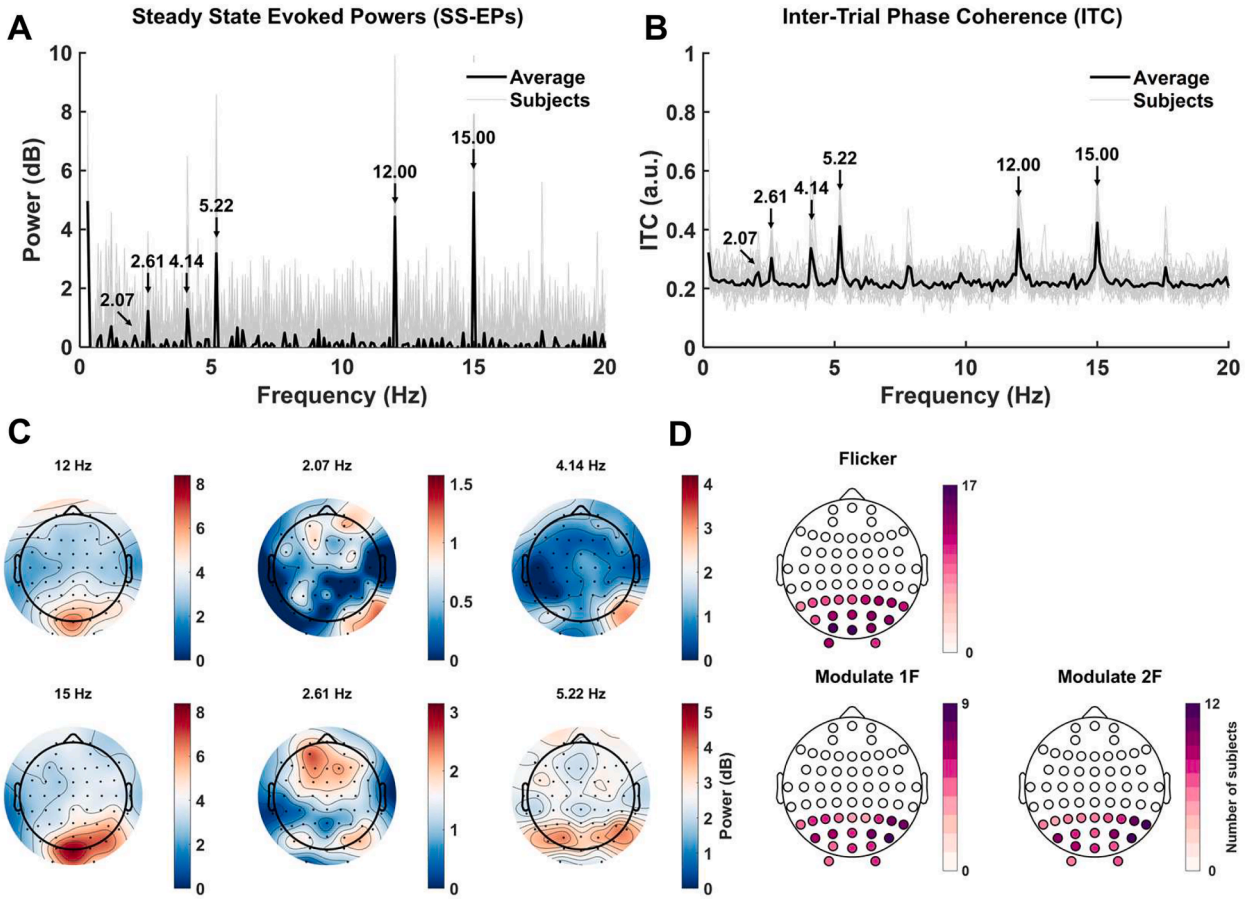


Fig. 4. An illustration of the SSRs in the rhythmic context. Normalized power spectra (A) and ITC spectra (B) averaged across posterior electrodes, conditions and participants. Arrows indicate modulate 1f (2.07 Hz and 2.61 Hz), modulate 2f (4.14 Hz and 5.22 Hz) and flicker frequency (12 Hz and 15 Hz), all demonstrating obviously higher power and ITC compared to other frequencies. (C) Topographic scalp of the SSVEPs at the tagging frequencies, the modulate 1f, the modulate 2f, and the flicker frequencies. (D) The number of electrodes selected for SSVEP and ITC statistical analysis summed across participants, drawn with color gradient.

significant ($Fs < 1.995$, $ps > 0.173$, $BF_{Incl} < 0.468$).

3.6. Rhythmicity modulates the influence of audiovisual temporal congruency on visual processing

It appeared that audiovisual temporal congruency selectively modulated flicker-evoked SSVEPs interactively with selective attention exclusively in rhythmic contexts. To confirm the role of rhythmicity, we combined the SSVEP data at the flicker frequency from both rhythmic and unrhythmic contexts and carried out a three-way repeated measures ANOVA (rhythmicity × selective attention × audiovisual temporal congruency). As expected, the results revealed a significant three-way interaction (F (1, 43) = 4.087, $p = 0.049$, $\eta_p^2 = 0.087$, $BF_{Incl} = 5.726$), supporting that the temporal structure of audiovisual stream indeed affected whether a congruent auditory stream selectively enhances the unattended SSVEPs compared to an incongruent one. The results further revealed that continuous audiovisual interaction depends not only on moment-to-moment alignment, but also on the temporal regularity of the stimulus streams.

Other significant and marginal significant effects included the main effect of selective attention (F (1, 43) = 101.288, $p < 0.001$, $\eta_p^2 = 0.702$, $BF_{Incl} = 3.283 \times 10^9$), the interaction between rhythmicity and audiovisual temporal congruency (F (1, 43) = 6.640, $p = 0.013$, $\eta_p^2 = 0.134$, $BF_{Incl} = 2.881$), and the interaction between rhythmicity and selective attention (F (1, 43) = 3.626, $p = 0.064$, $\eta_p^2 = 0.078$, $BF_{Incl} = 2.008$). All the remaining effects were not significant ($Fs < 2.354$, $ps > 0.132$, $BF_{Incl} < 1.110$).

3.7. Behavioral-neural correlations in rhythmic and unrhythmic contexts

We then assessed whether RTs facilitation driven by audiovisual temporal congruency was associated with related SSR measures (Fig. 7A). Specifically, we normalized the RT reduction for congruent versus incongruent conditions (ΔRT), and the SSR congruent effects for attended versus unattended streams (ΔPower and ΔITC). Only the correlation between ΔRT and ΔITC at the flicker frequency in the rhythmic context was significant and with moderate or stronger Bayesian evidence ($r = 0.53$, $p = 0.013$, $BF_{10} = 4.975$). Other significant or marginally significant correlations but with anecdotal Bayesian evidence included a positive one between ΔRT and ΔITC at modulate 1f frequency ($r = 0.54$, $p = 0.039$, $BF_{10} = 2.231$), a positive one between ΔRT and ΔPower at modulate 2f frequency ($r = 0.56$, $p = 0.039$, $BF_{10} = 2.295$), and a positive one between ΔRT and ΔPower at the flicker frequency ($r = 0.42$, $p = 0.055$, $BF_{10} = 1.520$), all in the rhythmic context. These results indicate that, in the rhythmic context, sound-induced shortening of reaction time to the attended targets is in proportion to the contrast between ITC congruency effects for the attended and the unattended streams, especially at the flicker frequency. Such correlation disappeared in the unrhythmic context (all $ps > 0.537$, $BF_{10} < 0.316$, Fig. 7B), further underscoring the critical role of rhythmicity in facilitating audiovisual interaction.

4. Discussion

Previous studies have demonstrated that both selective attention and audiovisual temporal congruency can enhance visual processing;

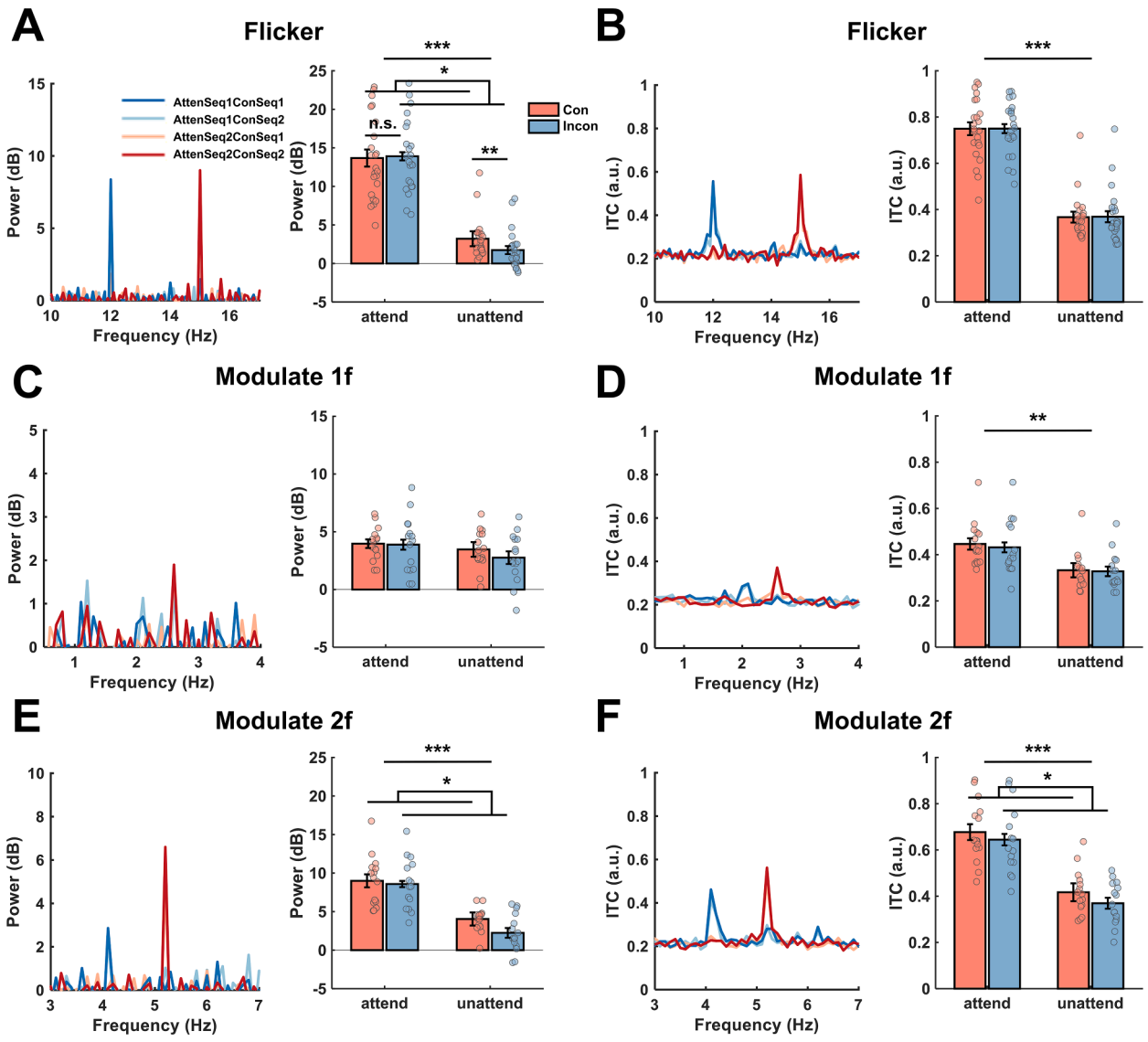


Fig. 5. The SSVEP and ITC results in rhythmic context. The power spectra and the SSVEP at the flicker frequency (A), at the modulate 1f frequency (C), and at the modulate 2f frequency (E) under different conditions. The ITC spectra and the ITC at the flicker frequency (B), at the modulate 1f frequency (D), and at the modulate 2f frequency (F) under different conditions. The attended and congruent sequences in the power and ITC spectra panels were labeled following the rule that indicates which sequence is attended and congruent with. For instance, AttenSeq1ConSeq1 indicates that sequence 1 was attended and temporally congruent with. Audiovisual temporal congruency was defined as temporal alignment between the tone and the disc at the tagged frequency. Error bars showed standard errors of the mean. *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$. n.s., non-significant. The circles indicate individual data points.

however, the relationship between these two processes remains controversial (Ahmed et al., 2023; Alsius et al., 2014, 2005; Bertelson et al., 2000; Covic et al., 2017; Keitel and Mueller, 2015; Spence and Driver, 2000; Talsma et al., 2007, 2010; Vroomen et al., 2001a, 2001b). The present findings revealed that selective attention and audiovisual temporal congruency simultaneously exert independent and interactive influences on visual processing, relying on two critical factors: the relevance of visual features to audiovisual temporal congruency and the temporal regularity of continuous audiovisual streams. First, selective attention and audiovisual temporal congruency independently enhanced SSVEPs and ITCs at the 2nd harmonics of the shape modulation (features relevant to audiovisual temporal congruency). Second, they interactively enhanced SSVEPs at the flicker frequency (features orthogonal to audiovisual temporal congruency), with audiovisual temporal congruency significantly boosting responses for unattended sequences but not for attended sequences. Third, larger flicker-induced ITC congruency effects for attended compared to unattended streams were associated with shorter RTs to deviants in the attended streams. Lastly, all observed crossmodal influences were restricted to rhythmic contexts with regularly structured audiovisual streams.

4.1. The impact of audiovisual temporal congruency on the visual deviants detection

Extensive research has demonstrated that temporally congruent visual or auditory inputs can enhance sensory processing in another modality (Ahmed et al., 2023; Atilgan et al., 2018; Haider et al., 2024; Iordanescu et al., 2008; Kvasova et al., 2019; Maddox et al., 2015; Park et al., 2016; Peng et al., 2023; Reisinger et al., 2025; Shen et al., 2023a; Van der Burg et al., 2008, 2011). Consistently, in our study the reaction times shortened for detection of deviants in the attended stream when a temporally congruent tone was played compared to an incongruent one. Notably, the finding appears in stark contrast with previous studies employing the same paradigm (Covic et al., 2017; Keitel and Mueller, 2015) which showed that when a tone's pitch changed congruently with the spatial-frequency change of an attended Gabor, deviant detection

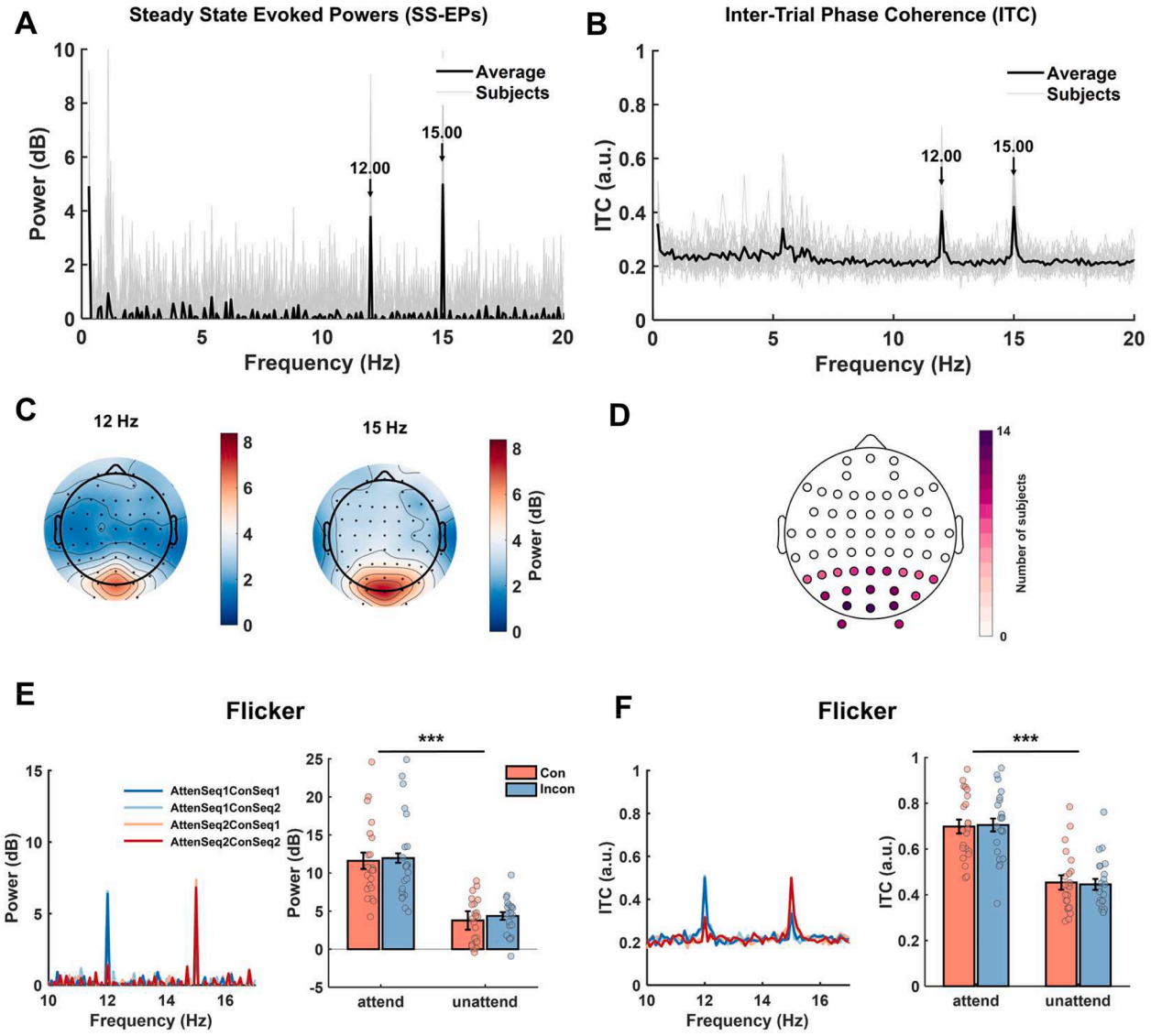


Fig. 6. The SSVEP and ITC results in unrhythmic context. (A)(B) depict the normalized power spectra and ITC spectra averaged across posterior electrodes, conditions and participants respectively, in which the arrows point at the flicker frequencies. (C) Topographic scalp of SSVEPs at the flicker frequency. (D) The number of electrodes selected for the SSVEP and ITC statistical analysis summed across participants. (E) The power spectra and the SSVEP at the flicker frequency under different conditions. (F) The ITC spectra and the ITC at the flicker frequency under different conditions. Error bars showed standard errors of the mean. *** $p < 0.001$. The circles indicate individual data points.

RTs increased relative to temporally incongruent tones.

Several key differences between their studies and ours may account for these divergent findings. First, they used pitch-spatial-frequency audiovisual streams, whereas we used pitch-shape ones. As stated in the introduction, this coupling may be more ideal for an underlying neural enhancement to occur (also see Section 4.2 in Discussion). Second, the tagging frequencies differed. Covic et al. (2017), Keitel and Mueller (2015) used flicker frequencies of ~14/17 Hz and modulate frequencies at 3.14/3.62 Hz, whereas our study employed lower flicker (12/15 Hz) and modulate (2.07/2.61 Hz) frequencies. Both modulate frequencies fall below the ~4 Hz limit for crossmodal temporal congruency discrimination (Fujisaki and Nishida, 2005, 2010). However, the lower frequencies used in our study may facilitate easier discrimination, and improved temporal congruency discrimination has been shown to be requisite for audiovisual interaction with dynamic stimuli in competitive settings (Atilgan and Bizley, 2021). Third, the task design differed. We asked participants to concurrently attend to the auditory stream and perform a deviant detection task within it, whereas they assigned no task to the auditory modality. Consequently, their participants may have devoted less attention to the task-irrelevant modality. Prior studies have shown that attention to a task-irrelevant modality can boost audiovisual interactions (Talsma et al., 2007; van Ee et al., 2009), and some studies revealing crossmodal facilitation have employed this approach (Atilgan and Bizley, 2021; Kim et al., 2012; Maddox et al., 2015).

It is worth noting that the RT reduction here may reflect either facilitation of visual processing at the attended side, distraction at the unattended side, or both, as the tone temporally congruent with the attended stream was necessarily incongruent with the unattended stream (and vice versa) in our study. Currently, we cannot determine which effect provides the dominant explanation for the observed RT reduction without a baseline, such as a neutral tone temporally incongruent with both the attended and unattended visual streams or a vision-only condition. However, two observations suggest that the shortened detection RT is not merely due to the congruency effect at the attended side. First, the deviants were defined by sudden luminance changes of

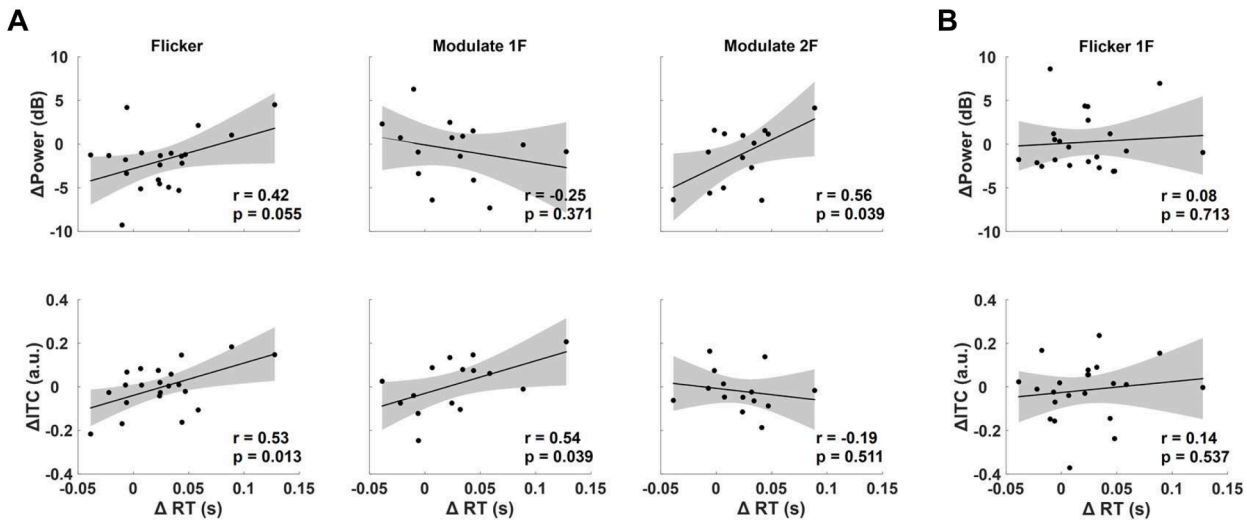


Fig. 7. Behavioral-neural correlations. (A) and (B) represent the correlations between ΔRT and ΔPower, ΔITC in the rhythmic and unrhythmic contexts, respectively. The shaded region reflects the 95% confidence interval. The circles denoted individual data points.

visual features orthogonal to audiovisual temporal congruency. The flicker-induced SSVEPs, which served as the neural index of orthogonal visual processing, were selectively enhanced by temporally congruent sound at the unattended side rather than at the attended side. Second, the RT reduction correlated with the ITC congruency effects between the attended versus the unattended side, indicating that neuro-behavioral prediction requires a combination of neural indices for both sides.

4.2. The impact of selective attention and audiovisual temporal congruency on visual processing: features relevant or orthogonal to audiovisual temporal congruency matter

In rhythmic contexts, temporally congruent sound significantly enhanced the SSVEP and ITC in response to visual features relevant to audiovisual temporal congruency (shape modulation) independent of selective attention. Although the significant effects were supported by anecdotal Bayesian evidence (see Results), they were highly consistent with findings by Covic et al. (2017) and Keitel and Mueller (2015), who also showed that enhanced SSVEPs and ITCs by auditory temporal congruency at the 2nd harmonic of spatial-frequency modulation independently at both the attended and unattended side. On the other hand, audiovisual temporal congruency enhanced the SSVEP tagging to unattended rather than attended visual features of the same object while orthogonal to temporal congruency (flicker), demonstrating an interactive impact of audiovisual temporal congruency and selective attention on visual processing. The results align with the crossmodal binding framework proposed by Bizley et al. (2016), demonstrating that temporal coherence between co-varying auditory and visual features facilitates their binding into a crossmodal object, which further enhances representations of the orthogonal features belonging to the objecthood (Atilgan et al., 2018; Maddox et al., 2015). Moreover, the results reveal that the relationship between audiovisual temporal congruency and selective attention is not simply dichotomous as either independent or interactive; rather, both of them can coexist within a single experiment, depending on whether the stimulus feature is relevant to audiovisual temporal congruency or not; see Peng et al. (2023) for a similar finding on speech.

Evidence regarding whether SSVEPs can index audiovisual interaction for features orthogonal to temporal congruency has been mixed (Covic et al., 2017; Giani et al., 2012; Keitel and Mueller, 2015; Nozaradan et al., 2012; Sciortino and Kayser, 2023). Recently, a study replicated and demonstrated that enhancement of flicker-induced SSVEPs by sound may rely on the tightness of audiovisual correspondence: pitch-size rather than pitch-hue and pitch-saturation correspondences enables robust SSVEP enhancement (Sciortino and Kayser, 2023). Although the study used static rather than dynamic audiovisual stimuli, it inspired us to fine-tune the experimental paradigm by adopting a pitch-shape instead of a pitch-spatial-frequency co-variation to establish temporal congruency. As expected, our results successfully demonstrated auditory modulation of flicker-induced SSVEPs in rhythmic contexts. The power spectra around the tagging frequencies were also much narrower in Nozaradan et al. (2012) (their Fig. 2) and ours (Figs. 4A and 6A) than in those studies that failed to find such effects (Covic et al., 2017; Keitel and Mueller, 2015). This finding extends Nozaradan et al. (2012)'s work by showing that crossmodal facilitation can spread from relevant to orthogonal visual features in a competitive context. Furthermore, it offered another choice of dynamic audiovisual streams that could successfully evoke SSVEP enhancement, supplementing the auditory loudness-visual position coupling they used. We posit that dynamic changes in shape, or position, when paired with auditory pitch or loudness, may create periodic audiovisual "beats" that strengthen crossmodal binding, thereby facilitating the spread of modulation to orthogonal visual features. Additionally, it is necessary to reiterate that the aforementioned three paradigm differences between (Covic et al., 2017; Keitel and Mueller, 2015) and ours (Section 4.1) may also contribute to the distinction in flicker-induced SSVEPs. Together, these results highlight the importance of feature selection in revealing crossmodal interactions using SSVEPs.

The interaction between selective attention and audiovisual temporal congruency observed in flicker-induced SSVEPs contrasts with previous reports that attention typically boosts audiovisual interaction (Alsius et al., 2005; Fairhall and Macaluso, 2009; Seijdel et al., 2024; Senkowski et al., 2005; Talsma and Woldorff, 2005). However, our findings are consistent with prior findings that concurrent sounds more strongly modulate the processing of unattended visual stimuli (Van der Stoep et al., 2015; Zou et al., 2012), and enhance the SSVEP induced by ignored visual speech (Krause et al., 2012; Senkowski et al., 2008). In the framework proposed by Talsma et al. (2010), sparse auditory stimuli in the task-irrelevant modality are inherently more salient and can facilitate processing of temporally congruent visual stimuli in cluttered settings, thereby enhancing its ability to attract bottom-up attention (Talsma et al., 2010). To fully account for our results, however, two extensions to Talsma's framework are needed. First, it should be integrated with Bizley's crossmodal binding framework to explain the spread of crossmodal influences from features relevant to audiovisual temporal congruency to orthogonal features (Bizley et al., 2016).

Second, it should be combined with the well-known principle of inverse effectiveness in multisensory integration to explain why flicker-induced SSVEP enhancement occurred only at the unattended side. According to this principle, weaker unimodal stimuli are more susceptible to crossmodal influence (Noesselt et al., 2010; Senkowski et al., 2011; Stein et al., 2020). Since unattended stimuli—which are selectively suppressed from further processing—evoke weaker neural responses compared to attended stimuli (Covic et al., 2017; Heinze et al., 1994; Hopfinger and West, 2006; Keitel and Mueller, 2015; Luck et al., 1990), the unattended stream gains more from audiovisual interaction. Overall, for dynamic stimuli, the interplay between selective attention and multisensory interaction depends on whether features are orthogonal to those that establish audiovisual temporal congruency, and is governed by the principle of inverse effectiveness. Noteworthy, we cannot completely rule out an alternative possibility: attentional modulation may have been sufficiently strong to produce a ceiling effect, thereby limiting any additional gain from audiovisual interaction.

4.3. The critical role of rhythmicity in modulating the impact of audiovisual temporal congruency on visual processing

In line with previous findings (Heins et al., 2021; Marchant et al., 2012; ten Oever et al., 2014), our results underscore the importance of rhythmicity for crossmodal interactions between continuous stimuli, and further suggest that physical temporal congruency alone is insufficient for effective crossmodal interaction. A rhythmic (regular) temporal structure is also necessary. Rhythm enables synchronization of neural oscillations in the brain—particularly in low-frequency bands—to the external continuous stimulus stream (Bauer et al., 2020; Shen et al., 2025). Presenting task-irrelevant rhythmic sound modulates neural synchronization in visual cortices, thereby modulating the visual responses (Bauer et al., 2021; Yuan et al., 2021). Our findings are consistent with this phase alignment explanation. For example, only rhythmic auditory streams presented temporally congruent with the visual streams yielded higher ITC in parieto-occipital visual areas at the second harmonics of modulate frequency. Moreover, the differential ITC congruency effects at the attended and unattended flicker predicted the RT reduction. In contrast, the lack of regularity in unrhythmic audiovisual streams may prevent phase alignment of neural oscillations in the visual cortex to the external stimulus stream, thereby hindering crossmodal interaction.

Some quasi-rhythmic stimuli, especially speeches, the acoustic spectra of whose acoustic envelopes mainly cover a low frequency band of $<$15 Hz, are sensitive to influence from visual inputs (Giraud and Poeppel, 2012; Klatt et al., 2023; Zion Golumbic et al., 2013). But unlike speech which maintains temporal predictability between successive syllables, the unrhythmic stimulation in the current study dynamically changed its oscillatory frequency, creating a stream with low temporal predictability. Reduced temporal predictability may increase the perceptual difficulty of temporal congruency across continuous streams, thereby nullifying the audiovisual temporal congruency effect. As evidence, previous studies revealed that unrhythmic visual inputs can enhance the processing of attended auditory stimuli after training to improve the sensitivity to audiovisual temporal congruency (Atilgan and Bizley, 2021). Though we did not find significant auditory influence on visual processing in unrhythmic contexts, it remains unclear whether training the sensitivity of audiovisual temporal congruency would restore its effect. In summary, audiovisual streams with regular structures are indispensable for their interactions. As the majority of prior literature employed rhythmic or quasi-rhythmic stimuli (Ahmed et al., 2023; Covic et al., 2017; Haider et al., 2024; Keitel and Mueller, 2015; Nozaradan et al., 2012; Park et al., 2016; Reisinger et al., 2025; Zion Golumbic et al., 2013), few directly compared the audiovisual interaction in rhythmic versus unrhythmic streams (Heins et al., 2021; Marchant et al., 2012; ten Oever et al., 2014). The finding that rhythmicity plays a role in audiovisual interaction imposes an additional constraint, even when attention is fully engaged.

4.4. Limitations

Despite these findings, two limitations should be acknowledged. First, following previous studies (Covic et al., 2017; Keitel and Mueller, 2015; Nozaradan et al., 2012), we maximized the congruency effect by comparing congruent with incongruent audiovisual conditions. However, this design entails a limitation: it cannot determine whether benefits in congruent conditions reflect enhanced processing of the attended stream, reduced distraction from the unattended stream, or both, as discussed earlier. Empirical evidence for the direction and nature of the congruency effect calls for a unimodal condition (visual-only) or a neutral auditory baseline (an auditory stream incongruent with both visual streams). Future studies should incorporate such control conditions to clarify this distinction when adopting similar paradigms. Second, although significant and aligned with prior literature (Covic et al., 2017; Keitel and Mueller, 2015), the Bayesian evidence for some observed effects (e.g., SSVEP/ITC enhancements at modulate 2f) was not robust enough. As discussed, this may stem from the strict screening criteria that ensured included participants indeed had robust neural indices. Nevertheless, replication with larger sample sizes is warranted to confirm the reliability of the current findings.

5. Conclusion

The present study revealed both independent and interactive effects of selective attention and audiovisual interaction on visual processing, contingent upon the relevance of visual features to audiovisual temporal congruency, and the temporal regularity of the audiovisual streams. The findings provide supportive evidence for the crossmodal binding framework (Bizley et al., 2016) from the visual domain, showing that temporal congruency can enhance task-irrelevant features while further highlighting that attention modulates this spread in a feature-specific manner. By taking into consideration the temporal structure of continuous stimulus streams, this work explains why some continuous audiovisual streams fail to show interaction effects and advances our understanding of the complex interplay between selective attention and audiovisual interaction in shaping visual experience. Future studies exploring crossmodal interaction in continuous stimulus streams may benefit from the current findings regarding stimulus selection to maximize observable effects. These implications may also extend to technological domains, particularly in optimizing multisensory integration for virtual reality systems and other human-computer interfaces.

Data availability

All data that support the findings of this study are open access at https://www.scidb.cn/en/s/YRFbM3.

CRediT authorship contribution statement

Jieru Chen: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Wenjie Liu:** Writing – review & editing, Writing – original draft, Formal analysis, Data curation. **Shiqi Tan:** Writing – review & editing, Visualization, Validation, Methodology, Data curation. **Xiangyong Yuan:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Yi Jiang:** Writing – review & editing, Supervision, Software, Resources, Project administration, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgment

This research was supported by grants from the STI2030-Major Projects (2021ZD0203800, 2021ZD0204200), the National Natural Science Foundation of China (32430043, 31600884), the Interdisciplinary Innovation Team (JCTD-2021-06), Youth Innovation Promotion Association of the Chinese Academy of Sciences, the Key Research and Development Program of Guangdong, China (2023B0303010004), and the Fundamental Research Funds for the Central Universities.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2026.121873.

References

Ahmed, F., Nidiffer, A.R., O'Sullivan, A.E., Zuk, N.J., Lalor, E.C., 2023. The integration of continuous audio and visual speech in a cocktail-party environment depends on attention. Neuroimage 274, 120143. https://doi.org/10.1016/j.neuroimage.2023.120143.

Alsius, A., Möttönen, R., Sams, M.E., Soto-Faraco, S., Tiippana, K., 2014. Effect of attentional load on audiovisual speech perception: evidence from ERPs [Original Research]. Front. Psychol. 5. https://doi.org/10.3389/fpsyg.2014.00727.

Alsius, A., Navarra, J., Campbell, R., Soto-Faraco, S., 2005. Audiovisual integration of speech falters under high attention demands. Curr. Biol. 15 (9), 839–843. https://doi.org/10.1016/j.cub.2005.03.046.

Atilgan, H., Bizley, J.K., 2021. Training enhances the ability of listeners to exploit visual information for auditory scene analysis. Cognition 208. https://doi.org/10.1016/j.cognition.2020.104529.

Atilgan, H., Town, S.M., Wood, K.C., Jones, G.P., Maddox, R.K., Lee, A.K.C., Bizley, J.K., 2018. Integration of visual information in auditory cortex promotes auditory scene analysis through multisensory binding. Neuron 97 (3), 640–655. https://doi.org/10.1016/j.neuron.2017.12.034 e644.

Bauer, A.K.R., Debener, S., Nobre, A.C., 2020. Synchronisation of neural oscillations and cross-modal influences [Review]. Trends Cogn. Sci. 24 (6), 481–495. https://doi.org/10.1016/j.tics.2020.03.003.

Bauer, A.K.R., Ede, F.V., Quinn, A.J., Nobre, A.C., 2021. Rhythmic modulation of visual perception by continuous rhythmic auditory stimulation [Article]. J. Neurosci. 41 (33), 7065–7075. https://doi.org/10.1523/JNEUROSCI.2980-20.2021.

Bertelson, P., Pavani, F., Ladavas, E., Vroomen, J., de Gelder, B., 2000. Ventriloquism in patients with unilateral visual neglect. Neuropsychologia 38 (12), 1634–1642. https://doi.org/10.1016/S0028-3932(00)00067-1.

Bizley, J.K., Maddox, R.K., Lee, A.K.C., 2016. Defining auditory-visual objects: behavioral tests and physiological mechanisms [Review]. Trends Neurosci. 39 (2), 74–85. https://doi.org/10.1016/j.tins.2015.12.007.

Brainard, D.H., 1997. The psychophysics toolbox. Spat. Vis. 10 (4), 433–436. https://doi.org/10.1163/156856897X00357.

Covic, A., Keitel, C., Porcu, E., Schroeger, E., Mueller, M.M., 2017. Audio-visual synchrony and spatial attention enhance processing of dynamic visual stimulation independently and in parallel: a frequency-tagging study [Article]. Neuroimage 161, 32–42. https://doi.org/10.1016/j.neuroimage.2017.08.022.

Crosse, M.J., Butler, J.S., Lalor, E.C., 2015. Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. J. Neurosci. 35 (42), 14195–14204. https://doi.org/10.1523/jneurosci.1829-15.2015.

Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Methods 134 (1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009.

Drijvers, L., Jensen, O., Spaak, E., 2020. Rapid invisible frequency tagging reveals nonlinear integration of auditory and visual information. Hum. Brain Mapp. 42. https://doi.org/10.1002/hbm.25282.

Fairhall, S.L., Macaluso, E., 2009. Spatial attention can modulate audiovisual integration at multiple cortical and subcortical sites. Eur. J. Neurosci. 29 (6), 1247–1257. https://doi.org/10.1111/j.1460-9568.2009.06688.x.

Fujisaki, W., Nishida, S., 2005. Temporal frequency characteristics of synchrony-asynchrony discrimination of audio-visual signals [Conference Paper]. Exp. Brain Res. 166 (3–4), 455–464. https://doi.org/10.1007/s00221-005-2385-8.

Fujisaki, W., Nishida, S.Y., 2010. A common perceptual temporal limit of binding synchronous inputs across different sensory attributes and modalities. Proc. R. Soc. B Biol. Sci. 277 (1692), 2281–2290. https://doi.org/10.1098/rspb.2010.0243.

Giani, A.S., Ortiz, E., Belardinelli, P., Kleiner, M., Preissl, H., Noppeney, U., 2012. Steady-state responses in MEG demonstrate information integration within but not across the auditory and visual senses. Neuroimage 60 (2), 1478–1489. https://doi.org/10.1016/j.neuroimage.2012.01.114.

Giraud, A.L., Poeppel, D., 2012. Cortical oscillations and speech processing: emerging computational principles and operations. Nat. Neurosci. 15 (4), 511–517. https://doi.org/10.1038/nn.3063.

Haider, C., Park, H., Hauswald, A., Weisz, N., 2024. Neural speech tracking highlights the importance of visual speech in multi-speaker situations. J. Cogn. Neurosci. 36, 128–142. https://doi.org/10.1162/jocn_a_02059.

Heins, N., Pomp, J., Kluger, D.S., Vinbrüx, S., Trempler, I., Kohler, A., Kornysheva, K., Zentgraf, K., Raab, M., Schubotz, R.I., 2021. Surmising synchrony of sound and sight: factors explaining variance of audiovisual integration in hurdling, tap dancing and drumming. PLoS ONE 16 (7), e0253130. https://doi.org/10.1371/journal.pone.0253130.

Heinze, H.J., Mangun, G.R., Burchert, W., Hinrichs, H., Scholz, M., Münte, T.F., Gös, A., Scherg, M., Johannes, S., Hundeshagen, H., et al., 1994. Combined spatial and temporal imaging of brain activity during visual selective attention in humans. Nature 372 (6506), 543–546. https://doi.org/10.1038/372543a0.

Hopfinger, J.B., West, V.M., 2006. Interactions between endogenous and exogenous attention on cortical visual processing. Neuroimage 31 (2), 774–789. https://doi.org/10.1016/j.neuroimage.2005.12.049.

Iordanescu, L., Guzman-Martinez, E., Grabowecky, M., Suzuki, S., 2008. Characteristic sounds facilitate visual search. Psychon. Bull. Rev. 15 (3), 548–554. https://doi.org/10.3758/PBR.15.3.548.

Keitel, C., Keitel, A., Benwell, C.S.Y., Daube, C., Thut, G., Gross, J., 2019. Stimulus-driven brain rhythms within the alpha band: the attentional-modulation conundrum [Article]. J. Neurosci. 39 (16), 3119–3129. https://doi.org/10.1523/JNEUROSCI.1633-18.2019.

Keitel, C., Mueller, M.M., 2015. Audio-visual synchrony and feature-selective attention co-amplify early visual processing [Article]. Exp. Brain Res. 234 (5), 1221–1231. https://doi.org/10.1007/s00221-015-4392-8.

Kim, R., Peters, M.A., Shams, L., 2012. 0 + 1 >1: how adding noninformative sound improves performance on a visual task. Psychol. Sci. 23 (1), 6–12. https://doi.org/10.1177/0956797611420662.

Kim, Y.J., Grabowecky, M., Paller, K.A., Suzuki, S., 2011. Differential roles of frequency-following and frequency-doubling visual responses revealed by evoked neural harmonics. J. Cogn. Neurosci. 23 (8), 1875–1886. https://doi.org/10.1162/jocn.2010.21536.

Klatt, L.I., Begau, A., Schneider, D., Wascher, E., Getzmann, S., 2023. Cross-modal interactions at the audiovisual cocktail-party revealed by behavior, ERPs, and neural oscillations. Neuroimage 271, 120022. https://doi.org/10.1016/j.neuroimage.2023.120022.

Krause, H., Schneider, T.R., Engel, A.K., Senkowski, D., 2012. Capture of visual attention interferes with multisensory speech processing. Front. Integr. Neurosci. 6 (2012), 67. https://doi.org/10.3389/fnint.2012.00067.

Kvasova, D., Garcia-Vernet, L., Soto-Faraco, S., 2019. Characteristic sounds facilitate object search in real-life scenes. Front. Psychol. 10, 2511. https://doi.org/10.3389/fpsyg.2019.02511.

Lee, M.D., Wagenmakers, E.J., 2013. Bayesian Cognitive Modeling: A Practical Course. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759, 10.1017/CBO9781139087759.

Lenc, T., Keller, P.E., Varlet, M., Nozaradan, S., 2018. Neural tracking of the musical beat is enhanced by low-frequency sounds. Proc. Natl. Acad. Sci. 115 (32), 8221–8226. https://doi.org/10.1073/pnas.1801421115.

Luck, S.J., 2022. Applied Event-Related Potential Data Analysis. LibreTexts. https://doi.org/10.18115/D5QG92.

Luck, S.J., Heinze, H.J., Mangun, G.R., Hillyard, S.A., 1990. Visual event-related potentials index focused attention within bilateral stimulus arrays. II. Functional dissociation of P1 and N1 components. Electroencephalogr. Clin. Neurophysiol. 75 (6), 528–542. https://doi.org/10.1016/0013-4694(90)90139-B.

Maddox, R.K., Lee, A.K., Atilgan, H., Bizley, J.K., 2015. Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. Elife 2015 (4), 1–11. https://doi.org/10.7554/eLife.04995.001.

Marchant, J.L., Ruff, C.C., Driver, J., 2012. Audiovisual synchrony enhances BOLD responses in a brain network including multisensory STS while also enhancing target-detection performance for both modalities [Article]. Hum. Brain Mapp. 33 (5), 1212–1224. https://doi.org/10.1002/hbm.21278.

Natale, E., Marzi, C.A., Girelli, M., Pavone, E.F., Pollmann, S., 2006. ERP and fMRI correlates of endogenous and exogenous focusing of visual-spatial attention. Eur. J. Neurosci. 23 (9), 2511–2521. https://doi.org/10.1111/j.1460-9568.2006.04756.x.

Noesselt, T., Tyll, S., Boehler, C.N., Budinger, E., Heinze, H.J., Driver, J., 2010. Sound-induced enhancement of low-intensity vision: multisensory influences on human sensory-specific cortices and thalamic bodies relate to perceptual enhancement of visual detection sensitivity. J. Neurosci. 30 (41), 13609–13623. https://doi.org/10.1523/jneurosci.4524-09.2010.

Nozaradan, S., Peretz, I., Mouraux, A., 2012. Steady-state evoked potentials as an index of multisensory temporal binding [Article]. Neuroimage 60 (1), 21–28. https://doi.org/10.1016/j.neuroimage.2011.11.065.

Ouyang, G., Li, Y., 2025. Protocol for semi-automatic EEG preprocessing incorporating independent component analysis and principal component analysis. STAR Protoc. 6 (1), 103682. https://doi.org/10.1016/j.xpro.2025.103682.

Parise, C.V., Harrar, V., Ernst, M.O., Spence, C., 2013. Cross-correlation between auditory and visual signals promotes multisensory integration [Article]. Multisens. Res. 8 (3), 307–316. https://doi.org/10.1163/22134808-00002417.

Park, H., Kayser, C., Thut, G., Gross, J., 2016. Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility [Article]. Elife 5. https://doi.org/10.7554/eLife.14521.

Pelli, D.G., 1997. The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat. Vis. 10 (4), 437–442.

Peng, F., Bizley, J.K., Schnupp, J.W., Auksztulewicz, R., 2023. Dissociable neural correlates of multisensory coherence and selective attention [Journal Article]. J. Neurosci. https://doi.org/10.1523/JNEUROSCI.1310-22.2023.

Peter Rosenfeld, J., Olson, J.M., 2021. Bayesian data analysis: a fresh approach to power issues and null hypothesis interpretation. Appl. Psychophysiol. Biofeedback 46 (2), 135–140. https://doi.org/10.1007/s10484-020-09502-y.

Porcu, E., Keitel, C., Müller, M.M., 2013. Concurrent visual and tactile steady-state evoked potentials index allocation of inter-modal attention: a frequency-tagging study. Neurosci. Lett. 556, 113–117. https://doi.org/10.1016/j.neulet.2013.09.068.

Quintana, D.S., Williams, D.R., 2018. Bayesian alternatives for common null-hypothesis significance tests in psychiatry: a non-technical guide using JASP. BMC Psychiatry 18 (1), 178. https://doi.org/10.1186/s12888-018-1761-4.

Reisinger, P., Gillis, M., Suess, N., Vanthornhout, J., Haider, C., Hartmann, T., Hauswald, A., Schwarz, K., Francart, T., Weisz, N., 2025. Neural speech tracking contribution of lip movements predicts behavioral deterioration when the speaker's mouth is occluded. eNeuro. https://doi.org/10.1523/ENEURO.0368-24.2024. ENEURO.0368-0324.2024.

Sciortino, P., Kayser, C., 2023. Steady state visual evoked potentials reveal a signature of the pitch-size crossmodal association in visual cortex [Article]. Neuroimage 273, 120093. https://doi.org/10.1016/j.neuroimage.2023.120093. Article.

Seijdel, N., Schoffelen, J.M., Hagoort, P., Drijvers, L., 2024. Attention drives visual processing and audiovisual integration during multimodal communication. J. Neurosci. 44 (10). https://doi.org/10.1523/jneurosci.0870-23.2023.

Senkowski, D., Saint-Amour, D., Gruber, T., Foxe, J.J., 2008. Look who's talking: the deployment of visuo-spatial attention during multisensory speech processing under noisy environmental conditions. Neuroimage 43 (2), 379–387. https://doi.org/10.1016/j.neuroimage.2008.06.046.

Senkowski, D., Saint-Amour, D., Höfle, M., Foxe, J.J., 2011. Multisensory interactions in early evoked brain activity follow the principle of inverse effectiveness. Neuroimage 56 (4), 2200–2208. https://doi.org/10.1016/j.neuroimage.2011.03.075.

Senkowski, D., Talsma, D., Herrmann, C.S., Woldorff, M.G., 2005. Multisensory processing and oscillatory gamma responses: effects of spatial selective attention. Exp. Brain Res. 166 (3–4), 411–426. https://doi.org/10.1007/s00221-005-2381-z.

Shen, L., Li, S., Tian, Y., Wang, Y., Jiang, Y., 2025. Cortical tracking of hierarchical rhythms orchestrates the multisensory processing of biological motion. Elife 13. https://doi.org/10.7554/eLife.98701. RP98701.

Shen, L., Lu, X., Wang, Y., Jiang, Y., 2023a. Audiovisual correspondence facilitates the visual search for biological motion. Psychon. Bull. Rev. https://doi.org/10.3758/s13423-023-02308-z.

Shen, L., Lu, X., Yuan, X., Hu, R., Wang, Y., Jiang, Y., 2023b. Cortical encoding of rhythmic kinematic structures in biological motion. Neuroimage 268, 119893. https://doi.org/10.1016/j.neuroimage.2023.119893. Article.

Slagter, H.A., Prinssen, S., Reteig, L.C., Mazaheri, A., 2016. Facilitation and inhibition in attention: functional dissociation of pre-stimulus alpha activity, P1, and N1 components. Neuroimage 125, 25–35. https://doi.org/10.1016/j.neuroimage.2015.09.058.

Spence, C., Driver, J., 2000. Attracting attention to the illusory location of a sound: reflexive crossmodal orienting and ventriloquism. Neuroreport 11, 2057–2061.

Stein, B.E., Stanford, T.R., Rowland, B.A., 2020. Multisensory Integration and the Society for Neuroscience: then and now. J. Neurosci. 40 (1), 3–11. https://doi.org/10.1523/jneurosci.0737-19.2019.

Störmer, V.S., McDonald, J.J., Hillyard, S.A., 2009. Cross-modal cueing of attention alters appearance and early cortical processing of visual stimuli. Proc. Natl. Acad. Sci. 106 (52), 22456–22461. https://doi.org/10.1073/pnas.0907573106.

Talsma, D., Doty, T.J., Woldorff, M.G., 2007. Selective attention and audiovisual integration: is attending to both modalities a prerequisite for early integration? Cereb. Cortex 17 (3), 679–690. https://doi.org/10.1093/cercor/bhk016.

Talsma, D., Senkowski, D., Soto-Faraco, S., Woldorff, M.G., 2010. The multifaceted interplay between attention and multisensory integration. Trends Cogn. Sci. 14 (9), 400–410. https://search.ebscohost.com/login.aspx?direct=true&db=edsfra&AN=edsfra.23242131&lang=zh-cn&site=eds-live. htt ps://ww w.ncbi.nlm.nih.gov/ pmc/articles/PMC3306770/pdf/nihms219313.pdf.

Talsma, D., Woldorff, M.G., 2005. Selective attention and multisensory integration: multiple phases of effects on the evoked brain activity. J. Cogn. Neurosci. 17 (7), 1098–1114. https://doi.org/10.1162/0898929054475172.

ten Oever, S., Schroeder, C.E., Poeppel, D., van Atteveldt, N., Zion-Golumbic, E., 2014. Rhythmicity and cross-modal temporal cues facilitate detection. Neuropsychologia 63, 43–50. https://doi.org/10.1016/j.neuropsychologia.2014.08.008.

Van der Burg, E., Olivers, C., Bronkhorst, A., Theeuwes, J., 2008. Pip and pop: nonspatial auditory signals improve spatial visual search. J. Exp. Psychol. Hum. Percept. Perform. 34 (5), 1053–1065. https://core.ac.uk/download/pdf/15455974.pdf.

Van der Burg, E., Talsma, D., Olivers, C.N.L., Hickey, C., Theeuwes, J., 2011. Early multisensory interactions affect the competition among multiple visual objects. Neuroimage 55 (3), 1208–1218. https://doi.org/10.1016/j.neuroimage.2010.12.068.

Van der Stoep, N., Van der Stigchel, S., Nijboer, T.C.W., 2015. Exogenous spatial attention decreases audiovisual integration. Atten. Percept. Psychophys. 77 (2), 464–482. https://doi.org/10.3758/s13414-014-0785-1.

van Ee, R., van Boxtel, J.J., Parker, A.L., Alais, D., 2009. Multisensory congruency as a mechanism for attentional control over perceptual selection. J. Neurosci. 29 (37), 11641–11649. https://doi.org/10.1523/jneurosci.0873-09.2009.

Vroomen, J., Bertelson, P., de Gelder, B., 2001a. Directing spatial attention towards the illusory location of a ventriloquized sound. Acta Psychol. 108 (1), 21–33. https://doi.org/10.1016/S0001-6918(00)00068-8.

Vroomen, J., Bertelson, P., De Gelder, B., 2001b. The ventriloquist effect does not depend on the direction of automatic visual attention. Percept. Psychophys. 63 (4), 651–659. https://doi.org/10.3758/BF03194427.

Yuan, P., Hu, R., Zhang, X., Wang, Y., Jiang, Y., 2021. Cortical entrainment to hierarchical contextual rhythms recomposes dynamic attending in visual perception. Elife 10, 21. http://ir.psych.ac.cn/handle/311026/39701.

Yuan, Y., Wayland, R., Oh, Y., 2020. Visual analog of the acoustic amplitude envelope benefits speech perception in noise. J. Acoust. Soc. Am. 147 (3), El246. https://doi.org/10.1121/10.0000737.

Zion Golumbic, E., Cogan, G.B., Schroeder, C.E., Poeppel, D., 2013. Visual input enhances selective speech envelope tracking in auditory cortex at a “cocktail party”. J. Neurosci. 33 (4), 1417–1426. https://doi.org/10.1523/jneurosci.3675-12.2013.

Zou, H., Müller, H.J., Shi, Z., 2012. Non-spatial sounds regulate eye movements and enhance visual search. J. Vis. 12 (5), 2. https://doi.org/10.1167/12.5.2.