Research article

Crossmodal processing of environmental sounds and everyday life actions: An ERP study

Mirella Manfredi [a],[*], Pamella Sanchez Mello de Pinho [b], Lucas Murrins Marques [b], Beatriz de Oliveira Ribeiro [b], Paulo Sergio Boggio [b],[**]

[a] Department of Psychology, University of Zurich, Zurich, Switzerland
[b] Social and Cognitive Neuroscience Laboratory, Developmental Disorders Program, Center for Health and Biological Sciences, Mackenzie Presbyterian University, Sao Paulo, Brazil

ARTICLE INFO

ABSTRACT

To investigate the processing of environmental sounds, previous researchers have compared the semantic processing of words and sounds, yielding mixed results. This study aimed to specifically investigate the electrophysiological mechanism underlying the semantic processing of environmental sounds presented in a naturalistic visual scene. We recorded event-related brain potentials in a group of young adults over the presentation of everyday life actions that were either congruent or incongruent with environmental sounds. Our results showed that incongruent environmental sounds evoked both a P400 and an N400 effect, reflecting sensitivity to physical and semantic violations of environmental sounds' properties, respectively. In addition, our findings showed an enhanced late positivity in response to incongruous environmental sounds, probably reflecting additional reanalysis costs. In conclusion, these results indicate that the crossmodal processing of the environmental sounds might require the simultaneous involvement of different cognitive processes.

1. Introduction

In 1993, Ballas defined environmental sounds as produced by a real event and characterized by meaning by virtue of their causal relation to that event. Environmental sounds include a wide range of types, such as those generated by humans, tools, animals, liquids and objects (Lewis et al., 2004; Alluri and Kadiri, 2019). In the current study, we adopted a crossmodal approach to examine the semantic processing of environmental sounds and the human actions that produce them.

The semantic processing of environmental sounds has been investigated and analyzed by focusing on the N400, an event-related brain potential (ERP) component indexing semantic processing that usually peaks 400 ms after the stimulus onset. This component reflects the access to semantic meaning by a stimulus in relation to a previous context (Kutas and Federmeier, 2011). The amplitude of the N400 is larger in response to semantically incongruent words presented in semantic priming paradigms (Bentin et al., 1985; Kutas and Hillyard, 1980). Other types of information also elicit the N400 response, such as individual images, visual sequences and visual narratives (Cohn, 2012; Van Berkum

et al., 2003; Barrett and Rugg, 1990; Sitnikova et al., 2003, 2008; West and Holcomb, 2002). Although the N400 is sensitive to different types of information, this component's morphology and latency vary with different sensory modalities (Manfredi et al., 2018; 2020a, 2020b, 2021).

Previous ERP studies compared the semantic processing of speech and environmental sounds by combining them with semantically incongruent images, videos or visual narratives (Cummings et al., 2008; Liu et al., 2011; Manfredi et al., 2018; Plante et al., 2000; Puce et al., 2007; see also Dick et al., 2016). In 2011, Liu and colleagues analyzed the integration of natural/non-natural sound and visual information by presenting videos of real-world events with semantically congruent or incongruent natural sound/speech. The results revealed that videos with inconsistent natural sound elicited N400–P600 effects, while videos with inconsistent speech elicited N400-LPN. However, the findings of this investigation pointed out that the N400 effect in response to videos with semantically inconsistent speech was larger and later than that inconsistent with semantically inconsistent sounds. Therefore, these findings highlighted both commonalities and differences in the semantic processing of environmental sounds and words.

Similar results were found in a recent ERP work performed by our group (Manfredi et al., 2018). We analyzed the crossmodal semantic processing of auditory words and/or environmental sounds, semantically consistent or inconsistent with the critical visual event embedded in a visual narrative. We found that both incongruent sounds and words elicited an N400 effect, however this effect showed different distribution and latency between modalities: incongruent words elicited a centro-parietal scalp distributed effect, while sounds elicited a more fronto-central distributed N400 effect.

All in all, these previous studies indicated that although semantic processing of speech and environmental sounds are similar, differences in amplitude, latency and scalp distribution of the N400 effects suggest the existence of a cortical network sensitive to different modalities (Manfredi et al., 2018; Manfredi et al., 2017; Kutas and Federmeier, 2011; van Petten and Rheinfelder, 1995).

Nevertheless, other studies have not found such differences in the semantic processing of words and environmental sounds. For example, Cummings et al. (2008) compared behavioural and electrophysiological responses associated with semantic processing of words and environmental sounds in pre-adolescents, adolescents and young adults. They compared the N400 effect evoked by words and environmental sounds associated with semantically congruent and incongruent pictures. The results revealed that children showed longer latencies and larger amplitudes of the N400 effect for environmental sound when compared to adults. Nevertheless, unlike previous studies (van Petten and Rheinfelder, 1995; Manfredi et al., 2018), Cummings and colleagues did not find laterality and/or scalp differences in the processing of words and environmental sounds. The authors stated that a possible reason could be found in the different data analyses performed.

The studies mentioned above investigated the semantic processing of environmental sounds by comparing them with those elicited by auditory words. Although this comparison is worthful and offers important information about similarities and differences between these two classes of information in the semantic system, it could lead to confounding results. For example, potential confounding effects could be produced when the image of an object is presented in association with a sound and with a word in the same block of trials. In addition, since previous studies in which the processing of sounds has been compared with that related to words yielded mixed results (van Petten and Rheinfelder, 1995; Cummings et al., 2008), it would be worth investigating the semantic processing of environmental sounds by using a paradigm that isolates them from those related to words. Thus, in the present work:

- We tried to analyze the semantic processing of environmental sounds, distinguishing them from those related to words.
- We adopted a crossmodal approach to identify the specific electrocortical signature of environmental sounds presented in a naturalistic visual scene.
- We investigated the brain responses of healthy young adults during the presentation of environmental sounds and highly ecological pictures representing actions (un)related to the sounds.

In our design, environmental sounds were semantically congruent or incongruent with the action depicted in the picture. Based on previous evidence, we expected the congruent/incongruent contrast to elicit a fronto-central N400 effect, similar to the one found in our previous study (Manfredi et al., 2018).

2. Methods

2.1. Participants

Participants were thirty healthy, right-handed, native Portuguese speakers university students (15 females) recruited from the Presbyterian Mackenzie University of São Paulo (mean age = 23.6, SE = 4.7). They had a normal or corrected-to-normal vision and reported no history of neurological illness or drug abuse. We assessed their handedness by the Portuguese version of the Edinburgh Handedness Inventory (Oldfield, 1971), a laterality preference questionnaire reporting right-handedness dominance for all participants. We excluded six participants from the ERP statistical analyses because of EEG artefacts: excessive blinking, eye movements, and muscular movements. The final sample included 24 participants (13 females).

The study adhered to the Declaration of Helsinki guidelines and was approved by the institutional ethics committee of Mackenzie Presbyterian University (Brazil).

2.2. Stimuli

The materials consisted of 96 complex ecological scenes depicting everyday actions. The images were downloaded from Google Images (copy-right free).

Half pictures were assigned to the congruent condition, and the other half was assigned to the incongruent condition (Figure 1). An environmental sound was associated with each picture (mean duration = 1070 ms, SD = 207 ms), and it could be congruent or incongruent with the action (the complete dataset is available at https://osf.io/s3ucf/). The sound quality was standardized using the *normalize* function of the audio software *Ocenaudio* v. 3.3.10 (44.1 kHz, 16 bit, stereo).

The stimuli were previously presented to 10 judges with similar ages (mean age = 24.3, SE = 2.25) and educational levels as the participants who participated in the experimental task. We asked the judges to evaluate the semantic congruency between the action depicted in the pictures and the sound.

Congruent stimuli evaluated as incoherent by more than 20% of judges were discarded, as were incongruent sequences evaluated as coherent. The final stimulus set consisted of 48 stimuli per condition. We created two final lists (each consisting of 96 stimuli), with the two conditions counterbalanced such that participants viewed each stimulus only once in a list.

2.3. Procedure

Participants' task consisted in responding as accurately and quickly as possible to the presence of landscapes or interior (images without visible persons) by pressing a response key with the index finger of the left or right hand. The two hands were used alternately during the recording session, and the order of the hand and task conditions was counterbalanced across participants.

Participants faced a high-resolution monitor in an electrically shielded recording chamber. Before each picture, a fixation cross appeared for a duration of 1500 ms. The auditory sentence was time-locked to each picture that was presented at the centre of the screen. Pictures and sounds were presented for 3s, separated by an ISI of 1000 ms.

Participants were instructed not to blink or move during the experimental session. The experiment had five sections separated by breaks. A short training preceded experimental trials. The order of presentation of the stimuli was randomized before the beginning of each recording session. The total duration of participation took an average of 30 min, counting the time of preparation of the EEG and the accomplishment of the experimental task.

2.4. EEG recording parameters

We recorded the electroencephalogram (EEG) from 128 electrodes at a sampling rate of 250 Hz (bandpass 0.01–100 Hz). For both EEG recording and analyzing, we used the Net station software (*Geodesic EEG Net Station, EGI*, Eugene, OR). The impedance of all electrodes was maintained below 50 kΩ over the experiment. All recordings were referenced to Cz electrode during data acquisition. EEG epochs were synchronized with the onset of stimuli presentation.

Congruent

Mouse click sound

Inflating sound

Incongruent

Turning on a blender

Feet tapping sound

Figure 1. Examples of the stimuli used as experimental stimuli in the congruent (above) and incongruent conditions (below). We synchronized a congruent or incongruent environmental sound to the image depicting human actions.

2.5. Statistical analysis of ERP responses

Before averaging, we discarded trials contaminated by blinks, muscle tension (EMG), channel drift, and/or amplifier blocking. Approximately 9% of critical panel epochs were rejected due to such artefacts, with losses distributed approximately evenly across the four conditions. Participant's EEG was time-locked to the onset of the auditory stimuli, and ERPs from 100 ms before to 1500 ms after stimulus onset were averaged off-line (see Figure 2 for an overview of the entire process).

We analyzed two epochs of interest: the mean amplitude voltage of the N400 (or P400) in the 350–550 ms epoch and of the LP in the 550–750 ms epoch. These responses were measured at 72 electrode sites (24 in each region) in frontal (23, 24, 26, 32, 33,27, 34, 28, 18, 16, 10, 22, 9, 15, 21, 14, 3, 124, 2, 123, 122, 1, 117, 116, 39), central (40, 45, 46, 41, 47, 50, 51, 7, 106, 31, 80, 55, 30, 105, 79, 115, 109, 104, 103, 102, 98, 101, 97) and posterior (58, 59, 60, 70, 66, 65, 69, 64, 61, 62, 78, 67, 72, 77, 71, 76, 96, 85, 91, 83, 84, 90, 95, 89) regions (Figure 3).

We analyzed the mean amplitude of each component using repeated-measures ANOVAs with three factors within groups: Congruency (2 levels: Congruent, Incongruent), Region (frontal, central, posterior), Electrode (14 levels). Multiple comparisons of means were performed with post-hoc Fisher's tests.

3. Results

3.1. Electrophysiological results

3.1.1. N/P400 (350–550 ms)

The ANOVA across regions for the mean amplitude of the N400 component showed a significant interaction between Congruency x Region (F (8, 176) = 7.53; p < .01; $\eta p^2 = 0.25$) revealing a greater positivity in response to Incongruent stimuli than congruent ones in the frontal area (Congruent: -3.48 µV; SE = 0.48; Incongruent: -2.50 µV; SE = 0.51; p < .05) and to Incongruent stimuli than Congruent ones in the parietal area (Congruent: 3.90µV; SE = 0.46; p < .05; Incongruent: 3.48 µV; SE = 0.48) (Figure 4). We found no differences between the N400 response to congruent and incongruent stimuli in the central sites.

3.1.2. Later effects (550–750 ms)

The ANOVA across regions for the mean amplitude of the LP component showed a significant interaction between Congruency and Region (F (8, 184) = 2.57, p < .01; $\eta p^2 = 0.10$) revealing that incongruent stimuli elicited ERPs that were more positive-going compared to congruent ones over frontal areas (p < .01; Congruent = -3.06 µV, SE = 0.73; Incongruent = -2.16µV, SE = 0.76) (Figure 4). No differences were
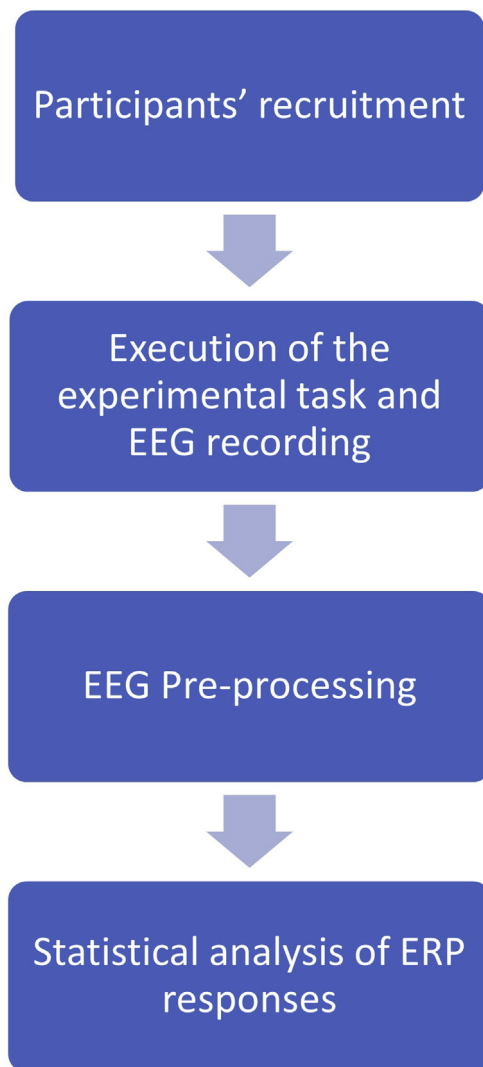
Figure 2. Block diagram describing the entire process.

Figure 3. Schematic showing the 128-channel array of scalp electrodes from which the EEG was recorded.

found between Congruent and Incongruent stimuli across the central and the parietal sites.

4. Discussion

The present work examined the semantic processing associated with the comprehension of environmental sounds associated with everyday life actions. To this aim, we analyzed the electrophysiological responses underlying the semantic processing of the environmental sounds presented in association with highly ecological pictures.

Our findings revealed that incongruent sounds elicited an N400 effect over centroparietal sites; however, we also observed an enhanced positivity for incongruous environmental sounds at frontal sites. This finding suggests that the semantic processing of the environmental sounds might elicit separate brain responses. Below, we further elaborate on this interpretation.

In the 350–550 ms time window, we observed a larger N400 response to incongruent sounds than congruent ones only in the parietal areas. The larger N400 to incongruent sounds with visual information reflects the more difficult retrieval process for semantic information that is semantically incongruent or unexpected (Kutas and Federmeier, 2011). These findings are consistent with previous results of N400s to semantic anomalies between environmental sounds and images or videos (Cummings et al., 2006; Liu et al., 2011; Plante et al., 2000; Puce et al., 2007).
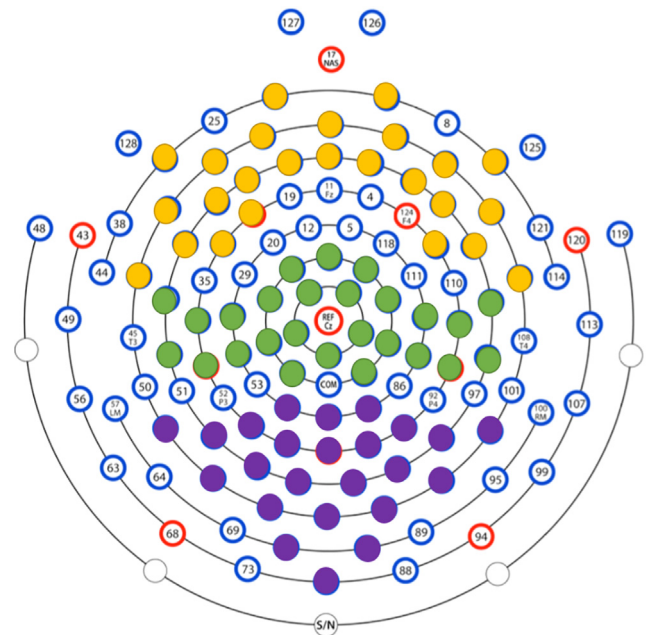
In 2006, Cummings and colleagues compared the behavioural and electrophysiological responses elicited by presenting auditory words, environmental sounds, and noises that were semantically congruent or not congruent with a visual context. The results demonstrated that both incongruent meaningful environmental sounds and words elicited an N400 effect in the centro-parietal areas, suggesting partially overlapping neural networks for auditory words and environmental sounds.

Moreover, a positive-going potential with a frontal distribution occurred in the same time window as the N400. We refer to this component as the P400. To our knowledge, this positive component occurred only in few studies. In 1984, McCallum and colleagues analyzed the brain responses of a group of young adults while they were presented with auditory sentences spoken by a male voice. The final word of the sentences could be semantically incongruous or grammatically correct but unexpectedly spoken by a female voice. The results pointed out that the semantically incongruent words elicited a large N400 component, while the physical (voice) incongruity evoked a late positive component (P416). According to the authors, the auditory P416 was elicited to the physical incongruity produced by the speaker's voice change.

A similar positive-going ERP response was found in a more recent ERP study performed by Puce et al. (2007). In this study, ERPs were recorded in a group of adults presented with dynamic human or monkey faces with associated congruent or incongruent vocalizations/sounds. The results revealed a large P400 response only when human faces were mismatched with a non-human sound. The authors concluded that the P400 effect was modulated only when the human face motion was paired with an incongruous auditory stimulus, reflecting a species-specific incongruity response.

Therefore, in our study, the increase of the P400 amplitude might reflect physical violation, elicited by the physical mismatch between the environmental sound and the representation of the associated action.

It is interesting to notice that these previous studies have only found a positive response, whereas the incongruent environmental sounds of our study evoked both a positive response and a negative response within the same time window but with a different cortical distribution. A possible explanation for such different responses could be found in the different types of violations. In the previous studies, the physical incongruity is produced by changing the gender of the speaker's voice or by replacing species-specific vocalization (Puce et al., 2007). Otherwise, either a
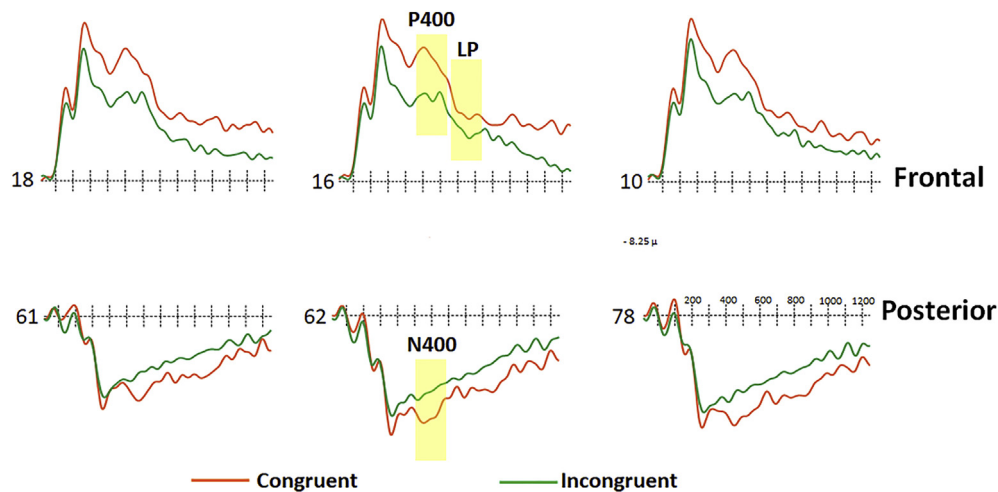
Figure 4. Grand-average ERP waveforms recorded at the frontal and posterior electrode sites in response to Congruent (orange) and Incongruent (green) stimuli.

physical or a semantic violation is present in our study. Specifically, it is possible that the P400 was evoked by the violation of the physical features of the expected environmental sound, whereas the N400 was rather evoked by the mismatch between the environmental sound and the representation of the associated action.

Moreover, our study showed a larger late positivity to incongruent sounds than the congruent ones. Incongruent stimuli elicited ERPs that were more positive-going compared to congruent ones over frontal. This late positivity has been associated with reanalysis processes of incongruent situations (Ibáñez et al., 2011; Sitnikova et al., 2003). Similarly to the N400, the LP can be evoked by meaningful but non-linguistic stimuli such as objects (Ganis and Kutas, 2003), pictures (Federmeier and Kutas, 2001; Ganis et al., 1996; McPherson and Holcomb, 1999), gestures (Kelly et al., 2004; Proverbio and Riva, 2009; Wu and Coulson, 2007).

In their study, Liu et al. (2011) found a large P600 in response to videos with semantically inconsistent natural sound and an increased LPN in response to videos with inconsistent speech. The authors suggested that videos with natural sound and videos with speech might involve different late cognitive processing: while for natural sounds, the brain could directly extract the semantic meanings and perform the integration processing with the visual information, for the speech, the brain should first convert the speech into natural sound, a process that could be reflected in the LPN response.

In our study, the late positivity response observed in response to incongruent sound-pictures stimuli might reflect reanalysis costs. In line with the previous study, our results suggest that incongruent sounds require additional processing costs to process the semantically incongruent combination of everyday human actions and environmental sounds.

One limitation of this study is that our task did not differentiate the semantic processing of auditory words and environmental sounds associated with everyday actions. In the previous study performed by our group (Manfredi et al., 2018), we found different distribution and latency of the N400 effect in response to auditory sounds and words associated with events in visual narratives. In a future study, it would be interesting to compare semantic processing evoked by the two classes of stimuli (i.e. environmental sounds and auditory words) presented in association with highly ecological images depicting human actions.

In conclusion, our results showed that incongruent environmental sounds evoked both a P400 and an N400 effect, suggesting a double dissociation between the physical and the semantic properties of environmental sounds. These results indicate that the crossmodal processing of the environmental sounds might require the simultaneous involvement of different brain networks involved in processing different features of these stimuli.

References

Alluri, V., Kadiri, S.R., 2019. Neural correlates of timbre processing. In: Timbre: Acoustics, Perception, and Cognition. Springer, Cham, pp. 151–172.

Ballas, J.A., 1993. Common factors in the identification of an assortment of brief everyday sounds. J. Exp. Psychol. Hum. Percept. Perform. 19 (2), 250–267.

Barrett, S.E., Rugg, M.D., 1990. Event-related potentials and the semantic matching of pictures. Brain Cognit. 14 (2), 201–212.

Bentin, S., McCarthy, G., Wood, C.C., 1985. Event-related potentials, lexical decision and semantic priming. Electroencephalogr. Clin. Neurophysiol. 60 (4), 343–355.

Cohn, N., 2012. Explaining "I can't draw": parallels between the structure and development of language and drawing. Hum. Dev. 55 (4), 167–192.

Cummings, A., Ceponiene, R., Koyama, a, Saygin, A.P., Townsend, J., Dick, F., 2006. Auditory semantic networks for words and natural sounds. Brain Res. 1115 (1), 92–107.

Cummings, A., Ceponiene, R., Dick, F., Saygin, A.P., Townsend, J., 2008. A developmental ERP study of verbal and non-verbal semantic processing. Brain Res. 1208, 137–149.

Dick, F., Krishnan, S., Leech, R., Saygin, A.P., 2016. Environmental sounds. In: Neurobiology of Language. Academic Press, pp. 1121–1138.

Federmeier, K.D., Kutas, M., 2001. Meaning and modality: influences of context, semantic memory organization, and perceptual predictability on picture processing. J. Exp. Psychol. Learn. Mem. Cognit. 27 (1), 202–224.

Ganis, G., Kutas, M., Sereno, M.I., 1996. The search for "common sense": an electrophysiological study of the comprehension of words and pictures in reading. J. Cognit. Neurosci. 8 (2), 89–106.

Ganis, G., Kutas, M., 2003. An electrophysiological study of scene effects on object identification. Cognit. Brain Res. 16 (2), 123–144.

Ibáñez, A., Toro, P., Cornejo, C., Urquina, H., Hurquina, H., Manes, F., Weisbrod, M., Schröder, J., 2011. High contextual sensitivity of metaphorical expressions and gesture blending: a video event-related potential design. Psychiatr. Res. 191 (1), 68–75.

Kelly, S.D., Kravitz, C., Hopkins, M., 2004. Neural correlates of bimodal speech and gesture comprehension. Brain Lang. 89 (1), 253–260.

Kutas, M., Hillyard, S.A., 1980. Reading senseless sentences: brain potentials reflect semantic incongruity. Science 207 (4427), 203. LP – 205.

Kutas, M., Federmeier, K.D., 2011. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annu. Rev. Psychol. 62, 621–647.

Lewis, J.W., Wightman, F.L., Brefczynski, J. a, Phinney, R.E., Binder, J.R., DeYoe, E.a., 2004. Human brain regions involved in recognizing environmental sounds. Cerebr. Cortex 14 (9), 1008–1021 (1991).

Liu, B., Wang, Z., Wu, G., Meng, X., 2011. Cognitive integration of asynchronous natural or non-natural auditory and visual information in videos of real-world events: an event-related potential study. Neuroscience 180, 181–190.

Manfredi, M., Cohn, N., Kutas, M., 2017. When a hit sounds like a kiss: an electrophysiological exploration of semantic processing in visual narrative. Brain Lang. 169.

Manfredi, M., Cohn, N., De Araújo Andreoli, M., Boggio, P.S., 2018. Listening beyond seeing: event-related potentials to audiovisual processing in visual narrative. Brain Lang. 185, 1–8.

Manfredi, M., Cohn, N., Ribeiro, B., Sanchez, P., Fernandes, E., Pereira, R., Sergio, P., 2021. Brain and Cognition the Electrophysiology of Audiovisual Processing in Visual Narratives in Adolescents with Autism Spectrum Disorder. 151(February).

Manfredi, M., Cohn, N., Sanchez Mello, P., Fernandez, E., Boggio, P.S., 2020a. Visual and verbal narrative comprehension in children and adolescents with autism spectrum disorders: an ERP study. J. Autism Dev. Disord.

Manfredi, M., Proverbio, A.M., Sanchez Mello de Pinho, P., Ribeiro, B., Comfort, W.E., Murrins Marques, L., Boggio, P.S., 2020b. Electrophysiological indexes of ToM and non-ToM humor in healthy adults. Exp. Brain Res. 238 (4), 789–805.

McCallum, W.C., Farmer, S.F., Pocock, P.V., 1984. The effects of physical and semantic incongruites on auditory event-related potentials. Electroencephalogr. Clin. Neurophysiology Evoked Potentials Sect. 59 (6), 477–488.

McPherson, W.B., Holcomb, P.J., 1999. An electrophysiological investigation of semantic priming with pictures of real objects. Psychophysiology 36 (1), 53–65.

Plante, E., Petten, C.V., Senkfor, a J., 2000. Electrophysiological dissociation between verbal and nonverbal semantic processing in learning disabled adults. Neuropsychologia 38 (13), 1669–1684. http://www.ncbi.nlm.nih.gov/pubmed/11099725.

Proverbio, A.M., Riva, F., 2009. RP and N400 ERP components reflect semantic violations in visual processing of human actions. Neurosci. Lett. 459 (3), 142–146.

Puce, A., Epling, J. a, Thompson, J.C., Carrick, O.K., 2007. Neural responses elicited to face motion and vocalization pairings. Neuropsychologia 45 (1), 93–106.

Sitnikova, T., Holcomb, P.J., Kiyonaga, K.A., Kuperberg, G.R., 2008. Two Neurocognitive Mechanisms of Semantic Integration during the Comprehension of Visual Real-World Events. 2037–2057.

Oldfield, R.C., 1971. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9, 97–113.

Sitnikova, T., Kuperberg, G., Holcomb, P.J., 2003. Semantic integration in videos of real-world events: an electrophysiological investigation. Psychophysiology 40 (1), 160–164. http://www.ncbi.nlm.nih.gov/pubmed/12751813.

Van Berkum, J.J.A., Zwitserlood, P., Hagoort, P., Brown, C.M., 2003. When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. Cognit. Brain Res. 17 (3), 701–718.

van Petten, C.K., Rheinfelder, H., 1995. Conceptual relationships between words and environmental sounds: event-related brain potentials measures. Neuropsychologia 33 (4), 485–508. http://www.sciencedirect.com/science/article/pii/002839329400133A.

West, W.C., Holcomb, P.J., 2002. Event-related potentials during discourse-level semantic integration of complex pictures, 13, pp. 363–375.

Wu, Y.C., Coulson, S., 2007. How iconic gestures enhance communication: an ERP study. Brain Lang. 101 (3), 234–245.