

OPEN UNIVERSITY OF CYPRUS

School of Pure and Applied Sciences

Joint Master's Programme in collaboration with the Department of Psychology and the Department of Computer Science of the University of Cyprus: *MSc. Cognitive Systems*

Master's Dissertation



Semantic content effects on the perception of movieclips

Anastasia Maria Kesoglou

Supervisor

Dr. Kyriaki Mikellidou

May 2023

OPEN UNIVERSITY OF CYPRUS

School of Pure and Applied Sciences

Joint Master's Programme in collaboration with the Department of Psychology and the Department of Computer Science of the University of Cyprus: *MSc. Cognitive Systems*

Master's Dissertation

Semantic content effects on the perception of movieclips

Anastasia Maria Kesoglou

Supervisor

Dr. Kyriaki Mikellidou

The present Master's Dissertation was submitted in partial fulfilment of the requirements for the postgraduate degree in MSc. Cognitive Systems School of Pure and Applied Sciences of the Open University of Cyprus.

May 2023

Summary

Our brain is skilled with the ability to perceive and process multimodal stimuli. This process known as crossmodal perceptual integration, has been in the research spotlight for a long time, providing evidence for the integration of information coming from different modalities. Prior experiments on the field mostly utilized pictures and were limited in the semantic content of a single sound or word. The present study aims to investigate crossmodal perceptual integration in realistic conditions using short movieclips (1500ms) and auditory meaningful three-word sentences in cases of target detection judgments. This study (N=36) is the first to introduce trials without a target that always include target-related information, which was present, either only through vision or audition (incongruent movieclips) or through both (congruent movieclips). For each target condition (present or absent) the movieclips were made up of a combination of 12 videos and 12 sentences, which were repeated in a pseudorandomized order four times for each participant (total trials= 288). The results from the two-way repeated measures ANOVA indicate a similar pattern between the two modalities for semantically incongruent movieclips, with statistically lower accuracy scores in trials where the target was present only in one modality ($M_{\text{audio}}=0.647$, $SD_{\text{audio}}=0.305$; $M_{\text{visual}}=0.841$, $SD_{\text{visual}}=0.235$), whereas in target absent trials both showed superior performance ($M_{\text{audio}}=0.931$, $SD_{\text{audio}}=0.038$; $M_{\text{visual}}=0.986$, $SD_{\text{visual}}=0.018$). On the other hand, we observed the opposite pattern for semantically congruent movieclips (Target present trials: $M_{\text{audiovisual}}=0.981$, $SD_{\text{audiovisual}}=0.036$ vs. Target absent trials: $M_{\text{audiovisual}}=0.898$, $SD_{\text{audiovisual}}=0.111$). Reaction times were the same for the two modalities ($F(2,70)=0.384$, $p=0.683$). In accordance with previous research using images and single words, our results show that when auditory and visual information is congruent, performance is superior and when the target is only present through audio but visual information is incongruent, performance is evidently compromised, and vice versa. Regarding the role of semantics, when the audio sentence included a target-related noun accompanied by a semantically incongruent video, accuracy in judgements was statistically better compared to when it was a verb ($t_{\text{incVerb vs. incNoun}}=-8.428$, $p<.001$; $t_{\text{conVeb vs. incNoun}}=-4.256$, $p<.001$). The present results could provide more evidence regarding the role of complexity of semantics, and especially the different role verbs and nouns could play in crossmodal perceptual integration in more realistic situations. Our findings can enrich the content of learning techniques, as well as the design of AI models, by taking advantage of the supporting role of semantic audiovisual information, while taking into consideration the confusion that the complexity in semantic information could cause to perception experience.

Key words: Psychology, Perception, Crossmodal integration, Audiovisual integration, Semantic congruency, Semantic audiovisual movieclips

Περίληψη

Ο εγκέφαλός μας έχει την ικανότητα να αντιλαμβάνεται και να επεξεργάζεται πολυαισθητηριακά ερεθίσματα. Αυτή η διαδικασία γνωστή ως διατροφική αντιληπτική ολοκλήρωση, ήταν στο επίκεντρο της έρευνας για μεγάλο χρονικό διάστημα, παρέχοντας στοιχεία για την ενσωμάτωση πληροφοριών που προέρχονται από διαφορετικά αισθητηριακά μέσα. Τα προηγούμενα πειράματα χρησιμοποιούσαν κυρίως εικόνες και περιορίζονταν στο σημασιολογικό περιεχόμενο ενός μόνο ήχου ή λέξης. Η παρούσα μελέτη στοχεύει στη διερεύνηση της διατροφικής αντιληπτικής ολοκλήρωσης σε ρεαλιστικές συνθήκες χρησιμοποιώντας σύντομα βίντεο κλιπ (1500ms) και ακουστικές νοηματικές προτάσεις τριών λέξεων σε περιπτώσεις κρίσεως ανίχνευσης στόχου. Η μελέτη μας (N=36) είναι η πρώτη που εισήγαγε δοκιμασίες χωρίς στόχο που περιλαμβάνουν όμως πάντα πληροφορίες σχετικές με το στόχο, οι οποίες ήταν παρούσες, είτε μόνο μέσω της όρασης ή ακοής (σημασιολογικά αντικρουόμενα βίντεο) είτε μέσω και των δύο (σημασιολογικά σύμφωνα βίντεο). Για κάθε συνθήκη στόχου (παρών ή απών) τα κλιπ ταινιών αποτελούνταν από έναν συνδυασμό 12 βίντεο και 12 προτάσεων, οι οποίες επαναλήφθηκαν με ψευδοτυχαία σειρά τέσσερις φορές για κάθε συμμετέχοντα (σύνολο δοκιμασιών = 288). Τα αποτελέσματα που προέκυψαν από την Ανάλυση Διακύμανσης Επαναλαμβανόμενων Μετρήσεων με δύο μεταβλητές (ANOVA), υποδεικνύουν ένα παρόμοιο μοτίβο μεταξύ των δύο αισθητηριακών οδών για τα σημασιολογικά αντικρουόμενα βίντεο κλιπ, με στατιστικά χαμηλότερες βαθμολογίες στην ακρίβεια σε δοκιμασίες όπου ο στόχος ήταν παρών μόνο σε μία αισθητηριακή οδό ($M_{\text{audio}} = 0.647$, $SD_{\text{audio}} = 0.305$; $M_{\text{visual}} = 0.841$, $SD_{\text{visual}} = 0.235$), ενώ σε δοκιμασίες εν απουσία στόχου και οι δύο έδειξαν ανώτερη απόδοση ($M_{\text{audio}} = 0.931$, $SD_{\text{audio}} = 0.038$; $M_{\text{visual}} = 0.986$, $SD_{\text{visual}} = 0.018$). Από την άλλη πλευρά, παρατηρήσαμε το αντίθετο μοτίβο για σημασιολογικά σύμφωνα βίντεο κλιπ (Δοκιμές εν παρουσία στόχου: $M_{\text{audiovisual}} = 0.981$, $SD_{\text{audiovisual}} = 0.036$ vs. Δοκιμές εν απουσία στόχου: $M_{\text{audiovisual}} = 0.898$, $SD_{\text{audiovisual}} = 0.111$). Οι χρόνοι αντίδρασης ήταν οι ίδιοι για τις δύο οδούς ($F(2,70) = 0.384$, $p = 0.683$). Σε συμφωνία με την έως τώρα έρευνα βασισμένη στην χρήση εικόνων και μεμονωμένων λέξεων, τα αποτελέσματά μας δείχνουν ότι όταν οι ακουστικές και οπτικές πληροφορίες είναι σύμφωνες, η απόδοση είναι καλύτερη και όταν ο στόχος είναι παρών μόνο μέσω ήχου αλλά η οπτική πληροφορία είναι ασύμβατη, η απόδοση αποδεδειγμένα υποβαθμίζεται και το αντίστροφο. Όσον αφορά στο σημασιολογικό περιεχόμενο, παρατηρήσαμε ότι όταν η ηχητική πρόταση περιλάμβανε ένα ουσιαστικό που σχετίζεται με τον στόχο συνοδευόμενη από το σημασιολογικά αντικρουόμενο βίντεο του, η ακρίβεια στις κρίσεις ήταν στατιστικά καλύτερη σε σύγκριση με όταν περιλάμβανε ρήμα ($t_{\text{incVerb vs. incNoun}} = -8.428$, $p < .001$; $t_{\text{conVerb vs. incNoun}} = -4.256$, $p < .001$). Τα παρόντα αποτελέσματα θα μπορούσαν να παρέχουν περισσότερες ενδείξεις σχετικά με το ρόλο της πολυπλοκότητας της σημασιολογίας, και ειδικά τον διαφορετικό ρόλο που θα μπορούσαν να παίξουν τα ρημάτα και τα ουσιαστικά στη διατροφική αντιληπτική ολοκλήρωση υπό πιο ρεαλιστικές καταστάσεις. Τα ευρήματά μας μπορούν να εμπλουτίσουν το περιεχόμενο των τεχνικών μάθησης, καθώς και το σχεδιασμό μοντέλων τεχνητής νοημοσύνης, εκμεταλλευόμενοι τον υποστηρικτικό ρόλο των σημασιολογικών σύμφωνων οπτικοακουστικών πληροφοριών, λαμβάνοντας

παράλληλα υπόψη τη σύγχυση που θα μπορούσε να προκαλέσει η πολυπλοκότητα στη σημασιολογία στην εμπειρία αντίληψης.

Λέξεις κλειδιά: Ψυχολογία, Αντίληψη, Διατροφικής ολοκλήρωσης, Οπτικοακουστική ολοκλήρωση, Σημασιολογική συνάφεια, Σημασιολογικά οπτικοακουστικά βίντεο

Acknowledgments

As this long and tough journey comes to an end, I feel grateful and proud for all that I accomplished through this master thesis, and this master programme in general. The knowledge and skills that I acquired will guide me to reach my professional goals and the career change I was seeking for. I am grateful to the people I was lucky to meet, and I will forever remember with respect.

I am deeply grateful to my advisor, Dr. Kyriaki Mikellidou, for her unwavering support and guidance throughout my master's dissertation. The valuable time she spent on a weekly, and many times, on a daily basis in order to advise me in every little and big step I made during the writing, experimental design, data collection and statistical analysis, motivated me to overcome every obstacle. Her expertise and patience have been invaluable to me and have played a crucial role in the success of this thesis.

I also wish to thank Dr. Konstantinos Tsagkaridis and Dr. Vasilis Pelekanos for serving on my thesis committee and providing valuable feedback and suggestions.

To all the people that volunteered to participate in my study I would like to extend my sincere gratitude. Their willingness to run the experimental task has been invaluable to my research and has helped to make this thesis a success. Thank you for your time and contribution.

Finally, I would like to thank my friends and family for their love and support during this process, and most of all my beloved partner Paris, for being there for me the whole time patiently.

Sincerely grateful to everyone.

Without your support, this thesis would not have been possible.

Table of Contents

Chapter 1 Introduction	9
1.1.Perception and Senses.....	9
1.2.The Perception of Crossmodal Stimuli.....	10
1.2.1.Crossmodal Integration Effect on Perception of Audiovisual Stimuli.....	11
1.3.Neural Signals Related to Crossmodal Perceptual Integration Effect.....	12
1.4.The Role of Semantics.....	13
Purpose of the Study	17
Chapter 2 Materials and Methods	18
2.1.Participants.....	18
2.2.Experimental Stimuli.....	18
2.3.Procedure.....	19
2.4.Design/Experimental Conditions.....	20
2.5.Data Analysis.....	23
Chapter 3 Results	24
3.1.Mean Proportion Correct.....	24
3.1.1.Semantics.....	25
3.2.Mean Reaction Times.....	27
3.2.1.Semantics.....	28
Chapter 4 Discussion	30
4.1.Target Present Trials.....	30
4.2.Target Absent Trials.....	31
4.3.The Role of Semantics (Verb vs. Noun) in Target Absent Trials	32
4.4.General Comments and Limitations.....	33
Chapter 5 Conclusion	34
Appendix	35
Appendix Tables.....	35
Appendix Figures.....	39
References	41

Chapter 1

Introduction

1.1. Perception And Senses

We perceive the world with our five senses: vision, audition, taste, olfaction, and touch -which includes tactile and temperature sense, as well as pressure. In a neuroanatomical manner, a “sense” refers to a system consisting of sensory receptor cells which respond to a specific physical stimulus and relay neural “messages” to particular brain regions in the form of electrical signals (Privitera, 2023). Due to the world’s causal and complex multisensory nature, human brain procedures as well as those of other animals have been developed over the course of time to efficiently perceive world events often through reliability weighting of the available crossmodal sensory information (Cao et al., 2019). The concept of sensory modality in perceiving the world is often being associated with what is being visually perceived, giving in this way more weight to the visual modality (Hutmacher, 2019). However, from our personal experience, we are aware of the complementary role of the diverse sensory characteristics of the objects surrounding us. For example, when we perceive stimuli in our environment such as a fruit tree in a garden, we do not perceive it only by seeing its form and colors, but we can also feel the texture of its trunk, leaves and fruits by touching them and even taste its fruits, smell its flowers or even hear its leaves shaking because of the breeze. Thus, the picture of the fruit tree is a holistic representation built from all information that we receive from distinct sensory pathways. Studies indicate the importance of more than just a single modality for the efficacy of perception. Precisely, the binding of crossmodal information is linked to the observable complementary effect of signals presented in one modality on the signals presented in another i.e., the enhanced effect of sound on a visual stimulus (Cox & Hong, 2015). The process referring to the temporal and spatial binding of two or more perceptual features rooting from the same or different sensory modalities refers to what is known as crossmodal integration (Lalanne & Lorenceau, 2004; Lachs, 2023). Discussion in Lalanne and Lorenceau (2004) was not limited to this binding of multisensory stimuli, but expanded to its neural results, which is the coherent spatiotemporal and object representations that are built after the information binding of multimodal stimuli of various reliability levels. The latter detail in the definition of crossmodal integration reveals the significant role of each sensory piece of information in experiencing materials or events as a whole percept (Schifferstein & Wastiels, 2014; Lachs, 2023).

1.2. The Perception of Crossmodal Stimuli

In general, there is a lot of discussion in the field of crossmodal information processing and the effect of one modality cue on the perception of the other, since it is an effect that we experience very commonly. In this way, we are often able to taste a specific flavor not only because of our taste buds and tongue, but even before we have a right bite, as the perception of taste can be enhanced by the smell of the food (Narumi et al., 2011). Evidence is also provided for smell perception induced by other sensory modalities such as visual cues (Koubaa & Eleuch, 2021). Thus, our everyday experience of chemical sensations -taste and smell- can provide evidence of the unavoidable multimodal interaction of simultaneously processed unimodal and crossmodal cues. (Kakutani et al., 2017; Narumi et al., 2011).

Examples of how multimodal interactions work and of their outcomes come from experimental settings utilizing all possible crossmodal and multimodal stimulus pairings. For instance, many researchers focused on the level of enhancement of one modality over the other during a crossmodal perceptual task, where crossmodal signals were either congruent (temporally/spatially) or incongruent. (Stein et al., 1996; Shams, Kamitani, & Shimojo, 2000; Calvert, Campbell, & Brammer, 2000; Vroomen & de Gelder, 2000; Morein-Zamir, Soto-Faraco & Kingstone, 2003; Laurienti et al., 2004; Ro et al, 2004; Zampini et al., 2005; Chen & Spence, 2010; van de Groen, 2013; Cox & Hong, 2015; Sakai et al., 2015; Li et al., 2019; Brandman et al., 2020; Rekow et al., 2022; Woods et al., 2023). A suppression effect of sensory cues contributes to a class of perceptual illusions (Tsuchiya, 2008). According to Bruns (2019) these perceptual illusions arise when, for example, the highest weight is being given on visual cues, influencing in this way, the perception of an audio cue (i.e., its location). Perceptual illusions have been mostly reported in settings that include the visual modality (Violentyev, Shimojo, & Shams; 2005; Kammers et al., 2009; Moscatelli et al. 2015; Kang, Sah, & Lee, 2021). Bresciani et al. (2008) used sequences of events to investigate the interaction between simultaneously presented visual flashes, haptic taps and auditory beeps. Participants were asked to count the number of events presented in the target modality and ignore all stimuli in other modalities presented as background sequences. Comparisons between the nine combinations of vision, touch and audition were conducted while taking into account any background stimuli biases. These showed immediate integration of multimodal stimuli, which was found to depend on the respective contributions of the three modalities, and in turn, on their relative reliability. Interestingly they provided evidence of the increased influence of task-irrelevant stimuli when presented in two modalities.

According to what has been discussed above, we observe that different research directions among studies are linked with different descriptive references of crossmodal perception, like “multimodal perceptual enhancement”, “suppression effect in perception” or “crossmodal illusory perception”. In particular, multimodal perceptual enhancement gives more weight to the supplementary role of congruent crossmodal stimuli that leads to the enhancement of performance (see Ball, Nentwich, & Noesselt,

2022). As we can interpret from relevant studies, such as this of Hidaka & Ide (2015), the term “suppression effect” in perception limits the definition of multimodal interaction to the dominance of one modality among the others. On the contrary, the term crossmodal illusory perception focuses more on the falsified outcome observed during perception of crossmodal stimuli, leaving aside the enhanced nature of multimodal stimuli in perception (see Bolognini et al., 2013). In the present study, the term crossmodal perceptual integration is preferred, because of its focus on the outcome of the integrative process across sensory modalities, meaning the conjunction and/or summation of multiple modality information cues (Lalanne, & Lorenceau, 2004; Xie et al., 2017).

1.2.1. Crossmodal Integration Effects on The Perception of Audiovisual Stimuli

Multiple studies have examined the diverse conditions under which crossmodal integration of visual and auditory signals occurs. Early research in the field showed that auditory information can qualitatively and quantitatively alter the perception of a visual stimulus. The visual illusion of multiple flashes because of multiple auditory beeps is an example of qualitative change of visual input caused by an audio-cue (Shams, Kamitani, & Shimojo, 2000). Respectively, the influence of sound on perception of light intensity applies as an example of quantitative change (Stein, London, Wilkinson, Price, 1996). However, this is only one part of the various effects that have been observed during the audiovisual perceptual integration process. A very common observation regarding the case of audio–visual interaction, is that audition suppresses vision in temporal perception, while vision prevails over audition for spatial perception (Wada, Kitagawa, & Noguchi, 2003; Ortega, 2014). In addition to this, Morein-Zamir, Soto-Faraco & Kingstone (2003), suggested a ‘temporal ventriloquism’ phenomenon analogous to spatial ventriloquism, showing that visual temporal order judgments were interfered when sounds were appearing between the two lights, resulting to lower accuracy scores. Similar observations have been made for speed judgments between asynchronous audiovisual stimuli, supporting the notion that auditory and visual signals tend to become perceptually integrated in a temporal manner (Arnold, Johnston & Nishida, 2005; Keetels & Vroomen, 2011). Soto-Faraco, Spence, and Kingstone (2004) demonstrated a dynamic capture effect on perception of auditory and visual motion, which expands above the static nature of the ventriloquism effect. Their experiments were conducted to extend their previous results in Soto-Faraco et al. (2002) and indicated that visual and auditory motion streams appearing in opposite directions lead to an illusory reversal in motion perception of the audio cue influenced by the direction of the visual cue.

Hidaka and Ide (2015) showed decreased performance during visual orientation discrimination tasks due to the appearance of spatially and temporally congruent sounds (white noise bursts) through headphones. They explained their findings on the basis of potentially direct and close interactions of neural responses occurring across modalities during the observation of audiovisual stimuli. In an earlier study, Teder-Sälejärvi et al.

(2002) reported better performance (both accuracy and response times RTs) during a detection task when both brief noise bursts and visual flashes were presented compared to unimodal presentation of crossmodal signals. All stimuli were centrally located and appeared at irregular intervals (600-800ms). Participants were asked to respond by pressing a button when an irrelevant stimulus appeared such as a more intense noise burst, a brighter flash, or both. Recently, Uno & Yokosawa (2022) took into consideration the spatiotemporal congruency effects between modalities, that have been discussed in detail in our Introduction, and examined their potential relation with temporal recalibration of audiovisual stimuli during perceptual integration processing. For this purpose, they conducted simultaneity judgment tasks consisting of audiovisual pairs (audio pitch either high or low, and visual cycle either presented above fixation point or below) that were synchronous or asynchronous. In each block, participants were initially completing adaptation trials for 60s where alternating auditory and visual stimuli were presented with no time interval differences. Their results showed selective recalibration of asynchronous but semantically congruent audiovisual signals.

1.3. Neural Signals Related to Crossmodal Perceptual Integration Effect

In this phase, it is worth bringing into the conversation the processes that facilitate information binding in the brain. Since integration or segregation of signals from the real-world depends on whether or not they arise from a common source, we can assume that perception of a real-world environment depends critically on its causal structure (Noppeney et al., 2018; Mihalik & Noppeney, 2020). In this way, our continuous exposure to simultaneously presented multisensory stimuli leads to the formation of associations by binding asymmetries between various signals, such as objects or sentences (Kubovy & Schutz, 2010) and interfering with our subsequent actions (Jensen et al., 2020). Formations refer to perceptual estimates, dynamically encoded across sensory processing hierarchies, that follow the principles of Bayesian Causal Inference (Rohe, Ehlis, & Noppeney, 2019). According to this framework, stronger prior belief of a shared common cause is correlated with a greater degree of perceptual binding, and in turn, with greater audiovisual integration (Tong et al., 2020).

Sensory systems possess the receptors that transfer information to the body regarding the external and internal environment (e.g., nerve damage or dysfunction of organs) and are characterized by response dynamics which can determine the perceived position, velocity or acceleration of stimulus (Feher, 2012). As previously stated, perception of multimodal stimuli requires the combination of different sensory inputs, an ability observed to exist in the nervous system. Expanded research related to the neural activity recorded during multisensory perception processes, indicates that since neurons are interrelated to each other, the level of activity in the central nervous system is not limited to modality-specific pathways, but rather by the interaction of simultaneous stimuli in multiple modalities (Stein, London, Wilkinson, & Price, 1996; Macaluso et al., 2004; Choi, Lee, & Lee, 2018) which in turn impacts the perceptual interpretation of simultaneous signals from multisensory modalities (Bushara, et al., 2003).

Evidence from examining sound influence on visual motion perception indicate higher activity in multimodal areas compared to predominantly unimodal areas, suggesting a two-way interaction where multimodal and unimodal areas compete against each other to perceptually interpretate simultaneous signals coming from multiple sensory modalities (Bushara et al., 2003). Moreover, early positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) studies reported a dominant activation of the claustrum when processing synchronized audiovisual information, indicating that the linkage between senses is achieved through a subcortical relay area. (Calvert, 2001; Olson, Gatenby, & Gore, 2002). In addition, the wider semantic network regions and areas related to extralinguistic sensory, as well as perceptual and cognitive processing have been shown to be involved in multisensory integration during speech (Ross et al., 2022). By examining the neural code that underlies responses of motor networks to different sensory conditions during feeding action, fluorescence imaging showed that modality differences are encoded in the combination of activated neurons (excitation or inhibition) and can be altered when a simultaneous activation of both pathways occurs (Follmann, Goldsmith, & Stein, 2018). Temporal and spatial congruency were highlighted as critical factors for the effectiveness of crossmodal integration, as indicated by early PET studies such as in Macaluso et al. (2004). In their study audiovisual synchrony showed to affect ventral areas in speech identification, whereas the spatial multisensory interactions were mostly associated with dorsal areas stimulation. According to Molholm et al. (2004), evidence of behavioral enhancement in the case of matched audiovisual inputs is accompanied with evoked potential changes in the latency range and general topographic region of N1—a visual evoked component related to early feature processing in the ventral visual stream. More interestingly, Molholm et al. (2004) suggested that auditory stimuli modulate regional lateral-occipital-cortex processing, a brain area related to visual object perception and the discrimination between face/non-face stimuli (Nagy, Greenlee, & Kovács, 2012). A recent N1/P2 event-related potential paradigm indicates the great dynamic of theta-band activity induced by audiovisual integration of sine-wave speech (SWS) (Lindborg et al., 2019). SWS refers to a highly degraded speech signal, constructed from frequency- and amplitude-modulated sinusoids (Rosen & Hui, 2015). Researchers have also tried to enlighten the significance of neural oscillations in the audiovisual perceptual integration of affective signals, suggesting its relationship with oscillation activities of early evoked sub-additive theta, as well as sustained induced supra-additive delta and beta, irrespective of affective content (Gao et al., 2021).

1.4. The Role of Semantics

The presence of multiple crossmodal cues can enhance behavioral performance by speeding responses, increasing accuracy, and improving stimulus detection. (Laurienti et al., 2004). Studies focusing on brain activation patterns during perception of semantically congruent and incongruent audiovisual stimuli suggest that there is a strong relation between congruent audiovisual stimuli and facilitation of neural representations of

semantic categories or concepts (Li et al., 2011). Thus, it is of great interest to examine the role of semantic content in the perception of audiovisual stimuli.

The importance of semantic content in crossmodal effect of sound on perception of visual stimuli has been acknowledged by researchers over the last years. As Laurienti et al. suggested in 2004, through this next level investigation, the research community could gain more insight into neural operations and constituents of crossmodal information processing. There have been a few but very interesting studies which focused on the impact of semantic sounds on the perception of pictures and the semantic congruency between crossmodal stimuli. Williams et al. (2022) used a psychophysical task consisting of pairs of naturalistic sounds and noisy visual images. In their first experiment an ambiguous visual image which referred to two distinct objects (identical shadows of two different objects) switched from obscured to clear view while a naturalistic sound played. The incidental sound was either coherent with one of the two objects or completely irrelevant. Participants were asked to recreate the target morph they previously saw as accurately as possible using a continuous report line and press the button when they were ready to submit their answer. No sound was playing during the response phase. In the second experiment, sounds were played during the continuous report phase whereas in the second part of the experiment half of all blocks had no sound. A further experiment was conducted to examine potential high-level semantic representations implicated by sounds, presenting the full length of the naturalistic sound followed by the visual morph after a 3-s delay. The authors observed a continuous integration of temporally and semantically congruent audiovisual inputs and explained the perception of visual objects as a function of naturalistic auditory context, since the latter provides and enriches visual perception with complementary information, which is both independent and diagnostic (Williams et al., 2022). Results from a crossmodal and a visual feature discrimination task indicate that complementary to the role of spatial and temporal relationships between multisensory stimuli and this of their physical effectiveness, the semantic content plays also a significant role in multisensory information binding, in a goal-directed manner (Laurienti et al. 2004). Moreover, Chen and Spence (2010) conducted a series of experiments to assess the effect of audiovisual semantic congruency specified on the identification of masked visual targets. The results indicate a shared semantic system in which neural representations of crossmodal stimuli interact, and that this crossmodal semantic interaction depends on a short-term buffer responsible for handling semantic representations.

In the case of unimodal stimuli, there is evidence regarding the influence of language on a specific visual attribute when the content of the presented written sentence and visual attribute are semantically congruent (Pelekanos & Moutoussis, 2011). As for semantic congruency effects on crossmodal stimuli, past research focused on two factors regarding crossmodal semantic congruency effects of auditory stimuli on the processing of visual cues: synchronization and categorical specificity. In such Chen and Spence (2018), used either naturalistic sounds or spoken words in combination with pictures or printed words. Seven stimulus onset asynchronous (SOA) conditions were utilized in the

experiments, and the task given to the participants required speeded categorization judgements to whether each cue presented belonged to the living or nonliving objects category. Both congruency and inhibitory effects were reported for different SOA conditions. In respect with the unity assumption theory –according to which modulation of multisensory integration is the result of one observer’s assumption or beliefs that multiple unisensory signals root from the same source (Chen & Spence, 2017), Thomas and Shiffrar (2013) discussed the circumstances under which the temporal relationship between auditory and visual stimuli modulates this inhibitory effect in perception. For this reason, they conducted two experiments. In their first experiment, participants were asked to identify covered-up point-light walkers while they were listening to footsteps that were either synchronous or out of-phase with the point-light footfalls. In the second experiment the rhythm factor was added in the auditory and visual streams, so this time, participants were again asked to detect point- light walkers but the auditory cues were either footsteps or tone sounds, synchronous with the point-light footfalls or temporally random (completely decoupled from the motion). The findings suggest that in all conditions relative timing of auditory and visual stimuli was not a critical factor for enhanced visual sensitivity in detecting actions, but semantic congruency was. (Thomas & Shiffrar, 2013). Eg and Behne (2015) focused on temporal integration in complex settings by adding experimental conditions closer to realistic environments and comparing perceived synchrony for long-running and eventful audiovisual sequences to single audiovisual events, for the three different contents of action, music, and speech. The researchers report better detection of asynchrony in long-running stimuli, and explain this finding on the basis of timing cues potentially arising from correspondences between multiple audiovisual events, while the variance in subjective simultaneity points between the three contents suggests that visual scene content influences temporal perception of events. Viggiano et al. (2017) examined the influence of audiovisual semantic congruency in the identification of visual stimuli belonging in different categories (living vs. non-living things) in children (6-13 years old) and adults. Four conditions were tested with only visual, congruent audiovisual, incongruent audiovisual, and only noise stimuli. The beneficial role of multisensory presentation in speed identification was not observed for children under 12 years old, but rather the interfered role of incongruent crossmodal stimuli for all children was, especially during identification of living entities.

The data collected by Viggiano et al. (2017) suggest that the level on which audiovisual interactions facilitate semantic factors is not developmentally stable, but changes across ages, referring late childhood as the starting-point for stabilization of adult-like multisensory processing. Many studies focused on the developmental trajectory of multisensory integration, examining the efficacy of audiovisual integration effects in visual perception and its role in efficient encoding and memory performance in diverse age groups (Grossmann, Striano, & Friederici, 2006; de Boer-Schellekens & Vroomen, 2013; Fiacconi et al., 2013; Adams, 2016; Ujiie et al., 2018). Specifically, Heikkilä and Tiippana (2016) investigated the impact of audiovisual encoding on recognition memory in children ($N=114$; 8, 10, 12 years old) who were asked to memorize auditory or visual

stimuli that were presented in combination with a semantically congruent, incongruent or non-semantic stimulus of the other modality. Stimuli were either pictures and sounds of natural objects, written and spoken words, visual and auditory noise. Their findings indicate that children's memory performance can be enhanced when exposed to semantically congruent audiovisual information during the encoding stage. Maguinness et al., (2011) examined the beneficial role of multisensory information in perception for older people by the addition of the semantic congruency factor in audiovisual speech. In the experiment, audiovisual sentences were presented in which the visual cue was either blurred or not blurred. The sentences were presented through digital video recordings containing a target word which was two to five words away from the end of each sentence and this word was either meaningful or meaningless (i.e. semantically congruent or incongruent with the rest sentence). The participants' task was to repeat the sentence aloud. The results suggest that additional visual information facilitates resolving auditory information, enhances the representation and, in turn, the memory recall of an unpredictable speech signal.

The significant impact of semantics on audiovisual integration in perception has been examined in various perceptual contexts, such as the perception of bistable figures. In such, Hsiao et al. (2012) indicate that background auditory soundtrack such as the voice of a young or old female can alter the predominant perception of a bistable figure such as a "wife" or "mother-in-law" figure. This crossmodal semantic effect occurred in respect to manipulation of visual fixation and showed to interact with voluntary attention. In another experiment, Fujisaki et al. (2014) found strong associations in material perception (e.g., glass, plastic, ceramic, paper) of simultaneous audiovisual stimuli. In particular, participants' material categorization of an object shown on video was influenced by the material sound (e.g., the sound of vegetable's surface when it was hit by a wooden mallet or the sound of glass when hit by a wooden mallet, etc.). Their results indicated that in irrelevant audiovisual signals the perception of the material was modified depending on the combination of both the audio sound and the visual clip (for example the sound of vegetable hit by a wooden mallet and the visual clip of a glass were combined and perceived as a video that shows a plastic bottle).

Purpose of the study

Overall, our knowledge so far supports the significant role of semantic content of a crossmodal stimulus in determining the functions and components of crossmodal information processing in the nervous system (Laurienti, et al., 2004; Xi et al., 2020). Moreover, research is also at a state to propose mechanisms underlying the process of crossmodal integration and its effects –as such of the central executive (Xie et al., 2017). However, prior experiments on the field utilized mostly stable visual cues such as pictures and were limited in the semantic content of a naturalistic sound or a single word. So, it is of great interest to examine whether the same observations occur in more realistic conditions, such as movieclips combined with auditory meaningful sentences.

The aim of the current study is to introduce more realistic aspects in the examination of crossmodal integration by simultaneously presenting short movieclips and auditory three-word sentences. The main hypothesis refers to whether and how much audiovisual semantic information affects speed and accuracy in judgments regarding the presence or absence of a target. Complementary to this, the present study is the first to introduce target absent trials which always contain target-related information either presented through vision, audition, or both senses. In such trials, audio sentences include a semantic target-related noun or verb. To date no other study investigated the potential influence of the type of semantics (verb vs. noun) on the perception of audiovisual stimuli, especially of movieclips. If target-related information is considered useful for the neural system to identify the absence of the target then we expected that performance (through speed and accuracy) would be improved, otherwise if it is considered noise, we expected to observe a compromised performance.

Chapter 2

Materials and Methods

2.1. Participants

Thirty-eight volunteers participated. Data from two subjects were excluded, one wished to not be included in the data analysis after completing the task, and the other reported that he/she faced technical issues -latency of sound during the experiment, thus, the final number of participants was Thirty-six (mean age 29.909, age range 18-60 with three missing values; 24 female and 12 male). All were with normal hearing and normal or corrected-to-normal vision. Each participant was provided written detailed instructions regarding the experimental online task, and the equipment that was required in order to run the task.

2.2. Experimental Stimuli

Visual clips

There were 12 short scenes of 1.500 ms cut out from the short movie “37 Days” directed by Nikoleta Leousi (from which we took permission of use after written conversation through her social media account -Facebook), standardized on familiarity and complexity. Standardized on familiarity refers to the scenes since they were all coming from real-word routines like walking, cooking, etc., while standardized on complexity means that all videos included at least one and not more than two movements in the scene presented. As shown in the *Table 1* (see Appendix) these were: **(a)** a hand stirring lemon juice in a glass with lemon slices on the foreground, **(b)** a hand cutting an onion, **(c)** pregnant woman caressing her belly, **(d)** children with parents at a waiting hall, **(e)** a woman and an old man sitting in a bus, **(f)** a pregnant woman arriving at a crowded bus stop, **(g)** a woman going up the stairs, **(h)** a woman walking on an uphill, **(i)** two hairdressers with the one taking of a towel from a woman in a hair salon, **(j)** a sitting man wiping a telephone on his apron, **(k)** a pregnant woman staring at a mirror and a hairdryer, **(l)** a pregnant woman arriving at a hair salon. They were presented in full-screen mode on a grey background. The size of the screen varied due to the nature of the experiment (online).

Audio clips

There were 12 semantic audio sentences in Greek language that consisted of three-word sentences, spoken by an AI voice. The AI voice sentences were conducted using the mobile application "Text To Speech" developed by STCodesApp and provided by Google

Commerce Ltd (pitch and volume was defined in 50%, and speed in 26%). In particular, the 12 sentences (translated in English) are: **(a)** “Someone cut lemons”; **(b)** “He is using the knife”; **(c)** “She is wearing a wedding ring”; **(d)** “They have three children”; **(e)** “They are sitting in the bus”; **(f)** “She arrived at the bus stop”; **(g)** “She is going up the stairs”; **(h)** “She is walking uphill”; **(i)** “She removes the towel”; **(j)** “He is wiping the telephone”; **(k)** “She found the hairdryer”; **(l)** “She arrived at the hair salon”. As shown in *Table 2*, the sentences were distinguished in two basic semantic categories according to the type of the word (verb vs. noun) which was semantically related to the target expected in the movieclip. The category *verb* included the sentences **a** for the target “knife”, **d** for the target “ring”, **j** for the target “towel”, and the category *noun* included the sentences **f** for the target “bus”, **h** for the target “stairs”, **l** for the target “hairdryer”. The duration of all sentences was identical to the duration of the short movieclips had a duration of 1500 ms.

Table 2. The experimental conditions for semantics

Target	Semantics	
	Verb	Noun
Knife	(a) “Someone <u>cut</u> lemons”	
Wedding ring	(d) “They <u>have</u> three children”	
Towel	(j) “He is <u>wiping</u> the telephone”	
Bus		(f) “She arrived at the <u>(bus) stop</u> ”*
Stairs		(h) “She is walking <u>uphill</u> ”
Hairdryer		(l) “She arrived at the <u>hair salon</u> ”

*in the Greek language the bus stop is called “στάση” which refers to the English word “stop”

2.3. Procedure

While the experiment was provided to the participants, a briefly description of the instructions was given, and they were asked whether they had any question regarding the task. They were also reminded to remove any distracting item, to run the task in a quiet place alone, and to ensure their headphones were plugged in, if using any. When participants were ready, the URL was provided to them in order to start the experimental task. Participants were first asked to insert their age, gender, and country; then detailed instructions about the task were shown through an AI voice and pictures, and finally an example trial was provided. As soon as participants were ready to start the main trials, they responded by pressing the right arrow on their keyboard. The experimental timeline is shown in detail in *Figure 1*. A fixation point (400 ms) was firstly presented, followed by a line drawing of the target stimulus that participants had to perceive in the movieclip. This was presented simultaneously with a written and a verbal label (1000 ms). Another fixation point (500 ms) followed, and the movieclip (1500 ms) was presented right after. Finally, a line drawing of the two possible response key-arrows appeared accompanied by the label “Did you see and/or hear the target?”. Participants were instructed to

answer as quickly and accurately as possible whether they perceived the target in the movieclip (through either vision/audition or both).

Figure 1. The experimental timeline

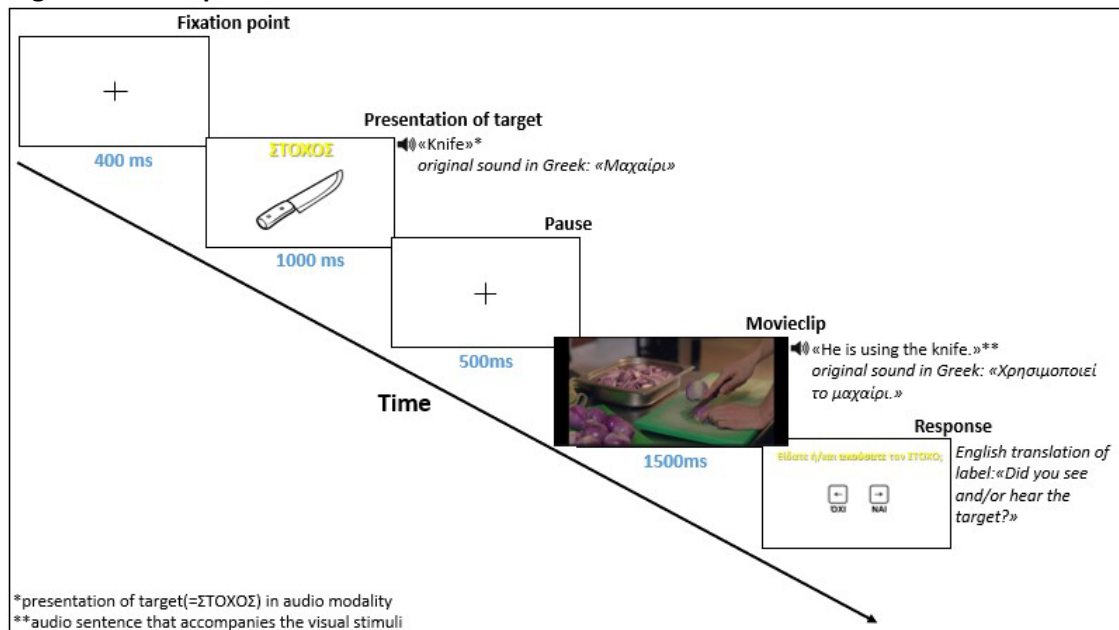


Figure 1. The example given here corresponds to a trial in which both the audio and visual stimuli included the target *Knife* (congruent movieclip). After every movieclip the picture with the draw-line response key-arrow keys appeared immediately. Response recording and reaction times (RT) started right from the moment this picture appeared (RT = inter-trial interval (ITI)).

The task was self-paced, producing a variable inter-trial interval (ITI), meaning that a new trial was not starting until a response was made. There were no breaks available since the task could not be paused. However, each participant was informed that the task can be terminated at any time by pressing the escape button on their keyboard if they wish not to continue the experiment.

2.4. Design/Experimental Conditions

The experimental task was designed using the PsychoPy software (Peirce et al., 2019) and was online running through pavlovia.org. Accuracy in responses and RT for each trial were recorded by the built-in keyboard backend PsychToolbox -Psychophysics Toolbox extensions (Kleiner et al., 2007), for data collection of keyboard input in PsychoPy. The experiment was based on two target conditions; target present and target absent, where the target absent condition included only trials with at least one modality presenting information semantically related to the target (thus, we refer to these as target-related stimuli). In addition, the target and target-related stimuli were equiprobably presented through modalities as follows: *i)* target presented only in the visual modality, *ii)* target presented only in the audio modality, *iii)* target presented in both audio and visual

modalities, *iv*) target-related only in the visual modality, *v*) target-related only in the audio modality, *vi*) target-related in both audio and visual modalities. By applying the types of stimulus (target and target-related) in the above six possible conditions of modality, we received the following audiovisual combinations: *(I)* target/target-related stimulus only presented in an audio clip accompanied by an incongruent visual clip; *(II)* target/target-related only presented in a visual clip accompanied by an incongruent audio clip; *(III)* target/target-related stimulus presented in both modalities. The first two (*I* and *II*) combinations have been grouped under the term “incongruent audiovisual stimuli” whereas the latter under the term “congruent audiovisual stimuli” (see *Table 3*). Thus, an example for incongruent audiovisual movieclip with the target stimulus in: *1a) visual modality* is a visual clip including the target “knife” accompanied by an audio sentence that does not have the target-word “knife”; *2a) audio modality* is an audio sentence including the target-word “knife” accompanied by a visual clip that does not include the target “knife”; whereas an example for congruent audiovisual movieclip with the target stimulus in *3a) both visual and audio modalities* is a visual clip that includes the target “knife” accompanied by an audio sentence including the target-word “knife”. Respectively, an example for incongruent audiovisual movieclip with the target-related stimulus in: *1b) visual modality* is a target-related visual clip for the target “knife” (i.e., a hand stirring lemon juice in a glass with lemon slices on the foreground) accompanied by an audio sentence that does not have the target-word “knife” or a target-related word to it, *2b) audio modality* is a target-related audio sentence for the target-word “knife” accompanied by a visual clip that does not include the target “knife” or a target-related object or action to it; whereas an example for congruent audiovisual movieclip with the target-related stimulus in *3b) both visual and audio modalities* is a target-related visual clip for the target “knife” accompanied by the target-related audio sentence for the target-word “knife”.

Table 3. The experimental conditions







Type of stimulus	Movieclips conditions		Type of audiovisual stimuli
Target present	1a) Target only in visual modality		Incongruent
	2a) Target only in audio modality		Incongruent
	3a) Target in both modalities		Congruent
Target absent	1b) Target-related only in visual modality		Incongruent
	2b) Target-related only in audio modality		Incongruent
	3b) Target-related in both modalities		Congruent

Table 3. The experimental conditions were distinguished in target present and target absent trials. Movieclips in target present trials were distinguished according to the

modality in which the target stimulus was presented: 1a, 2a, 3a. In target absent trials movieclips were distinguished according to the modality in which the target-related information was presented: 1b, 2b, 3b. When the target or target-related information was presented only in one of the two modalities, the movieclips were characterized as incongruent, while when it was presented in both, the movieclips were characterized as congruent.

Target absent and target trials occurred equally. The number of stimuli presented during the task was for the **A**) target present condition: six audio sentences, six visual clips, with 36 pairings of the visual and audio stimuli. However, for the **B**) target absent_(=target-related) condition there were again six audio sentences, six visual clips, but with 30 pairings plus the repetition of the six congruent audiovisual pairings (see *Table 4*). The experimental task was running in four repetitions (4 blocks). Therefore, the total number of trials was 288 (144 target trials and 144 target-related trials), which were equally distributed for each audiovisual target and target-related condition: 48 trials with target and 48 with target-related stimuli presented in audio modality, 48 trials with target and 48 with target-related stimuli presented in visual modality, 48 trials with target and 48 with target-related stimuli presented in both modalities. For the target-related trials, audio sentences were also equally distinguished with respect to the category of semantics they contained. That is, whether the one word, which was semantically related to the expected target, was a verb or a noun (24 audio sentences including a semantically related verb and 24 including a semantically related noun).

Table 4. Total number of movieclips













Type of stimulus	Combinations of stimuli			Total trials in a block:
Target present	12 visual clips with target	 	12 audio sentences without target	3 x 12 = 36
	12 visual clips without target	 	12 audio sentences with target	
	12 visual clips with target	 	12 audio sentences with target	
Target absent	6 x2 visual clips with target-related stimulus	 	12 audio sentences without target/target-related stimulus	3 x 12 = 36
	12 visual clips without target/target-related stimulus	 	6 x2 audio sentences with target-related stimulus	
	6 x2 visual clips with target-related stimulus	 	6 x2 audio sentences with target-related stimulus	

Table 4. Movieclips in target present trials were built out of the combinations of target and target absent audio and visual stimuli, following the rule that at least one modality includes the target. In target absent trials, movieclips were the outcome of the combinations of target-related and target absent stimuli, following the rule that at least one modality includes target-related information. In target absent trials, we replaced the targets with related information only for six targets. Therefore, the total target-related clips used for the combinations in each modality condition was six.

2.5. Data Analysis

Data for each participant were collected online and automatically saved in excel files from the pavlovia platform. For individual participants (for each condition), mean proportion correct and the mean reaction times (RT) were calculated and analysed using the JASP software version 0.17.1. To test for an effect of target presence on proportion correct and RT, we performed two separate two-way repeated measures analyses of variance (ANOVA), with Modality (three levels: audio, visual, audiovisual) and Target presence (two levels: target present, target absent) as factors.

We also wanted to test whether the type of a semantically related word (verb or noun) in a three-word sentence affects RT and correct responses of participants when the target is absent in both modalities. Thus, another group of two-way repeated measures ANOVA with Congruency (congruent vs. incongruent) and Semantics (verb vs. noun) as factors were conducted. In this way, we tried to test potential differences in results (proportion correct of responses and RTs) between congruent audiovisual clips when containing a target-related verb vs. noun, as well as between the incongruent and the congruent audiovisual clips with a semantically related verb, and noun respectively.

For the analysis, outlier trials that were 3 standard deviations above or below each participant's mean RT were removed. All trials (correctly and erroneously answered) were used for the analysis of RTs. In the Appendix, the RT analysis of only correctly answered trials can also be found (*Tables A12-A15* and *Figures A1* and *A2*).

Chapter 3

Results

3.1. Mean Proportion Correct

The results indicate a statistically significant difference in mean proportion correct responses depending on the presence of the target. *Figure 3* shows individual (a-c) and mean proportion (d) of correct responses for target present and target absent trials in respect to the modality in which target or target-related information was presented. *Figure 3a* and *3b* show results from movieclips with incongruent information from the visual and auditory modalities and *Figure 3c* shows results from movieclips with congruent information from these two modalities. For incongruent movieclips, we observed a similar pattern when the target was present in either of the two modalities. Specifically, when the target was present only in the audio or in the visual modality, mean proportion of correct responses was reduced ($M_{\text{audio}}= 0.647$, $SD_{\text{audio}}= 0.305$; $M_{\text{visual}}=0.841$, $SD_{\text{visual}}= 0.235$), compared to when the target was absent altogether ($M_{\text{audio}}= 0.931$, $SD_{\text{audio}}= 0.038$; $M_{\text{visual}}= 0.986$, $SD_{\text{visual}}= 0.018$) where variance in responses was also less. For the congruent condition, we observed the opposite pattern of results. When the target was present in both modalities mean proportion of correct answers was high ($M_{\text{audiovisual}}= 0.981$, $SD_{\text{audiovisual}}= 0.036$). However, in target absent trials, the mean proportion of correct responses was reduced ($M_{\text{audiovisual}}= 0.898$, $SD_{\text{audiovisual}}= 0.111$) - Appendix *Table A2*).

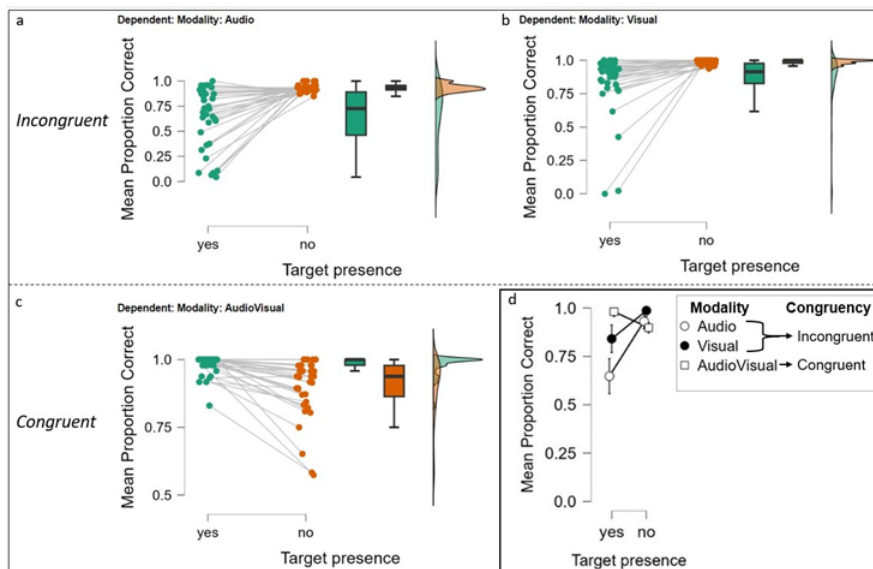


Figure 3. Mean proportion correct responses for target presence and target absence. Box-and-whisker plots in (a), (b) and (c) represent the interquartile ranges (IQRs); central, bold horizontal lines in (a), (b) and (c) represent the medians; white squares, black circles and

white circles in (d) represent the group mean proportion correct ($N=36$) in the modalities used for the movieclips. Green and orange points represent the mean proportion of correct responses of individual participants. Note that range in (c) starts from 0.5. a) shows mean proportion correct individual scores in target present and target absent incongruent trials when target/target-related information was only heard. B) shows mean proportion correct individual scores in target present and target absent incongruent trials when target/target-related information was only seen. c) shows mean proportion correct individual scores in target present and target absent congruent trials (target/target-related information was both heard and seen). d) shows mean proportion correct total scores in target present and target absent incongruent trials (target/target-related information was either only heard or only seen) vs. congruent trials (target/target-related information was both heard and seen).

Note that in target absent trials the target was completely absent in both modalities while target-related information was presented in one of the two modalities (incongruent movieclips) or in both modalities (congruent movieclips). Our results show that despite the fact that target-related information was present, this did not compromise participants' performance in incongruent movieclips (*Figure 3a, 3b*).

In a two-way repeated-measures ANOVA (Appendix *Table A3*), the main effects of Modality (audio/visual/audiovisual) and Target presence (yes/no) were found to be statistically significant ($F(2,70)=25.837$, $p < .001$ and $F(1,35)=26.940$, $p < .001$ respectively). The interaction between Modality and Target presence was also statistically significant ($F(2,70)=20.265$, $p < .001$). The post-hoc pairwise comparisons (Appendix *Table A4*) showed that the performance was statistically worse when target or target-related information was presented in audio modality during incongruent movieclips compared to visual ($t=-5.567$, $p < .001$) or audiovisual ($t=-6.722$, $p < .001$). Further post-hoc pairwise comparisons showed that the performance was superior in congruent movieclips (target present in both modalities) compared to incongruent movieclips (target present only through audio or only through vision) ($t_{A \text{ vs. } AV}=-9.094$, $p < .001$; $t_{V \text{ vs. } AV}=-3.819$, $p=0.002$). We also observed that in incongruent movieclips the presence of target, affects the accuracy in responses on a statistically significant level (target present vs. Target absent trials for audio: $t=-7.044$, $p < .001$; and for visual: $t=-3.619$, $p=0.004$). We also found a correlation between participants' age and performance (proportion of correct answers) -see *Figure A3* in Appendix.

3.1.1. Semantics

We tested for the effect of semantics only in target absent trials, since only those included target-related information (verb or noun) either in one of the modalities (incongruent movieclips) or in both (congruent movieclips). As mentioned in previous sections, in target absent trials incongruent movieclips did not contain the target in either modality, but rather one modality included information (a verb or a noun) which was semantically related to the target. For example, the target "knife" was not presented

in the visual clip nor heard in the audio sentence, rather the audio sentence included the verb “cut” (which is semantically related to the word “knife”). Respectively, in congruent movieclips the information presented in both modalities was semantically related to the target “knife”, i.e., participants were hearing the sentence “Someone cut the lemons” while watching two hands mixing a glass of juice with many lemon slices in the front ground.

Figure 4 below shows the difference in individual (a,b), as well as in total sample (c) mean proportion of correct responses between incongruent and congruent target absent trials (or target-related trials) for verb vs. noun. We observe a different pattern when the target-related audio sentence included a semantically related noun. While the pattern of results is similar for the two congruent semantic conditions ($M_{\text{noun}} = 0.903$, $SD_{\text{noun}} = 0.130$ vs. $M_{\text{verb}} = 0.892$, $SD_{\text{verb}} = 0.115$), we observe statistically superior performance for incongruent trials where target-related information was available in the form of a noun ($M_{\text{noun}} = 0.994$, $SD_{\text{noun}} = 0.015$ vs. $M_{\text{verb}} = 0.867$, $SD_{\text{verb}} = 0.075$; $t_{\text{verb vs. noun}} = -8.428$, $p < .001$) -Appendix Table A5. A two-way repeated-measures ANOVA (Appendix Table A6) was conducted to test for the effects of congruency (incongruent/congruent) and semantics (verb/noun) on the accuracy of responses. In this case, testing for congruency effects means to test whether the semantically related audio sentences presented with their identical video clips (congruent movieclips) showed statistically significant difference in performance compared to when the semantically related audio sentences were presented together with a random video clip (incongruent movieclips). In our analysis congruency did not yield a significant result ($F(1,35) = 2.580$, $p = 0.117$). On the other hand, the main effect of Semantics ($F(1,35) = 29.877$, $p < .001$) was found to be statistically significant, so did the interaction of Congruency and Semantics ($F(1,35) = 49.683$, $p < .001$).

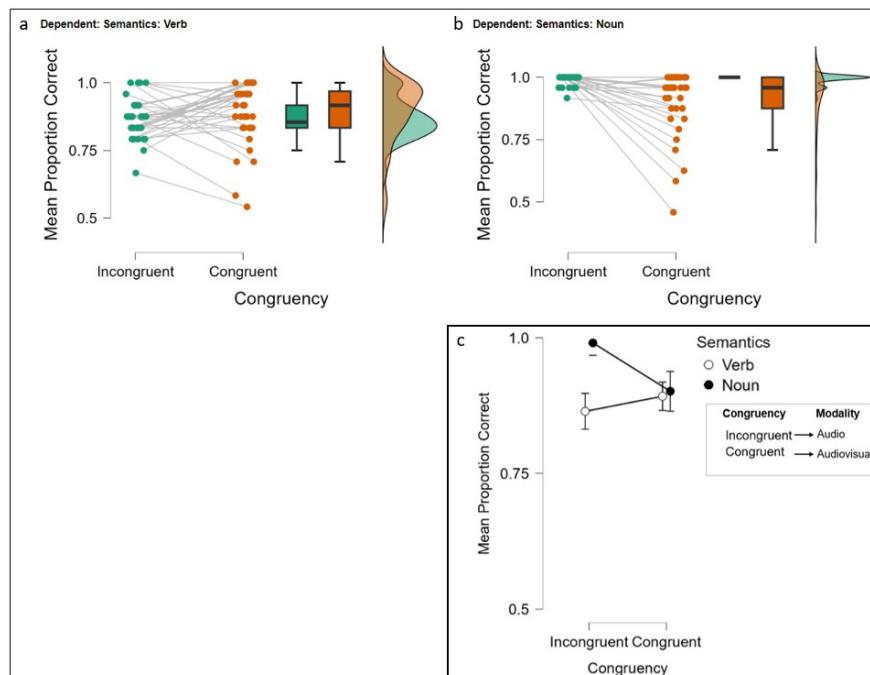


Figure 4. Mean proportion correct for incongruent or congruent semantic conditions in target absent trials. Box-and-whisker plots in (a) and (b) represent the interquartile ranges

(IQRs); central, bold horizontal lines in (a) and (b) represent the medians; black circles and white circles in (c) represent the mean proportion correct ($N=36$). Green and orange points represent the mean proportion of correct responses of individual participants. Note that the range starts from 0.5. a) shows mean proportion correct individual scores in target absent trials when target-related audio sentences included a verb accompanied by a semantically incongruent visual clip vs. a semantically congruent visual clip. b) shows mean proportion correct individual scores in target absent trials when target-related audio sentences included a noun accompanied by a semantically incongruent visual clip vs. a semantically congruent visual clip. c) shows mean proportion correct total scores in incongruent trials for verb vs. noun (audio sentence with a target related verb vs. noun accompanied by a semantically incongruent visual clip) vs. congruent trials (audio sentence with a target related verb vs. noun accompanied by a semantically congruent visual clip).

Semantically target-related noun was also associated with superior accuracy in responses when occurring in incongruent movieclips compared to when occurring in congruent movieclips regardless of whether audio included a noun ($M= 0.903$, $SD= 0.130$; $t_{\text{InNoun vs. CoNoun}}=4.144$, $p< .001$) or a verb ($M= 0.892$, $SD= 0.115$; $t_{\text{CoVerb vs. InNoun}}=-4.256$, $p< .001$) -Appendix *Table A7*, also see *Figure 4*.

3.2. Mean Reaction Times

We observed the fastest mean responses in congruent trials, meaning when the target was present audiovisually ($M_{\text{audiovisual}}= 0.520\text{s}$, $SD_{\text{audiovisual}}=0.230$). The second fastest scores were observed for the trials where the target was present only in the audio modality ($M_{\text{audio}}=0.590\text{s}$, $SD_{\text{audio}}=0.257$), whereas slower responses were shown in trials where the target was presented only in the visual modality ($M_{\text{visual}}=0.620\text{s}$, $SD_{\text{visual}}=0.358$). In target absent trials, meaning the trials where the target had been replaced by a target-related stimulus in one of the modalities or in both, mean RTs were generally slower compared to when the target was present. Here, faster responses were recorded when the target-related stimulus appeared in the visual modality ($M=0.670\text{s}$, $SD=0.302$), whereas slower responses were observed when the target-related stimulus was presented in the audio modality ($M=0.683\text{s}$, $SD=0.317$) and the slowest when presented in both modalities ($M=0.741\text{s}$, $SD=0.366$) -Appendix *Table A8*. The two-way repeated measures ANOVA (Appendix *Table A9*) showed that the performance regarding mean RTs was not associated with statistically significant differences between modality conditions ($F(2,70)=0.384$, $p=0.683$). However, the main effect of Target presence (yes/no) ($F(1,35)=20.825$, $p<.001$) as well as the interaction between Modality and Target presence ($F(2,70)=8.461$, $p< .001$) were statistically significant (see also *Figure 5*). Finally, we found a correlation between participants's age and reaction times (see *Figure A3* in Appendix).

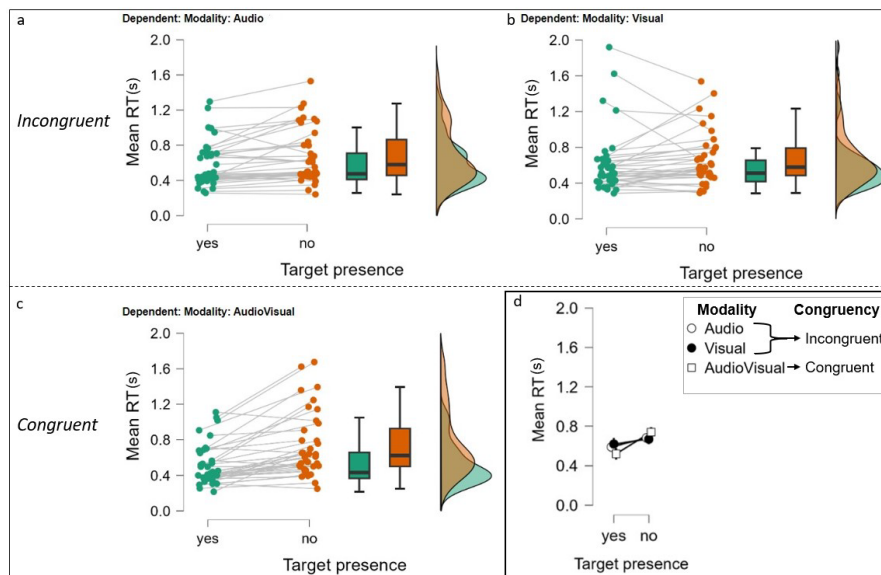


Figure 5. Mean RTs for target presence and target absence. Box-and-whisker plots in (a), (b) and (c) represent the interquartile ranges (IQRs); central, bold horizontal lines in (a), (b) and (c) represent the medians; white squares, black circles and white circles in (d) represent the group mean RTs ($N=36$) in the modalities used for the movieclips. Green and orange points represent the mean RT of individual participants. a) Mean RT individual scores in target present and target absent incongruent trials when target/target-related information was only presented through audio. b) Mean RT individual scores in target present and target absent incongruent trials when target/target-related information was only seen. c) Mean RT individual scores in target present and target absent congruent trials (target/target-related information was both heard and seen). d) Group mean RT in target present and target absent incongruent trials (target/target-related information was either only heard or only seen) vs. congruent trials (target/target-related information was both heard and seen).

3.2.1. Semantics

As we can see in *Figure 6*, the fastest mean RT scores were found in incongruent movieclips regardless of whether they included a semantically related noun ($M_{\text{noun}}=0.684\text{s}$, $SD_{\text{noun}}=0.336$) or a verb ($M_{\text{verb}}=0.681\text{s}$, $SD_{\text{verb}}=0.339$), compared to congruent movieclips ($M_{\text{verb}}=0.726\text{s}$, $SD_{\text{verb}}=0.352$; $M_{\text{noun}}=0.758\text{s}$, $SD_{\text{noun}}=0.386$) - Appendix *Table A10*.

A two-way repeated-measures ANOVA (Appendix *Table A11*) was conducted in target absent trials to test for the effects of congruency and semantics on mean RTs. A significant effect was found for Congruency ($F(1,35)=5.908$, $p=0.020$) but not for Semantics ($F(1,35)=0.782$, $p=0.382$) or the interaction between the two factors ($F(1,35)=0.410$, $p=0.526$).

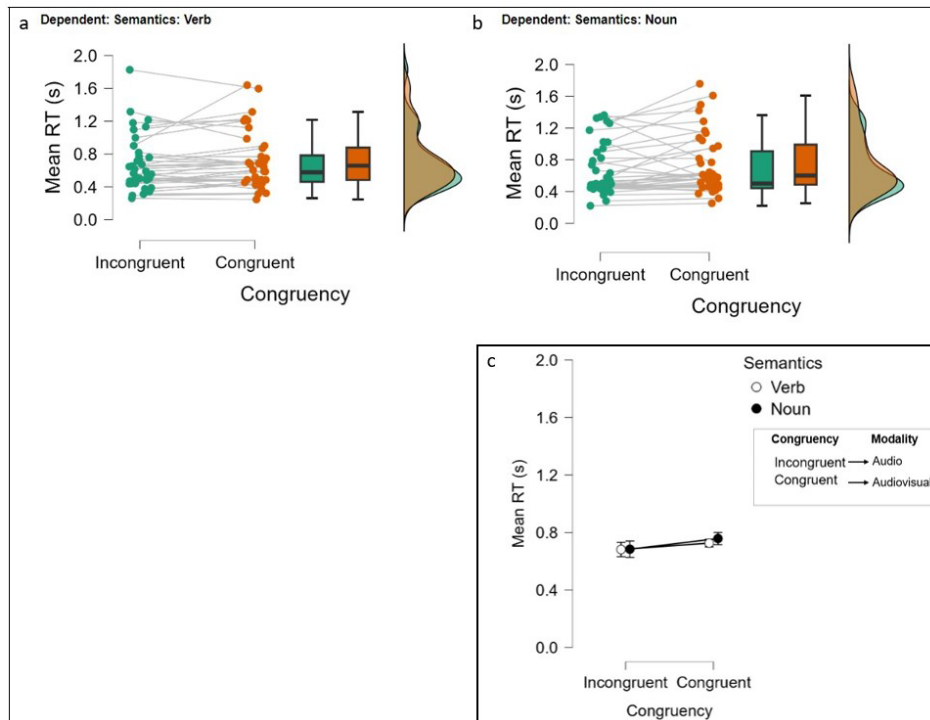


Figure 6. Mean RTs for incongruent or congruent semantic conditions when the target was absent. Box-and-whisker plots in (a) and (b) represent the interquartile ranges (IQRs); central, bold horizontal lines in (a) and (b) represent the medians; black circles and white circles in (c) represent the group mean RTs ($N=36$). Green and orange points represent the mean proportion of correct responses of individual participants. a) Mean RTs individual scores in target absent trials when target-related audio sentences included a verb accompanied by a semantically incongruent visual clip vs. by a semantically congruent visual clip. b) Mean RTs individual scores in target absent trials when target-related audio sentences included a noun accompanied by a semantically incongruent visual clip vs. by a semantically congruent visual clip. c) Mean group RTs in incongruent trials for verb vs. noun (audio sentence with a target related verb vs. noun accompanied by a semantically incongruent visual clip) vs. congruent trials (audio sentence with a target related verb vs. noun accompanied by a semantically congruent visual clip).

Chapter 4

Discussion

The present study was conducted to examine audiovisual integration in the perception of synchronous, and semantically congruent and incongruent audiovisual stimuli that include movement and spoken sentences. Furthermore, we investigated the role of semantic associations in the perception of congruent and incongruent crossmodal stimuli. Accuracy scores indicated a statistically significant difference between the modalities in which the target or target-related information was presented, as well as between target presence conditions. In accordance with previous studies, we also report a significant correlation between participants' age and accuracy scores, as well as between participants' age and mean RTs (Barrett & Newell, 2015; Smayda et al., 2016; Brooks et al., 2018).

4.1. Target Present Trials

In our study, judgements for targets' presence revealed that listening to spoken sentences that include the target stimuli while perceiving incongruent visual information (incongruent movieclips) was associated with lowest accuracy in responses. This finding is consistent with previous studies that examined the visual influences on auditory perception and revealed perceptual alterations of audio information due to their simultaneous presentation with conflicting but task-relevant visual information (Soto-Faraco et al., 2002; Soto-Faraco, Spence, & Kingstone, 2004; Bruns, 2019; Opoku-Baah et al., 2021). Our study revealed also that watching a video scene, where the target stimuli appeared while listening to incongruent spoken sentences (incongruent movieclips), was also associated with lower accuracy in judgments. Thus, our findings provide support for the suppressive effect that sound can have on visual perception (Shams, Kamitani, & Shimojo, 2000; Chen & Spence, 2010; Hidaka & Ide, 2015).

Moreover, our results indicate the complementary role of congruent audiovisual information (congruent movieclips) in perceiving target stimuli when presented in a crossmodal interface. In this case, the semantically congruent audiovisual stimuli resulted in significantly better accuracy in judgments but in not a significant difference in RTs compared to incongruent audiovisual stimuli. The superior performance when semantically congruent audiovisual stimuli were presented, has been previously reported in various experimental tasks (Laurienti et al., 2004; Molholm et al., 2004; Chen & Spence, 2010; Xie et al., 2017; Rekow et al., 2022). However, the absence of a significant difference in RTs between congruent and incongruent movieclips is somewhat surprising. Molholm et al. (2004), for example, observed a trend of

significantly better performance in an object detection task, where a combined influence of crossmodal inputs (such as line-drawing pictures and vocalizations of animals) was suggested. They reported not only significantly higher accuracy but also significantly faster target identification in congruent conditions when the picture and vocalization of same animal were matched, compared to when the target was presented only in one sensory modality. To date it is strongly supported that this significant difference in RTs is due to the semantical congruency between the two modality signals (Laurienti et al., 2004; Molholm et al, 2004; Mastroberardino, Santangelo, & Macaluso, 2015; Tsilioni & Vatakis, 2016). However, Letts, Basharat, and Barnett-Cowan (2022) bring the parameter of valence another significant factor in multisensory integration. Our results provide further evidence that semantic congruency cannot be a critical factor on its own for determining multisensory behavioural performance. We propose that the complexity of language structure (i.e., word or sentence) and its relation to the target stimulus may play the most important role, especially in real-word multimodal perception.

4.2. Target Absent Trials

In all trials where the target was absent, we included for the first time target-related information instead of completely irrelevant information. Although these trials included target-related information, this information did not compromise participants' performance. Specifically, the modality in which the target-related information appeared did not influence the performance significantly, as performance was near veridical whether target-related information was presented through vision or audition alone. As discussed so far, incongruent audio and visual information interact with each other to eventually perceive a coherent representation of that information (Roach, Heron, & McGraw, 2006; Tsilioni & Vatakis, 2016). Thus, during incongruent crossmodal signals, even though when the presented information in either modality is related to the object the observer is looking for, the coherent representation of the perceptual process remains accurate to detect the absence of the target. On the other hand, in the case of congruent movieclips, we observe decrease in performance. One possibility is that participants' judgements may have been affected by the relation between congruent crossmodal information and target-related information which in our analysis was found to be statistically significant. The presentation of target-related information in both modalities may resulted in a confusion or even an illusion of what has been seen and heard. This confusion may arise from the combination of two factors: 1) strong audiovisual integration that is formed temporally for each congruent trial, and its strength is due to the semantical congruency of audiovisual information, together with 2) strong semantic schemata/concepts that exist between objects, locations, actions, movements etc. As discussed so far regarding the first factor (1), congruent audiovisual stimuli tend to build stronger connections between the perceived information and therefore facilitate perception. These strong connections have been also indicated through their efficient neural representations in the neuroanatomical surface (Li et al.,

2011). As for the second factor (2): The strength of semantic concepts has been well-established for settings that examine attention performance such as visual search tasks, where research indicates the significant influence of semantic concepts even when there is low accuracy of the concept detectors (Long & Chang, 2014). In addition to this, a strong context-dependent association of audiovisual integration with multiple interactions in various brain regions has been reported (Diaconescu, Alain, & McIntosh, 2011; Gao et al, 2022). In this matter, by combining factors (1) and (2) we could assume that semantics could create an effect of semantic relativity on perceptual performance in congruent audiovisual movieclips depending on features such as the complexity of combined semantic information, the expectations of the observer, etc. which in turn activates more complex integrated brain processes. The significantly slower response times in our results when target-related information was presented in congruent movieclips compared to when the target was, could be another evidence of this effect.

4.3. The Role of Semantics (Verb vs. Noun) in Target Absent Trials

Furthermore, our experiment examined whether hearing a sentence that includes a word which is semantically related to the target stimuli, could affect judgements in incongruent and congruent movieclips. To date, it is the first study that used semantically related 'distractors' and replaced target absent trials with target-related ('distractor') trials. We grouped participant responses based on the type of the semantically target-related word that the audio sentences included: verbs or nouns. The results indicate that in congruent movieclips, regardless of whether they included a semantically target-related verb or noun, participant performance was similar and was accompanied with slower responses. These results suggest that the potential confusion during congruent movieclips in target absent trials, which was described above to explain the decrease in performance, was not associated with the type of semantics (whether a semantically target-related verb or noun was presented). On the contrary, in incongruent movieclips when listening to a sentence that includes a semantically related noun performance was almost impeccable (M=99.4%), whereas when the audio sentence includes a semantically related verb the accuracy of judgments decreased significantly. The first condition seems to agree with the general results of incongruent movieclips in target absent trials (4.2). Respective to what has been previously mentioned, this pattern could suggest that target-related nouns did not work as distractors for correctly judging the absence of the target in incongruent movieclips. This might be due to the fact that participants were able to make easier comparisons between the noun that they have heard while seeing an irrelevant clip and the target noun that they were looking to hear and/or see. In general, nouns differ from verbs in the information level they can add in a sentence, but both cooperate to establish neural representations of objects and events (Faroqi-Shah, Sebastian, & Woude, 2018). For instance, nouns are related to objects, or subjects who perform actions, and can complete the meaning of actions, whereas verbs refer to actions and events, including

also -in many languages- temporal information about the actions, and thus indicating the syntactic complexity of verbs (Geng et al., 2022; de Aguiar & Rofes, 2022). According to King and Gentner (2019), semantic context adaptations for verbs show to be driven by online adjustments whereas for nouns by sense-selection. Maguire et al. (2015) provide evidence for higher neural activity demands in action-verb based identification compared to object-noun. In addition to this, interesting assumptions arise by theories of semantic change regarding reinterpretation or form-meaning remapping of listeners depending on task demands (Dubossarsky, Weinshall, & Grossman, 2016). A frequent observation of verbs' meaning adaptations has been reported which does not depend on the polysemy of verbs, rather on semantic strain contexts (King & Gentner, 2019). Taking these into consideration, we can propose that our results may at some level provide further evidence of the semantic complexity of verbs and the flexible nature of their cognitive representations compared to the simpler and more stable nature of nouns.

4.4. General Comments and Limitations

As mentioned in the Data Analysis session, the analysis for RTs was conducted including all participants' trials (correctly and erroneously answered). However, even when analysing only the correctly answered trials the results were found to be similar (see Appendix *Tables A8-11* and *Tables 12-15*; Results *Figures 5* and *6* and Appendix *Figures A1* and *A2*).

Furthermore, it is important to mention that our analysis did not examine the movement factor, which was included in the real-word based movieclips, since our design did not give weight to effects of movement on crossmodal-based behavioural responses. However, research focusing on language accounts, suggests that meanings of words depend on their perceptual and motor representations (Faroqi-Shah, Sebastian, & Woude, 2018) Thus, it may be worthy to further investigate for potential influence of movement in perceived visual clips, especially for the case of semantical 'distractors' (target-related stimuli) and test whether motor representations linked with verbs influence in any way audiovisual integration.

Chapter 5

Conclusion

Our study tested for existence of semantic congruency effects, as well as the appearance of a target either only in the visual clip (video) or only in the audio sentence or in both influences the performance of human participants in target detection judgments. These results were compared with those of target absent trials, which included target-related information either only in the visual clip/the audio sentence or in both. Our findings come in alignment with previous research that support enhancement in performance when semantically congruent audiovisual information is simultaneously presented. Moreover, we provide further evidence of the effect of audio modality on visual and vice versa during crossmodal perceptual integration when using movieclips and, for the first time, audio sentences. To date it is also the first time a study involves target-related information to test how this may affect the perceptual integration process, in particular accuracy in judgments and RT. Thus, our approach focused on semantic congruency effects and complexity of semantics in target absent trials in order to examine whether seeing a target-related video accompanied by an audio sentence that includes a target-related verb vs. noun could impact performance. We tested this hypothesis also in comparison with the incongruent condition (non-target visual clip accompanied with audio sentence including target-related verb vs. noun). Our results indicate the critical role of complexity of semantics in crossmodal perceptual integration and could further support the assumption of cognitive representations induced by verbs being more elastic, and therefore, concluding in more complex associations with other verbs, nouns, objects, locations etc. compared to the simpler and more stable cognitive representations that depend on nouns.

The current study adds another important real-world aspect, this of sentences combined with durative-moving visual stimuli. The findings can be applied to improve teaching tools and methods by taking advantage of the information processing enhancement induced by semantically congruent audiovisual inputs. Research on the topic, could expand in examining methods that can take advantage of the impact of semantic complexity, and specifically the efficacy of related nouns, in incongruent audiovisual settings. Moreover, based on previous behavioural studies suggesting a visual illusion induced by the integration effect of sound, and other studies that report the influence of semantics in decision-making responses, our findings could in turn give rise to future research in AI design.

APPENDIX

Appendix Tables

Table A1. The audio and visual stimuli













	Modality		
	Audio	Visual	
Stimuli	(a)	"Someone cut lemons"	
	(b)	"He is using the knife"	
	(c)	"She is wearing a wedding ring"	
	(d)	"They have three children"	
	(e)	"They are sitting in the bus"	
	(f)	"She arrived at the bus stop"	
	(g)	"She is going up the stairs"	
	(h)	"She is walking uphill"	
	(i)	"She removes the towel"	
	(j)	"He is wiping the telephone"	
	(k)	"She found the hairdryer"	
	(l)	"She arrived at the hairsaloon"	

Table A2. Descriptives for mean proportion correct

Modality	Target presence	N	Mean	SD	SE	Coefficient of variation
Audio	yes	36	0.647	0.305	0.051	0.471
	no	36	0.931	0.038	0.006	0.041
Visual	yes	36	0.841	0.235	0.039	0.280
	no	36	0.986	0.018	0.003	0.018
AudioVisual	yes	36	0.981	0.036	0.006	0.037
	no	36	0.898	0.111	0.019	0.124

Table A3. two-way repeated measures ANOVA for mean proportion correct

Within Subjects Effects

Cases	Sphericity Correction	Sum of Squares	df	Mean Square	F	p
Modality	None	0.929 ^a	2.000 ^a	0.464 ^a	25.837 ^a	< .001 ^a
	Greenhouse-Geisser	0.929	1.452	0.640	25.837	< .001
Residuals	None	1.258	70.000	0.018		
	Greenhouse-Geisser	1.258	50.817	0.025		
Target presence	None	0.722	1.000	0.722	26.940	< .001
Residuals	None	0.938	35.000	0.027		
Modality * Target presence	None	1.232	2.000	0.616	20.265	< .001
	Greenhouse-Geisser	1.232	1.749	0.705	20.265	< .001
Residuals	None	2.128	70.000	0.030		
	Greenhouse-Geisser	2.128	61.199	0.035		

Note. Sphericity corrections not available for factors with 2 levels.

Note. Type III Sum of Squares

^a Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).**Table A4. Post Hoc Tests for mean proportion correct**

Post Hoc Comparisons - Modality

		Mean Difference	SE	t	Pholm
Audio	Visual	-0.124	0.022	-5.567	< .001
	AudioVisual	-0.150	0.022	-6.722	< .001
Visual	AudioVisual	-0.026	0.022	-1.155	0.252

Note. P-value adjusted for comparing a family of 3

Note. Results are averaged over the levels of: Target presence

Post Hoc Comparisons - Modality * Target presence

		Mean Difference	SE	t	Pholm
Audio, yes	Visual, yes	-0.193	0.037	-5.275	< .001
	AudioVisual, yes	-0.333	0.037	-9.094	< .001
	Audio, no	-0.284	0.040	-7.044	< .001
	Visual, no	-0.339	0.036	-9.490	< .001
	AudioVisual, no	-0.251	0.036	-7.017	< .001
Visual, yes	AudioVisual, yes	-0.140	0.037	-3.819	0.002
	Audio, no	-0.090	0.036	-2.528	0.101
	Visual, no	-0.146	0.040	-3.619	0.004
	AudioVisual, no	-0.057	0.036	-1.606	0.553
AudioVisual, yes	Audio, no	0.050	0.036	1.390	0.553
	Visual, no	-0.006	0.036	-0.162	0.872
	AudioVisual, no	0.083	0.040	2.051	0.257
Audio, no	Visual, no	-0.055	0.037	-1.512	0.553
	AudioVisual, no	0.033	0.037	0.899	0.741
Visual, no	AudioVisual, no	0.088	0.037	2.411	0.121

Note. P-value adjusted for comparing a family of 15

Table A5. Descriptives for mean proportion correct in semantics and congruency in target absent trials

Congruency	Semantics	N	Mean	SD	SE	Coefficient of variation
Incongruent	Verb	36	0.867	0.075	0.013	0.087
	Noun	36	0.994	0.015	0.002	0.015
Congruent	Verb	36	0.892	0.115	0.019	0.129
	Noun	36	0.903	0.130	0.022	0.143

Table A6. two-way repeated measures ANOVA for mean proportion correct in semantics and congruency in target absent trials

Within Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p
Congruency	0.038	1	0.038	2.580	0.117
Residuals	0.517	35	0.015		
Semantics	0.172	1	0.172	29.877	< .001
Residuals	0.201	35	0.006		
Congruency * Semantics	0.122	1	0.122	49.683	< .001
Residuals	0.086	35	0.002		

Note. Type III Sum of Squares

Table A7. Post Hoc Tests for mean proportion correct in semantics and congruency in target absent trials

Post Hoc Comparisons - Congruency * Semantics

		Mean Difference	SE	t	Pholm
Incongruent, Verb	Congruent, Verb	-0.026	0.022	-1.168	0.497
	Incongruent, Noun	-0.127	0.015	-8.428	< .001
	Congruent, Noun	-0.037	0.024	-1.531	0.394
Congruent, Verb	Incongruent, Noun	-0.102	0.024	-4.256	< .001
	Congruent, Noun	-0.011	0.015	-0.729	0.497
Incongruent, Noun	Congruent, Noun	0.091	0.022	4.144	< .001

Note. P-value adjusted for comparing a family of 6

Table A8. Descriptives for mean RTs

Modality	Target presence	N	Mean	SD	SE	Coefficient of variation
Audio	yes	36	0.590	0.257	0.043	0.437
	no	36	0.683	0.317	0.053	0.464
Visual	yes	36	0.620	0.358	0.060	0.577
	no	36	0.670	0.302	0.050	0.450
AudioVisual	yes	36	0.520	0.230	0.038	0.443
	no	36	0.741	0.366	0.061	0.494

Table A9. two-way repeated measures ANOVA for mean RTs

Within Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p
Modality	0.008	2	0.004	0.384	0.683
Residuals	0.696	70	0.010		
Target presence	0.797	1	0.797	20.825	< .001
Residuals	1.339	35	0.038		
Modality * Target presence	0.284	2	0.142	8.461	< .001
Residuals	1.176	70	0.017		

Note. Type III Sum of Squares

Table A10. Descriptives for mean RTs in semantics and congruency in target absent trials

Congruency	Semantics	N	Mean	SD	SE	Coefficient of variation
Incongruent	Verb	36	0.681	0.339	0.057	0.498
	Noun	36	0.684	0.336	0.056	0.491
Congruent	Verb	36	0.726	0.352	0.059	0.485
	Noun	36	0.758	0.386	0.064	0.510

Table A11. two-way repeated measures ANOVA for mean RTs in semantics and congruency in target absent trials
Within Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p
Congruency	0.127	1	0.127	5.908	0.020
Residuals	0.752	35	0.021		
Semantics	0.010	1	0.010	0.782	0.382
Residuals	0.462	35	0.013		
Congruency * Semantics	0.008	1	0.008	0.410	0.526
Residuals	0.673	35	0.019		

Note. Type III Sum of Squares

Table A12. Descriptives for mean RTs (only correct answers)

Modality	Target presence	N	Mean	SD	SE	Coefficient of variation
Audio	yes	35	0.572	0.274	0.046	0.479
	no	35	0.692	0.309	0.052	0.446
Visual	yes	35	0.632	0.428	0.072	0.677
	no	35	0.672	0.290	0.049	0.432
AudioVisual	yes	35	0.520	0.231	0.039	0.445
	no	35	0.692	0.308	0.052	0.445

Table A13. two-way repeated measures ANOVA for mean RTs (only correct answers)

Within Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p
Modality	0.074 ^a	2 ^a	0.037 ^a	2.231 ^a	0.115 ^a
Residuals	1.128	68	0.017		
Target presence	0.641	1	0.641	9.858	0.003
Residuals	2.211	34	0.065		
Modality * Target presence	0.158 ^a	2 ^a	0.079 ^a	4.202 ^a	0.019 ^a
Residuals	1.282	68	0.019		

Note. Type III Sum of Squares

^a Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).

Table A14. Descriptives for mean RTs in semantics and congruency in target absent trials (only correct answers)

Congruency	Semantics	N	Mean	SD	SE	Coefficient of variation
Incongruent	Verb	36	0.683	0.322	0.054	0.472
	Noun	36	0.684	0.334	0.056	0.489
Congruent	Verb	36	0.727	0.389	0.065	0.535
	Noun	36	0.768	0.447	0.075	0.582

Table A15. two-way repeated measures ANOVA for mean RTs in semantics and congruency in target absent trials (only correct answers)

Within Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p
Congruency	0.148	1	0.148	4.111	0.050
Residuals	1.260	35	0.036		
Semantics	0.016	1	0.016	1.243	0.272
Residuals	0.437	35	0.012		
Congruency * Semantics	0.015	1	0.015	0.871	0.357
Residuals	0.606	35	0.017		

Note. Type III Sum of Squares

Appendix Figures

Figure A1. Raincloud plots for Mean RTs (only correct answers)

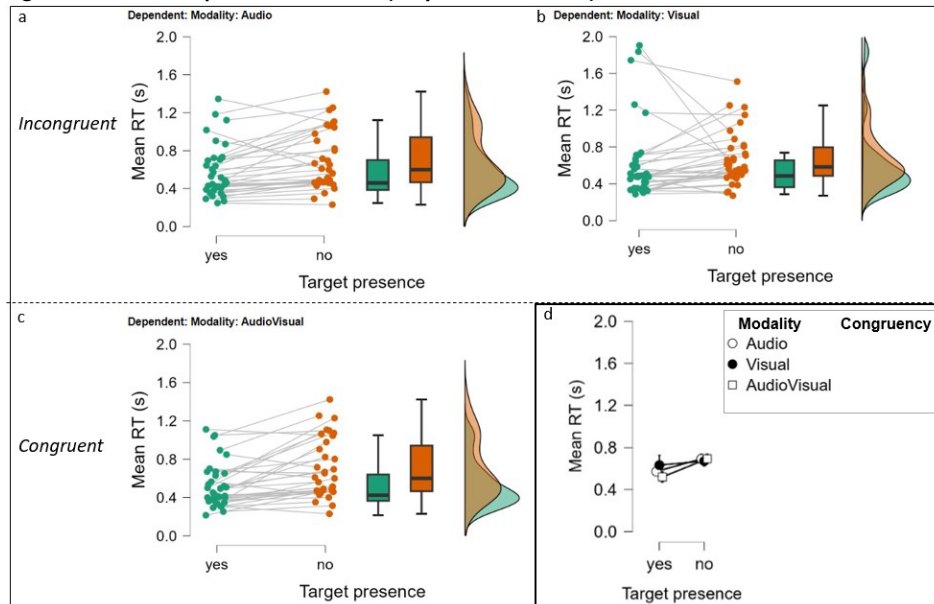


Figure A1. Mean RTs for target presence and target absence, analyzing only trials that were correctly answered. Box-and-whisker plots in (a), (b) and (c) represent the interquartile ranges (IQRs); central, bold horizontal lines in (a), (b) and (c) represent the medians; white squares, black circles and white circles in (d) represent the group mean RTs ($N=36$) in the modalities used for the movieclips. Green and orange points represent the mean RT of individual participants. a) Mean RT individual scores in target present and target absent incongruent trials when target/target-related information was only presented through audio. b) Mean RT individual scores in target present and target absent incongruent trials when target/target-related information was only seen. c) Mean RT individual scores in target present and target absent congruent trials (target/target-related information was both heard and seen). d) Group mean RT in target present and target absent incongruent trials (target/target-related information was either only heard or only seen) vs. congruent trials (target/target-related information was both heard and seen).

Figure A2. Raincloud plots for Mean RTs in target absent trials (only correct answers)

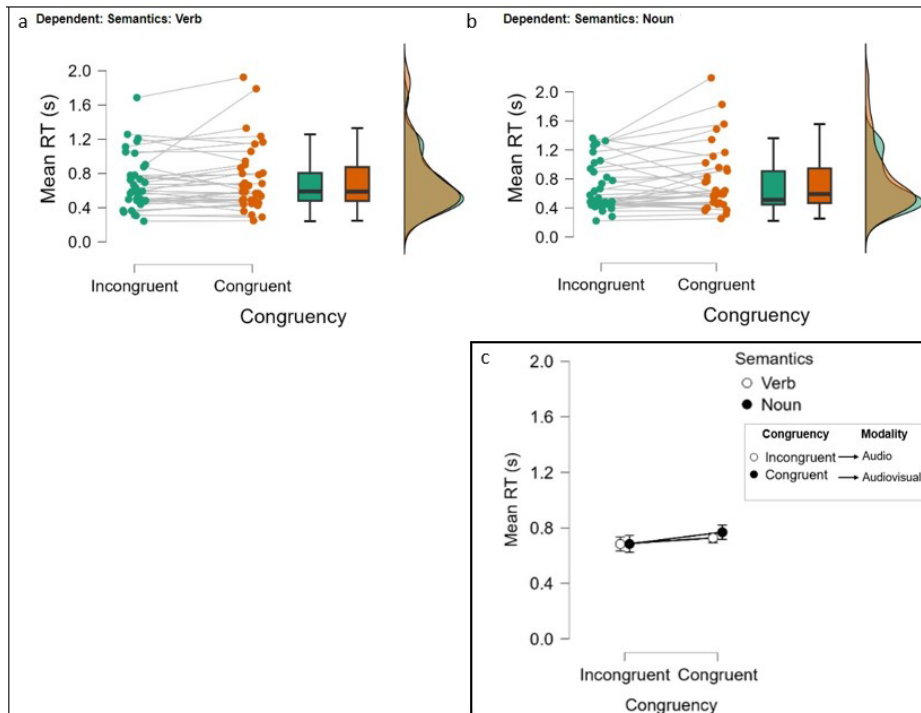


Figure A2. Mean RTs for incongruent or congruent semantic conditions when the target was absent, analyzing only trials that were correctly answered. Box-and-whisker plots in (a) and (b) represent the interquartile ranges (IQRs); central, bold horizontal lines in (a) and (b) represent the medians; black circles and white circles in (c) represent the group mean RTs ($N=36$). Green and orange points represent the mean proportion of correct responses of individual participants. a) Mean RTs individual scores in target absent trials when target-related audio sentences included a verb accompanied by a semantically incongruent visual clip vs. by a semantically congruent visual clip. b) Mean RTs individual scores in target absent trials when target-related audio sentences included a noun accompanied by a semantically incongruent visual clip vs. by a semantically congruent visual clip. c) Mean group RTs in incongruent trials for verb vs. noun (audio sentence with a target related verb vs. noun accompanied by a semantically incongruent visual clip) vs. congruent trials (audio sentence with a target related verb vs. noun accompanied by a semantically congruent visual clip).

Figure A3. Correlation plots of mean proportion correct and mean RT vs. age

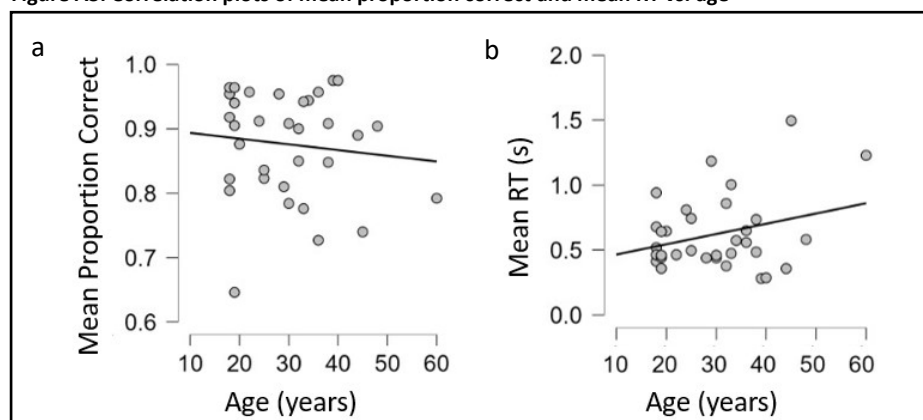


Figure A3. Mean proportion correct responses and mean RTs for age, respectively. The lines in (a) and (b) represent the direction of means of total sample, that is how much the y value (mean proportion correct/mean RT) increases or decreases across the x value (age). Grey points represent the means of individual participants. a) shows the correlation between total sample's ($n=36$) mean proportion correct scores and age. b) shows the correlation between total sample's mean RT scores and age.

References

- Adams, W. J. (2016). The Development of Audio-Visual Integration for Temporal Judgements. *PLOS Computational Biology*, 12(4), e1004865. <https://doi.org/10.1371/journal.pcbi.1004865>
- Arnold, D. H., Johnston, A., & Nishida, S. (2005). Timing sight and sound. *Vision Research*, 45(10), 1275–1284. <https://doi.org/10.1016/j.visres.2004.11.014>
- Ball, F., Nentwich, A., & Noesselt, T. (2022). Cross-modal perceptual enhancement of unisensory targets is uni-directional and does not affect temporal expectations. *Vision research*, 190, 107962. <https://doi.org/10.1016/j.visres.2021.107962>
- Barrett, M. M., & Newell, F. N. (2015). Task-Specific, Age Related Effects in the Cross-Modal Identification and Localisation of Objects. *Multisensory research*, 28(1-2), 111–151. <https://doi.org/10.1163/22134808-00002479>
- Bolognini, N., Convento, S., Fusaro, M., & Vallar, G. (2013). The sound-induced phosphene illusion. *Experimental brain research*, 231(4), 469–478. <https://doi.org/10.1007/s00221-013-3711-1>
- Brandman, T., Avancini, C., Leticevscaia, O., Peelen, M.V. (2020). Auditory and Semantic Cues Facilitate Decoding of Visual Object Category in MEG, *Cerebral Cortex*, 30(2), Pages 597– 606, <https://doi.org/10.1093/cercor/bhz110>
- Bresciani, J.-P., Dammeier, F., & Ernst, M. O. (2008). Tri-modal integration of visual, tactile and auditory signals for the perception of sequences of events. *Brain Research Bulletin*, 75(6), 753–760. <https://doi.org/10.1016/j.brainresbull.2008.01.009>
- Brooks, C. J., Chan, Y. M., Anderson, A. J., & McKendrick, A. M. (2018). Audiovisual Temporal Perception in Aging: The Role of Multisensory Integration and Age-Related Sensory Loss. *Frontiers in human neuroscience*, 12, 192. <https://doi.org/10.3389/fnhum.2018.00192>
- Bruns, P. (2019). The Ventriloquist Illusion as a Tool to Study Multisensory Processing: An Update. *Frontiers in Integrative Neuroscience*, 13(51). <https://doi.org/10.3389/fnint.2019.00051>
- Bushara, K. O., Hanakawa, T., Immisch, I., Toma, K., Kansaku, K., & Hallett, M. (2003). Neural correlates of cross-modal binding. *Nature neuroscience*, 6(2), 190–195. <https://doi.org/10.1038/nn993>
- Calvert, G. A. (2001). Crossmodal Processing in the Human Brain: Insights from Functional Neuroimaging Studies. *Cerebral Cortex*, 11(12), 1110–1123. <https://doi.org/10.1093/cercor/11.12.1110>

- Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current biology : CB*, *10*(11), 649–657. [https://doi.org/10.1016/s0960-9822\(00\)00513-3](https://doi.org/10.1016/s0960-9822(00)00513-3)
- Cao, Y., Summerfield, C., Park, H., Giordano, B. L., & Kayser, C. (2019). Causal Inference in the Multisensory Brain. *Neuron*, *102*(5), 1076-1087.e8. <https://doi.org/10.1016/j.neuron.2019.03.043>
- Chen, Y.-C., & Spence, C. (2010). When hearing the bark helps to identify the dog: Semantically congruent sounds modulate the identification of masked pictures. *Cognition*, *114*(3), 389–404. <https://doi.org/10.1016/j.cognition.2009.10.012>
- Chen, Y. C., & Spence, C. (2017). Assessing the Role of the 'Unity Assumption' on Multisensory Integration: A Review. *Frontiers in psychology*, *8*, 445. <https://doi.org/10.3389/fpsyg.2017.00445>
- Chen, Y.-C., & Spence, C. (2018). Audiovisual semantic interactions between linguistic and nonlinguistic stimuli: The time-courses and categorical specificity. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(10), 1488–1507. <https://doi.org/10.1037/xhp0000545>
- Choi, I., Lee, J.-Y., & Lee, S.-H. (2018). Bottom-up and top-down modulation of multisensory integration. *Current Opinion in Neurobiology*, *52*, 115–122. <https://doi.org/10.1016/j.conb.2018.05.002>
- Cox, D., & Hong, S. W. (2015). Semantic-based crossmodal processing during visual suppression. *Frontiers in Psychology*, *6*. <https://doi.org/10.3389/fpsyg.2015.00722>
- de Aguiar, V., & Rofes, A. (2022). The noun-verb distinction. *Handbook of clinical neurology*, *187*, 245–262. <https://doi.org/10.1016/B978-0-12-823493-8.00006-7>
- de Boer-Schellekens, L., & Vroomen, J. (2013). Multisensory integration compensates loss of sensitivity of visual temporal order in the elderly. *Experimental Brain Research*, *232*(1), 253–262. <https://doi.org/10.1007/s00221-013-3736-5>
- Diaconescu, A. O., Alain, C., & McIntosh, A. R. (2011). The co-occurrence of multisensory facilitation and cross-modal conflict in the human brain. *Journal of Neurophysiology*, *106*(6), 2896–2909. <https://doi.org/10.1152/jn.00303.2011>
- Dubossarsky, H., Weinshall, D., & Grossman, E. (2016). Verbs change more than nouns: a bottom-up computational approach to semantic change. *Lingue E Linguaggio*, *1*, 7–28. <https://doi.org/10.1418/83652>
- Eg, R., & Behne, D. M. (2015). Perceived synchrony for realistic and dynamic audiovisual events. *Frontiers in Psychology*, *6*, Article 736. <https://doi.org/10.3389/fpsyg.2015.00736>
- Faroqi-Shah, Y., Sebastian, R., & Woude, A. V. (2018). Neural representation of word categories is distinct in the temporal lobe: An activation likelihood analysis. *Human brain mapping*, *39*(12), 4925–4938. <https://doi.org/10.1002/hbm.24334>

- Feher, J. (2012). 4.3 - Cutaneous Sensory Systems. In *Quantitative Human Physiology: An Introduction* (pp. 321–331). 1st Edition. Elsevier Inc. <https://doi.org/10.1016/B978-0-12-382163-8.00035-9>
- Fiacconi, C. M., Harvey, E. C., Sekuler, A. B., & Bennett, P. J. (2013). The Influence of Aging on Audiovisual Temporal Order Judgments. *Experimental Aging Research*, *39*(2), 179–193. <https://doi.org/10.1080/0361073x.2013.761896>
- Follmann, R., Goldsmith, C. J., & Stein, W. (2018). Multimodal sensory information is represented by a combinatorial code in a sensorimotor system. *PLoS biology*, *16*(10), e2004527. <https://doi.org/10.1371/journal.pbio.2004527>
- Fujisaki, W., Goda, N., Motoyoshi, I., Komatsu, H., & Nishida, S. (2014). Audiovisual integration in the human perception of materials. *Journal of Vision*, *14*(4), 12–12. <https://doi.org/10.1167/14.4.12>
- Gao, C., Xie, W., Green, J. J., Wedell, D. H., Jia, X., Guo, C., & Shinkareva, S. V. (2021). Evoked and induced power oscillations linked to audiovisual integration of affect. *Biological psychology*, *158*, 108006. <https://doi.org/10.1016/j.biopsycho.2020.108006>
- Geng, S., Molinaro, N., Timofeeva, P., Quiñones, I., Carreiras, M., & Amoroso, L. (2022). Oscillatory dynamics underlying noun and verb production in highly proficient bilinguals. *Scientific Reports*, *12*(1). <https://doi.org/10.1038/s41598-021-04737-z>
- Grossmann, T., Striano, T., & Friederici, A. D. (2006). Crossmodal integration of emotional information from face and voice in the infant brain. *Developmental Science*, *9*(3), 309–315. <https://doi.org/10.1111/j.1467-7687.2006.00494.x>
- Heikkilä, J., & Tiippana, K. (2016). School-aged children can benefit from audiovisual semantic congruency during memory encoding. *Experimental Brain Research*, *234*(5), 1199–1207. <https://doi.org/10.1007/s00221-015-4341-6>
- Hidaka, S., & Ide, M. (2015). Sound can suppress visual perception. *Scientific reports*, *5*, 10483. <https://doi.org/10.1038/srep10483>
- Hutmacher, F. (2019). Why is there so much more research on vision than on any other sensory modality? *Frontiers in Psychology*, *10*. <https://doi.org/10.3389/fpsyg.2019.02246>
- Hsiao, J.-Y., Chen, Y.-C., Spence, C., & Yeh, S.-L. (2012). Assessing the effects of audiovisual semantic congruency on the perception of a bistable figure. *Consciousness and Cognition: An International Journal*, *21*(2), 775–787. <https://doi.org/10.1016/j.concog.2012.02.001>
- Jensen, A., Merz, S., Spence, C., & Frings, C. (2020). Perception it is: Processing level in multisensory selection. *Attention, perception & psychophysics*, *82*(3), 1391–1406. <https://doi.org/10.3758/s13414-019-01830-4>

- Kakutani, Y., Narumi, T., Kobayakawa, T., Kawai, T., Kusakabe, Y., Kunieda, S., & Wada, Y. (2017). Taste of breath: the temporal order of taste and smell synchronized with breathing as a determinant for taste and olfactory integration. *Scientific reports*, *7*(1), 8922. <https://doi.org/10.1038/s41598-017-07285-7>
- Kammers, M. P. M., de Vignemont, F., Verhagen, L., & Dijkerman, H. C. (2009). The rubber hand illusion in action. *Neuropsychologia*, *47*(1), 204–211. <https://doi.org/10.1016/j.neuropsychologia.2008.07.028>
- Kang, N., Sah, Y. J., & Lee, S. (2021). Effects of visual and auditory cues on haptic illusions for active and passive touches in mixed reality. *International Journal of Human-Computer Studies*, *150*, 102613. <https://doi.org/10.1016/j.ijhcs.2021.102613>
- Keetels, M., & Vroomen, J. (2011). Sound affects the speed of visual processing. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(3), 699–708. <https://doi.org/10.1037/a0020564>
- King, D., & Gentner, D. (2019). Polysemy and Verb Mutability: Differing Processes of Semantic Adjustment for Verbs and Nouns. *Annual Meeting of the Cognitive Science Society*. Retrieved from: https://groups.psych.northwestern.edu/gentner/papers/KingGentner_2019-Polysemy.pdf
- Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in psychtoolbox-3. *Perception*, *36*(14), 1-16
- Koubaa, Y., & Eleuch, A. (2020). Multimodal Perceptual Processing of Cues In Food Ads: Do You Smell What You See? *Journal of Advertising Research*, *61*(1), JAR-2020-006. <https://doi.org/10.2501/jar-2020-006>
- Kubovy, M., & Schutz, M. (2010). Audio-visual objects. *Review of Philosophy and Psychology*, *1*(1), 41–61. <https://doi.org/10.1007/s13164-009-0004-5>
- Lachs, L. (2023). Multi-modal perception. In R. Biswas-Diener & E. Diener (Eds), *Noba textbook series: Psychology*. Champaign, IL: DEF publishers. Retrieved from <http://noba.to/cezw4qyn>
- Lalanne, C., & Lorenceau, J. (2004). Crossmodal integration for perception and action. *Journal of physiology, Paris*, *98*(1-3), 265–279. <https://doi.org/10.1016/j.jphysparis.2004.06.001>
- Laurienti, P. J., Kraft, R. A., Maldjian, J. A., Burdette, J. H., & Wallace, M. T. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental Brain Research*, *158*(4). <https://doi.org/10.1007/s00221-004-1913-2>
- Letts, E., Basharat, A., & Barnett-Cowan, M. (2022). Evaluating the Effect of Semantic Congruency and Valence on Multisensory Integration. *Multisensory research*, *35*(4), 309–334. <https://doi.org/10.1163/22134808-bja10073>

- Li, Y., Wang, G., Long, J., Yu, Z., Huang, B., Li, X., Yu, T., Liang, C., Li, Z., & Sun, P. (2011). Reproducibility and discriminability of brain patterns of semantic categories enhanced by congruent audiovisual stimuli. *PloS one*, *6*(6), e20801. <https://doi.org/10.1371/journal.pone.0020801>
- Li, Q., Xi, Y., Zhang, M., Liu, L., & Tang, X. (2019). Distinct Mechanism of Audiovisual Integration With Informative and Uninformative Sound in a Visual Detection Task: A DCM Study. *Frontiers in Computational Neuroscience*, *13*. <https://doi.org/10.3389/fncom.2019.00059>
- Lindborg, A., Baart, M., Stekelenburg, J. J., Vroomen, J., & Andersen, T. S. (2019). Speech-specific audiovisual integration modulates induced theta-band oscillations. *PloS one*, *14*(7), e0219744. <https://doi.org/10.1371/journal.pone.0219744>
- Long, B., & Chang, Y. (2014). Chapter 4 - Visual Search Ranking. In *Relevance Ranking for Vertical Search Engines* (pp. 59–80). Elsevier. <https://doi.org/10.1016/B978-0-12-407171-1.00004-6>
- Macaluso, E., George, N., Dolan, R., Spence, C., & Driver, J. (2004). Spatial and temporal factors during processing of audiovisual speech: a PET study. *NeuroImage*, *21*(2), 725–732. <https://doi.org/10.1016/j.neuroimage.2003.09.049>
- Maguinness, C., Setti, A., Burke, K. E., Kenny, R. A., & Newell, F. N. (2011). The effect of combined sensory and semantic components on audio-visual speech perception in older adults. *Frontiers in aging neuroscience*, *3*, 19. <https://doi.org/10.3389/fnagi.2011.00019>
- Maguire, M. J., Abel, A. D., Schneider, J. M., Fitzhugh, A., McCord, J., & Jeevakumar, V. (2015). Electroencephalography theta differences between object nouns and action verbs when identifying semantic relations. *Language, Cognition and Neuroscience*, *30*(6), 673–683. <https://doi.org/10.1080/23273798.2014.1000344>
- Mastroberardino, S., Santangelo, V., & Macaluso, E. (2015). Crossmodal semantic congruence can affect visuo-spatial processing and activity of the fronto-parietal attention networks. *Frontiers in Integrative Neuroscience*, *9*, Article 45. <https://doi.org/10.3389/fnint.2015.00045>
- Mihalik, A., & Noppeney, U. (2020). Causal Inference in Audiovisual Perception. *The Journal of Neuroscience*, *40*(34), 6600–6612. <https://doi.org/10.1523/jneurosci.0051-20.2020>
- Molholm, S., Ritter, W., Javitt, D. C., & Foxe, J. J. (2004). Multisensory visual-auditory object recognition in humans: a high-density electrical mapping study. *Cerebral cortex (New York, N.Y. : 1991)*, *14*(4), 452–465. <https://doi.org/10.1093/cercor/bhh007>
- Morein-Zamir, S., Soto-Faraco, S., & Kingstone, A. (2003). Auditory capture of vision: examining temporal ventriloquism. *Cognitive Brain Research*, *17*(1), 154–163. [https://doi.org/10.1016/S0926-6410\(03\)00089-2](https://doi.org/10.1016/S0926-6410(03)00089-2)

Moscattelli, A., Hayward, V., Wexler, M., & Ernst, M. O. (2015). Illusory Tactile Motion Perception: An Analog of the Visual Filehne Illusion. *Scientific Reports*, 5(1). <https://doi.org/10.1038/srep14584>

Nagy, K., Greenlee, M. W., & Kovács, G. (2012). The lateral occipital cortex in the face perception network: an effective connectivity study. *Frontiers in psychology*, 3, 141. <https://doi.org/10.3389/fpsyg.2012.00141>

Narumi, T., Nishizaka, S., Kajinami, T., Tanikawa, T., Hirose, M. (2011). Meta Cookie+: An Illusion-Based Gustatory Display. In: Shumaker, R. (eds) *Virtual and Mixed Reality - New Trends*. VMR 2011. Lecture Notes in Computer Science, vol 6773. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-22021-0_29

Noppeney, U., Jones, S., Rohe, T. & Ferrari, A. (2018). See what you hear – How the brain forms representations across the senses. *Neuroforum*, 24(4), A169-A181. <https://doi.org/10.1515/nf-2017-A066>

Olson, I. R., Gatenby, J. C., & Gore, J. C. (2002). A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. *Cognitive Brain Research*, 14(1), 129–138. [https://doi.org/10.1016/S0926-6410\(02\)00067-8](https://doi.org/10.1016/S0926-6410(02)00067-8)

Opoku-Baah, C., Schoenhaut, A. M., Vassall, S. G., Tovar, D. A., Ramachandran, R., & Wallace, M. T. (2021). Visual Influences on Auditory Behavioral, Neural, and Perceptual Processes: A Review. *Journal of the Association for Research in Otolaryngology*, 22(4), 365–386. <https://doi.org/10.1007/s10162-021-00789-0>

Ortega, L., Guzman-Martinez, E., Grabowecky, M., & Suzuki, S. (2014). Audition dominates vision in duration perception irrespective of salience, attention, and temporal discriminability. *Attention, perception & psychophysics*, 76(5), 1485–1502. <https://doi.org/10.3758/s13414-014-0663-x>

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. (2019). PsychoPy2: Experiments in behavior made easy. *Behav Res* 51, 195–203. <https://doi.org/10.3758/s13428-018-01193-y>

Pelekanos, V., & Moutoussis, K. (2011). The effect of language on visual contrast sensitivity. *Perception*, 40(12), 1402–1412. <https://doi.org/10.1068/p7010>

Privitera, A. J. (2023). Sensation and perception. In R. Biswas-Diener & E. Diener (Eds), *Noba textbook series: Psychology*. Champaign, IL: DEF publishers. Retrieved from <http://noba.to/xgk3ajhy>

Rekow, D., Baudouin, J.-Y., Durand, K., & Leleu, A. (2022). Smell what you hardly see: Odors assist visual categorization in the human brain. *NeuroImage*, 255, 119181. <https://doi.org/10.1016/j.neuroimage.2022.119181>

Ro, T., Wallace, R., Hagedorn, J., Farné, A., Pienkos, E. (2004). Visual Enhancing of Tactile Perception in the Posterior Parietal Cortex. *J Cogn Neurosci* 2004; 16(1): 24–30. <https://doi.org/10.1162/089892904322755520>

- Rohe, T., Ehlis, A. C., & Noppeney, U. (2019). The neural dynamics of hierarchical Bayesian causal inference in multisensory perception. *Nature communications*, *10*(1), 1907. <https://doi.org/10.1038/s41467-019-09664-2>
- Roach, N. W., Heron, J., & McGraw, P. V. (2006). Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration. *Proceedings. Biological sciences*, *273*(1598), 2159–2168. <https://doi.org/10.1098/rspb.2006.3578>
- Rosen, S., & Hui, S. N. C. (2015). Sine-wave and noise-vocoded sine-wave speech in a tone language: Acoustic details matter. *The Journal of the Acoustical Society of America*, *138*(6), 3698–3702. <https://doi.org/10.1121/1.4937605>
- Ross, L. A., Molholm, S., Butler, J. S., Bene, V. A. D., & Foxe, J. J. (2022). Neural correlates of multisensory enhancement in audiovisual narrative speech perception: A fMRI investigation. *NeuroImage*, *263*, 119598. Advance online publication. <https://doi.org/10.1016/j.neuroimage.2022.119598>
- Sakai, N. (2005). The Effect of Visual Images on Perception of Odors. *Chemical Senses*, *30*(Supplement 1), i244–i245. <https://doi.org/10.1093/chemse/bjh205>
- Schifferstein, H. N. J., & Wastiels, L. (2014). Sensing Materials: Exploring the Building Blocks for Experiential Design. In E. Karana, O. Pedgley, & V. Rognoli (Eds.), *Materials Experience: Fundamentals of Materials and Design* (pp. 15–26). Butterworth-Heinemann. <https://doi.org/10.1016/B978-0-08-099359-1.00002-3>
- Shams, L., Kamitani, Y. & Shimojo, S. (2000). What you see is what you hear. *Nature* **408**, 788. <https://doi.org/10.1038/35048669>
- Smayda, K. E., Van Engen, K. J., Maddox, W. T., & Chandrasekaran, B. (2016). Audio-Visual and Meaningful Semantic Context Enhancements in Older and Younger Adults. *PLoS one*, *11*(3), e0152773. <https://doi.org/10.1371/journal.pone.0152773>
- Soto-Faraco, S., Lyons, J., Gazzaniga, M., Spence, C., & Kingstone, A. (2002). The ventriloquist in motion: Illusory capture of dynamic information across sensory modalities. *Cognitive Brain Research*, *14*(1), 139–146. [https://doi.org/10.1016/s0926-6410\(02\)00068-x](https://doi.org/10.1016/s0926-6410(02)00068-x)
- Soto-Faraco, S., Spence, C., & Kingstone, A. (2004). Cross-Modal Dynamic Capture: Congruency Effects in the Perception of Motion Across Sensory Modalities. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(2), 330–345. <https://doi.org/10.1037/0096-1523.30.2.330>
- Stein, B. E., London, N., Wilkinson, L. K., & Price, D. D. (1996). Enhancement of perceived visual intensity by auditory stimuli: a psychophysical analysis. *Journal of cognitive neuroscience*, *8*(6), 497–506. <https://doi.org/10.1162/jocn.1996.8.6.497>
- Teder-Sälejärvi, W. A., McDonald, J. J., Di Russo, F., & Hillyard, S. A. (2002). An analysis of audio-visual crossmodal integration by means of event-related potential (ERP)

recordings. *Cognitive Brain Research*, 14(1), 106–114. [https://doi.org/10.1016/s0926-6410\(02\)00065-4](https://doi.org/10.1016/s0926-6410(02)00065-4)

Thomas, J. P., & Shiffrar, M. (2013). Meaningful sounds enhance visual sensitivity to human gait regardless of synchrony. *Journal of Vision*, 13(14), Article 8. <https://doi.org/10.1167/13.14.8>

Tong, J., Li, L., Bruns, P., Röder, B. (2020). Crossmodal associations modulate multisensory spatial integration. *Atten Percept Psychophys* 82, 3490–3506. <https://doi.org/10.3758/s13414-020-02083-2>

Tsilionis, E., & Vatakis, A. (2016). Multisensory binding: is the contribution of synchrony and semantic congruency obligatory? *Current Opinion in Behavioral Sciences*, 8, 7–13. <https://doi.org/10.1016/j.cobeha.2016.01.002>

Tsuchiya, N. (2008). Flash suppression. *Scholarpedia*, 3(2):5640., revision #87576 doi:10.4249/scholarpedia.5640

Ujiie, Y., Yamashita, W., Fujisaki, W., Kanazawa, S., & Yamaguchi, M. K. (2018). Crossmodal association of auditory and visual material properties in infants. *Scientific reports*, 8(1), 9301. <https://doi.org/10.1038/s41598-018-27153-2>

Uno, K., & Yokosawa, K. (2022). Cross-modal correspondence between auditory pitch and visual elevation modulates audiovisual temporal recalibration. *Scientific reports*, 12(1), 21308. <https://doi.org/10.1038/s41598-022-25614-3>

van der Groen, O., van der Burg, E., Lunghi, C., & Alais, D. (2013). Touch Influences Visual Perception with a Tight Orientation-Tuning. *PLoS ONE*, 8(11), e79558. <https://doi.org/10.1371/journal.pone.0079558>

Vroomen, J., & de Gelder, B. (2000). Sound enhances visual perception: cross-modal effects of auditory organization on vision. *Journal of experimental psychology. Human perception and performance*, 26(5), 1583–1590. <https://doi.org/10.1037//0096-1523.26.5.1583>

Viggiano, M. P., Giovannelli, F., Giganti, F., Rossi, A., Metitieri, T., Rebai, M., Guerrini, R., & Cincotta, M. (2017). Age-related differences in audiovisual interactions of semantically different stimuli. *Developmental psychology*, 53(1), 138–148. <https://doi.org/10.1037/dev0000256>

Violentyev, A., Shimojo, S., & Shams, L. (2005). Touch-induced visual illusion. *NeuroReport: For Rapid Communication of Neuroscience Research*, 16(10), 1107–1110. <https://doi.org/10.1097/00001756-200507130-00015>

Wada, Y., Kitagawa, N., & Noguchi, K. (2003). Audio-visual integration in temporal perception. *International Journal of Psychophysiology*, 50(1-2), 117–124. [https://doi.org/10.1016/S0167-8760\(03\)00128-4](https://doi.org/10.1016/S0167-8760(03)00128-4)

Williams, J. R., Markov, Y. A., Tiurina, N. A., & Störmer, V. S. (2022). What You See Is What You Hear: Sounds Alter the Contents of Visual Perception. *Psychological Science*, 33(12), 2109–2122. <https://doi.org/10.1177/09567976221121348>

Woods, A. T., Poliakoff, E., Lloyd, D. M., Kuenzel, J., Hodson, R., Gonda, H., Batchelor, J., Dijksterhuis, G. B., & Thomas, A. (2011). Effect of background noise on food perception. *Food Quality and Preference*, 22(1), 42–47. <https://doi.org/10.1016/j.foodqual.2010.07.003>

Xi, Y., Li, Q., Gao, N., Li, G., Lin, W., & Wu, J. (2020). Co-stimulation-removed audiovisual semantic integration and modulation of attention: An event-related potential study. *International Journal of Psychophysiology*, 151, 7–17. <https://doi.org/10.1016/j.ijpsycho.2020.02.009>

Xie, Y., Xu, Y., Bian, C., & Li, M. (2017). Semantic congruent audiovisual integration during the encoding stage of working memory: an ERP and sLORETA study. *Sci Rep* 7, 5112. <https://doi.org/10.1038/s41598-017-05471-1>

Zampini, M., Guest, S., Shore, D. I., & Spence, C. (2005). Audio-visual simultaneity judgments. *Perception & Psychophysics*, 67(3), 531–544. <https://doi.org/10.3758/BF03193329>