NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

DEPARTMENT OF PHILOSOPHY AND HISTORY OF SCIENCE

DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

DEPARTMENT OF PSYCHOLOGY

DEPARTMENT OF PHILOLOGY

Using crossmodal correspondences to study the unity effect

By

Andromachi Melissari

Student Registration Number: 11M07

Thesis submitted

In partial fulfillment of the requirements for the Masters degree by the

Interdepartmental Graduate Program in Basic and Applied Cognitive Science

Advisory committee:

Athanasios Protopapas

Argiro Vatakis

Konstantinos Moutousis

Athens, Greece

September 2016

Advisory committee:

Athanasios Protopapas

Associate Professor

Department of Philosophy and History of Science

Argiro Vatakis

Researcher

Cognitive Systems Research Institute

Konstantinos Moutousis

Associate Professor

Department of Philosophy and History of Science

Acknowledgements

I would like to express my sincere thanks to Prof. Athanasios Protopapas, Prof. Konstantinos Moutousis, and Dr. Argiro Vatakis for their valuable comments on my thesis and the excellent collaboration we had during my Master's studies. It was a great honor for me to participate in this Master's degree program. My special thanks to Argiro Vatakis as this thesis would have not been realized without her constant support, guidance, and encouragement. She was always eager to discuss my difficulties and help me solve them. She is an exceptional person and a brilliant scientist and working with her was an unforgettable experience that I will always retain with the warmest feelings of gratitude and respect. Working with her was a great opportunity for me to extend my knowledge's and my mind. I would also like to thank my dear friend Maria Kostaki for the time she dedicated in order to prepare the code of the experiment and moreover for the time she has spent trying to encourage me. Last but not least, I would like to thank all the participants who took part in the experiments of this thesis.

Contents

Abstract

In this study, we investigate how crossmodal correspondences affect the “unity assumption” according to which an observer assumes that two different sensory signals refer to the same underlying multisensory event. Participants were exposed to audiovisual pairs of stimuli that they were presented with a range of different stimulus onset asynchronies (SOAs) using the method of constant stimuli. The presented audiovisual stimuli consisted of pitch-color, pitch-shape, and pitch-size pairs and participants were asked to make unspeeded temporal order judgments (TOJ) regarding witch modality, either the visual or the auditory, that had been presented first. According to the “unity assumption”, participants expected to have difficulty in judging the order of appearance of the stimuli, assuming they referring to the same underlying multisensory event (i.e., matched condition), while on the contrary, participants expected to finding it more easy to judge the order of appearance of the stimuli, assuming that they are not referring to the same underlying multisensory event (i.e., mismatched condition). Though, results in the audiovisual stimuli pair pith-size, and pitch-shape did not verify our hypothesis, as participants did not show any different performance between the matched and the mismatched condition. Therefore, it would seem fruitful to assume that our findings regarding the influence of the these specific crossmodal correspondences pairs we utilized to the “unity assumption” reflect some combination of both top-down and bottom-up factors influencing multisensory integration.

Περίληψη

Στη συγκεκριμένη έρευνα εξετάσαμε το πως οι διατροπικές αντιστοιχίες μπορούν να επιδράσουν στην απόφαση ενοποίησης, σύμφωνα με την οποία ένας παρατηρητής υποθέτει ότι δυο διαφορετικά αισθητηριακά σήματα αναφέρονται στο ίδιο πολυαισθητηριακό γεγονός. Οι συμμετέχοντες εκτέθηκαν σε ζεύγη οπτικοακουστικών ερεθισμάτων τα οποία παρουσιάστηκαν σε εύρος διαφορετικών χρονικών διαστημάτων, χρησιμοποιώντας τη μέθοδο των σταθερών ερεθισμάτων. Τα συγκεκριμένα οπτικοακουσικά ερεθίσματα αποτελούνται από τόνο-χρώμα, τόνο-σχήμα και τόνο-μέγεθος. Ζητήθηκε από τους συμμετέχοντες να αποφασίσουν τη σειρά έλευσης του ερεθίσματος που παρουσιάστηκε πρώτο (ακουστικό ή οπτικό). Βάσει της θεωρίας ενοποίησης, οι συμμετέχοντες αναμένεται να δυσκολευτούν στη διάκριση έλευσης του ερεθίσματος, θεωρώντας ότι ανήκει στο ίδιο πολυαισθητηριακό γεγονός (συνθήκη ταιριάσματος ερεθισμάτων) ενώ, αντίθετα αναμένεται να διακρίνουν πιο εύκολα το ερέθισμα που παρουσιάστηκε πρώτο, θεωρώντας ότι τα παρουσιαζόμενα ερεθίσματα δεν ανήκουν στο ίδιο πολυαισθητηριακό γεγονός (συνθήκη μη ταιριάσματος ερεθισμάτων). Τα αποτελέσματα για τα ερεθίσματα τόνος-μέγεθος και τόνος-σχήμα, δεν επιβεβαίωσαν την υπόθεσή μας, καθώς οι συμμετέχοντες δεν έδειξαν καμία διαφορά στην επίδοσή τους, είτε στη συνθήκη ταιριάσματος, είτε στη συνθήκη μη-ταιριάσματος. Το γεγονός αυτό ενδεχομένως αντικατοπτρίζει έναν συνδυασμό ανωφερούς (βάσει χαρακτηριστικών των ερεθισμάτων) κατωφερούς (βάσει γνωστικών παραγόντων) τρόπου επεξεργασίας της πληροφορίας.

Using crossmodal correspondences to study the unity effect

2016

1. Introduction

Definition of Crossmodal correspondences

Crossmodal correspondences (also referred to as: synesthetic correspondences, Parise & Spence, 2009; synesthetic associations, Parise & Spence, 2008; crossmodal mappings, Evans & Treisman, 2010) refer to the ability that humans (presumably and other species too; e.g., chimps, Ludwig, Adachid, & Matsuzawad, 2011) have to associate or integrate information about specific properties of stimulus features (e.g., pitch, color) that originate from different sensory modalities. In particular, crossmodal correspondences refer to the association between different and seemingly unassociated features of stimuli (Parise, 2015; Spence, 2011). For example, high pitch sounds are often associated with small in size visual objects (e.g., Evans & Treisman, 2010; Gallace & Spence, 2006; Keetels & Vroomen, 2011; Mondloch & Maurer, 2004; Parise & Spence, 2008, 2009). Crossmodal correspondences can occur at both levels either derived from the low-level amodal redundant stimuli characteristics (defused between modalities; e.g., duration, spatial location) as well as from high-level modal unrelated stimuli characteristics (Parise, 2015; Spence, 2011). It has to be noted, that two or more characteristics are redundant when they provide information about the same physical property (i.e., visual and haptic size characteristics provide information about the same physical property, which is the size of an object), while on the contrary, unrelated are those who are independent with each other (i.e., visual lightness is not associated to haptic conformity; Parise, 2015).

Crossmodal correspondences learned or low-level based

According to cross-cultural studies, the existence of crossmodal correspondences is believed to be universal. For example, it was found that indigenous people in Namibia lacking written language and having restricted contact with western culture, experienced the

"bouba-kiki effect" (whereby people associate the meaningless word "bouba" with a rounded shape object and the meaningless word "kiki" with an angular shape object; Bremner, Caparos, Davidoff, Fockert, Linnell, & Spence, 2013; Maurer, Pathman, & Mondloch, 2006; Parise & Spence, 2012). Moreover, according to developmental studies crossmodal correspondences appear to be present from early childhood. For example, the "bouba-kiki effect" was demonstrated in 2.5 year old children (Maurer, Pathman, & Mondloch, 2006). Likewise, infants between 4 and 6 months demonstrated crossmodal association for pitch and size (Fernandez-Prieto, Navarra, & Pons, 2015). Thus, according to the aforementioned cross cultural and developmental studies, crossmodal correspondences appear to be innate as they have been demonstrated both in foreign cultures and infants.

Furthermore, crossmodal correspondences have been studied through a variety of different methods and experimental designs such as, speeded classification method, where participants have to discriminate (i.e., task - relevant) a particular stimulus characteristic in one modality (e.g., size of visual object; Evans & Treisman, 2010; Gallace & Spence, 2006) as fast as possible (i.e., Reaction Times, RTs) while at the same time are asked to ignore (i.e., task - irrelevant) any other distractor stimulus characteristic presented in other modality. It has been documented that participants' responses are undoubtedly influenced by the ignored condition differentiating their RT from faster to lower, based on the compatibility or incompatibility of the presented stimuli (Evans & Treisman, 2010). Particularly, participants were exposed in direct (i.e., in the auditory task, they had to discriminate whether the presented tone was high or low, while in the visual task, they had to discriminate whether the grating object was small or large) and indirect (i.e., in the auditory task, participants were asked whether the tone was produced either by violin, or piano, while on the visual task, they were asked whether the grating was right or left oriented) condition. The finding that participants systematically associated small size with high pitch as big size with low pitch in

both conditions (i.e., direct and indirect) led the researchers to the conclusion that the specific association of crossmodal correspondence reflects an intrinsic association at a perceptual level (Evans &Treisman, 2010).

Moreover, the indirect implicit method (Implicit Association Test, IAT) is used in order to reveal peoples' unconscious associations among different stimuli attributes, where in the simple version participants are asked to respond as fast they can to a set of presented stimuli (e.g., two auditory and two visual stimuli) that are being assigned in two respond keys. It has been documented that participants RT is faster when the stimuli to a respond key are tightly connected with each other (the congruence condition), than being unconnected (the incongruence condition; Parise & Spence, 2012). It was found that participants responded extremely fast when a big visual cycle and a low pitch tone were assigned to the same respond key as well as a small cycle and a high pitch tone respectively (congruence condition), while on the contrary their reaction time was slower when a big cycle and high pitch tone were assigned to the same respond key as well as a small cycle and low tone, respectively (incongruence condition; Parise & Spence, 2012). This outcome led researchers to the conclusion that conversely to previous findings especially those reported by Evans and Treisman (2010), the crossmodal correspondences do not seem to have any effect at a perceptual level, but it is rather possible to reflect an effect of response selection, due to the experimental design that allowed only one stimulus at the time to be presented to the participants. As a result researchers argued that crossmodal correspondences function on both levels, perceptual and response selection, revealing the magnitude of the effect that crossmodal correspondences have on information processing (Parise & Spence, 2012). Thus, taken together all the above, researchers utilizing different experimental designs argued, due to the impact that crossmodal correspondences had on participants' accuracy and rapidity responses during the given tasks, that crossmodal correspondence are likely to operate in an

automatic manner, suggesting the existence of underlying mechanism supporting the crossmodal congruency (Evans & Treisman; Parise & Spence, 2012).

The experimental design Temporal Order Judgment (TOJ-task) has been used by researchers in order to investigate crossmodal correspondences, where participants are requested to judge the temporal order of presented stimuli (i.e., which stimulus has been presented first, either the visual, or the auditory one; Parise & Spence, 2009). For example, it was found that participants had difficulty in judging the temporal order of presented stimuli when they were presented in matched (i.e., big or small visual circles combined with low or high tone, respectively) condition, rather than mismatched (i.e., big or small visual circles combined with high or low tone, respectively) condition (Parise & Spence, 2009). The particular audiovisual pair is argued by the researchers to constitute a form of synesthetic association between stimuli presented by different modalities, thus synesthetic correspondence is depicted on participants' reliability on audiovisual TOJs and spatial localization judgments (Parise & Spence, 2009).

All these different methods adopted by the researchers have been used in order to answer questions regarding the nature of the crossmodal correspondences (i.e., whether are learned or are innate; Ludwig, Adachid, & Matsuzawad, 2011) and the impact they have at the perceptual system. One of the oldest questions in the field of crossmodal correspondences concerning their origin is whether they reflect statistical properties of the environment in which humans live and evolved (Parise, 2015). For example, the connection between auditory pitch and visual size is possible to rely on the properties of acoustic resonance (i.e., pitch and size are inversely related by the physical law of acoustic resonance; Parise, 2015) and if that is the case it can be explained how these properties (pitch-size) are constituted a special dyad that is interpreted by an observer depending on the circumstances as one multisensory event despite, that auditory pitch either high or low does not provide solid information about an

objects' size (i.e., Evans & Treisman, 2010; Parise, 2015; Parise & Spence, 2013; Spence, 2011).

In conclusion, crossmodal correspondences have been demonstrated between many different pairs of stimuli (e.g., audiovisual stimuli pair, which by the way has concentrated the greatest interest of the research) and their dimensions (Spence, 2011; Spence & Deroy, 2013). Auditory pitch has been documented to be congruent with different visual stimuli aspects such as size (i.e., experiments 3-4 in Evans & Treisman, 2010; Parise & Spence, 2009), shape (e.g., Maurer, Pathman, & Mondloch, 2006), and color (e.g., Hubbard, 1996; Klapetek, Ngo, & Spence, 2012; Marks, 1987; Martino & Marks, 1999; Melara, 1989; Mondloch & Maurer, 2004). Moreover, crossmodal correspondences have already been documented by cross cultural (e.g., Bremner, et al., 2013) and developmental (e.g., Fernandez-Prieto, Navarra, & Pons, 2015; Maurer, Pathman, & Mondloch, 2006) studies. Furthermore, they have been demonstrated by a variety of methodologies such as the speeded classification method (i.e., size of visual object combined with pitch; Evans & Treisman, 2010; Gallace & Spence, 2006), indirect implicit method (i.e., size of visual object combined with pitch; Parise & Spence, 2012), TOJ-task (i.e., size of visual object combined with pitch; Parise & Spence, 2009). Crossmodal correspondences have been used from researchers in order to shed light to questions, regarding their origin (i.e., whether are learned or are innate; Ludwig, Adachid, & Matsuzawad, 2011) and the influence they have on the perceptual system (i.e., Evans & Treisman, 2010; Parise, 2015; Parise & Spence, 2013; Spence, 2011).

Taken together all the above, crossmodal correspondences are argued to operate in automatic manner, suggesting the existence of underlying mechanism supporting the crossmodal congruency. In addition, crossmodal congruency has also been demonstrated in more complex audiovisual stimuli that were used in studying the "unity assumption".

Unity assumption

When people are presented with different stimuli originating from different modalities they may perceive them either as two separate sensory events, or, on the contrary, as one unitary multisensory event. Unity assumption is referring to the decision an observer makes regarding whether or not the presented stimuli constitute one unitary multisensory event (Vatakis & Spence, 2007). The decision is made based on the consistency of the available information of each sensory modality and on the perceptual grouping (Vatakis, Ghazanfar, & Spence, 2008; Vatakis & Spence, 2007, 2008).

At first, research on unity effect, had focused on the role of spatiotemporal variables, designating the spatial ventriloquism effect, where the perceived location of the audio stimulus is captured by the visual stimulus, due to their temporal coincidence (Vroomen & Keetels, 2006). Later, research focused on the temporal ventriloquism analogous to spatial ventriloquism, where the auditory stimulus influences the visual one, within a temporal window, by altering the perception of the temporal occurrence of the visual stimulus (Morein-Zamir, Soto-Faraco, & Kingstone, 2003; Vroomen & Keetels, 2006).

Moreover, early studies utilized more complex stimuli, common in everyday life (i.e., steam whistle and steaming kettle; Jackson, 1953), unfortunately, lack of validity, because it is ambiguous whether participants have really experienced any multisensory perceptual event, or on the contrary they have led to bias responses (Vatakis & Spence, 2007). In other words it is possible that participants assumed that the presented stimuli were constituted a unified event, rather than having experienced it. The above constrain was excluded later in a study where participants were presented with audiovisual speech stimuli, together with a TOJ-task (Vatakis & Spence, 2007).

Specifically, researchers hypothesized that according to the "unity assumption", participants' should have difficulty in discriminating the temporal order of the presented audiovisual speech stimuli, either the visual or the auditory, in a TOJ - task, when these audiovisual speech stimuli, have previously been perceived as one unitary multisensory event. The first experiment consisted of visual stimuli (video clip) of male and female figure uttered speech auditory syllables, either matched or mismatched, where participants had to decide the temporal order of their appearance. It was expected that in matched condition, participants should have difficulty in discriminating the temporal order due to the perceptual association of the audiovisual stimuli as referring to the same underlying multisensory event conversely on the mismatched condition they should have better performance. According to researchers, it was the first empirical demonstration of how unity assumption can facilitate the crossmodal correspondence at a perceptual level, while on the same time the experimental design (gender matched – mismatched stimuli) they used along with a TOJ – Task, warranted the outcome, by ruling out at a decisional level any potential bias responses, encountered in other studies where the use of simultaneously judgement task, might have mislead participants in bias responses (Vatakis & Spence, 2007). Results of the second experiment (gender images uttered the words happy and odor), of the third experiment (dubbed matched and dubbed mismatched syllables) and of the forth experiment (syllables uttered by female, in order to eliminate any possibility of gender bias responses in the mismatched condition) verified the finding of the first experiment, and researchers found strong evidence to support the idea that unity assumption can facilitate crossmodal correspondence by using auditory and visual speech stimuli. Though, it has to be mentioned that by many researchers (e.g., Baart, Stekelenburg, & Vroomen, 2014; Jones & Jarick, 2006; Tuomainen, Andersen, Tiippana, & Sams, 2005) speech represents a very special kind of stimulus, that is well established among humans.

In a follow up study, Vatakis and Spence (2008) utilized non-speech complex stimuli representatives of daily life, such as playing music and object manipulation, with purpose to see whether the results of their previous study, could be replicated. This would imply that "unity assumption" is referring not only to speech stimuli, but also to non-speech stimuli, originated from the environment we interact with. The first experiment consisted of two video clips presenting a man smashing a block of ice with a hammer and a man bouncing a ball. These video clips where either matched or mismatched with the auditory sounds and participants had to indicate which of the stimuli (visual – auditory) had been presented first. Researchers hypothesized that according to the "unity assumption", participants would have difficulty in discriminate the temporal order in the matched condition, while on the mismatched condition participants expected to have better performance. Unfortunately, the analysis showed no difference in participants' performance in both matched – mismatched conditions, thus, results did not verify the previous studies, failed to confirm the claim that the "unity assumption" can modulate the crossmodal not speech binding. The second experiment consisted of musical stimuli, (piano and guitar video clip, piano and guitar "a" & "r" notes). The above audiovisual stimuli presented either matched or mismatched and participants had to discriminate the temporal order of their appearance. It has to be noted here, that music is thought to be more representative and closely related to speech comparatively to objects manipulation involving time perception (Vatakis & Spence, 2008). However, the JND analysis showed no difference in participants' performance, thus, failing to confirm the claim that the "unity assumption" can modulate the crossmodal binding. The third experiment consisted only from piano audiovisual stimuli, in order to eliminate any possible answer attributed to differences easily distinguishable by participants, such as the shape and the appearance of the instruments. Moreover, participants were experts, taking lessons for many years, thus, overlearned users. Though, again the JND analysis showed no

difference in their performance. Consequently, in this particular study researchers by utilizing non-speech stimuli did not manage to provide any evidence of the "unity assumption", similar to the previous study (i.e., utilized speech stimuli; Vatakis & Spence, 2007), indicating that only speech stimuli can modulate the crossmodal binding, while the non-speech stimuli cannot. This finding led researchers to the conclusion that speech is "special" due the fact that participants had difficulty in discriminating the temporal order of the presented stimuli, either visual, or auditory in both matched and mismatched conditions. They argued that this finding presumably depicts our daily experience with speech phenomena from early childhood, made us experts in this area. It also ca explain the fact that humans have accomplished the ability to detect and furthermore to distinguish every single difference draw our attention in speech phenomena, in facial expressions, articulation gestures and sounds (phonemes) because speech constitutes a basic element on interaction and communication with other people, thus, is "special" (Vatakis, Ghazanfar, & Spence, 2008; Vatakis & Spence, 2008).

In a sequence of four experiments researchers investigated the possibility that unity effect can modulate audiovisual non-speech vocalization (Vatakis, Ghazanfar, & Spence, 2008). The first experiment consisted of video clip showing two rhesus monkeys uttering vocalizations ('coo' or 'grunt'), presented either matched or mismatched, where participants had to indicate the temporal order of the stimuli. The results failed to reveal any support of the "unity effect" in non-speech vocalization. The second experiment consisted of video clip showing two rhesus monkeys uttering different vocalizations ('coo' or 'threat'), presented either matched or mismatched, where participants had to indicate the temporal order of the stimuli. The results failed to reveal any support of the "unity effect" in non-speech vocalization. In order to see if there would be any difference between participants' performance referring to speech and vocalization non-speech stimuli, researchers conducted

the third experiment where the stimuli composed of video clips of either a man or a rhesus monkey uttering the same vocalization ('coo') presented either matched or mismatched where the participants had to indicate the temporal order of the audiovisual stimuli. Findings of the third experiment failed to reveal any support of the "unity effect" in non-speech vocalization, even though a male uttered the vocalization, which may reflect the fact that only speech can modulate the multisensory integration as speech has the power to accomplish the unity assumption (vatakis & Spence, 2007; 2008). In the fourth experiment, (either a male or female uttered the sound 'a') the findings were similar to the third experiment of Vatakis and Spence (2007), showing that the "unity effect" can modulate the multisensory integration of audiovisual speech stimuli, due to the putatively "special" nature of the speech (Vatakis, Ghazanfar, & Spence, 2008).

Summarizing, the above findings concerning the influence of the "unity assumption" on audiovisual speech stimuli, have led researchers to the conclusion that the "unity assumption" can affect the audiovisual integration of speech stimuli and, thus, speech is "special". Consequently, according to these researchers top-down processes play an important role in facilitating integration, where semantic congruency regarding the signals' underlying event, facilitates integration. This factor contributes to the perceptual system's decision as to whether multimodal signals originate from common underlying causes or events —a process known as the unity assumption.

Contrary to these findings, more recently, it has been suggested that judging audiovisual temporal order in speech is not affected by whether the auditory and visual streams are paired. For example, in a study where participants were presented with audiovisual sine-wave speech stimuli, it was found that audiovisual temporal sensitivity was no different for participants regardless of whether they had perceived the stimuli as speech, or sine-waves (Vroomen & Stekelenburg, 2010). This result is quite compelling evidence against the "unity

assumption" since it is the first study on intersensory synchrony in which pairing between the auditory and visual streams was manipulated while all contributions from low-level stimulus differences were equated (Vroomen, & Stekelenburg, 2010). These authors show that intersensory pairing of a sound and lipread information is affected by whether the sound is heard as speech. Crucially, though, the pairing in the phonetic domain did not affect judgments of audiovisual temporal order. They argued that the pairing is mostly based on the low-level temporal correlation and coincidence between the two information streams. Only if there is prudent support for "same object/event", the content of the two information streams is perceived as "synchronous" and subsequently merged at the phonetic level. They also interpreted these compelling (between their and the previous studies of Vatakis, Ghazanfar, & Spence, 2008; Vatakis, & Spence, 2007, 2008) findings as a result of the existence of different stimulus factors to contribute to the perception of audiovisual synchrony in speech and non-speech, and that mismatching information affects them differently. Thus, they argued that in normally-matched audiovisual continuous speech, there is the continuous temporal correlation between the time-varying characteristics of the auditory and visual streams that may induce a "temporal ventriloquist" effect which may explain why sensitivity for audiovisual temporal order is better for incongruent than congruent audiovisual speech, due the likelihood that the fine temporal correlation in incongruent speech is disrupted so that small lags in continuous audiovisual speech to become unnoticeable (Vroomen, & Stekelenburg, 2010). On the other hand, for discrete events there is no such inherent time-varying correlation between the auditory and visual streams, thus, perceivers will have to rely primarily on the temporal coincidence of auditory and visual transient onsets. The monkey calls used in the study by Vatakis, Ghazanfar, and Spence (2008) contained short of transient onsets with almost no visual anticipatory information. Here also, temporal judgments may thus likely be based on the temporal coincidence of the onsets rather than the time-varying

co-modulation. They concluded that Judging temporal order in audiovisual speech may thus differ from non-speech not because speech is "special", but because speech has a fine temporal correlation between sound and vision that induces temporal ventriloquism, and judging temporal order in audiovisual speech may for that reason be difficult (Vroomen & Stkelenburg, 2010).

In conclusion, contrary to the argument that speech is "special" the statement that speech is "not special" suggests that differences between speech and non-speech in temporal ventriloquism as a function of audiovisual congruency have less to do with higher order percepts of 'unity', and more to do with low-level differences between these stimulus classes.

Besides, more recently Chuen and Schutz (2016) by utilizing musical stimuli, they found strong influence of the "unity assumption" on temporal cross-modal binding for non-speech musical stimuli. Together with the former findings (i.e., Vroomen & Stekelenberg, 2010), these findings indicate that speech is "not special" when it comes to audiovisual temporal sensitivity. Thus, contrary to former researchers, who argued that higher order interpretations of unity do not facilitate temporal ventriloquism, this particular work extended evidence of the influence of the "unity assumption" in audiovisual integration by demonstrating that complex modality-specific information (e.g., timbre) influences multisensory causal inference (Chuen & Schutz, 2016).

Parise and Spence (2009) in their study, investigated the putative influence of synesthetic (i.e., other term to describe the crossmodal correspondences) correspondences on multisensory integration, and they performed a sequence of three experiments with a TOJ-Task design, utilizing classic audiovisual pairs, including pitch-size (i.e., small circle-high pitch tone, big circle-low pitch tone), pitch-shape (i.e., rounded shape 7 pointed star-low pitch tone, angular shape 7 pointed star-high pitch tone), and pitch-Gaussian blob (small visual

object-high pitch tone, big visual object-low pitch tone; Parise & Spence, 2009). In all the three experiments it was found that participants had better performance on congruent rather than incongruent condition. Thus, results led researchers to the conclusion that these correspondences can really affect multisensory integration as demonstrated by the increased reliability of participants' audiovisual TOJs and spatial localization judgments (Parise & Spence, 2009).

In the present study, we utilized three classic and extensively used pairs of crossmodal correspondences consisted of color-pitch, shape-pitch, and size-pitch stimuli, in order to investigate the effect, if any, they would have to the "unity assumption". The aforementioned crossmodal audiovisual pairs were presented, in both matched (i.e., based on their features congruency) and mismatched (i.e., based on their features incongruence) condition. It has to be noted that the matching condition was based on the literature, according to which a white visual object is congruent with a high pitch tone, while a black visual object is congruent with a low pitch tone (e.g., Hubbard, 1996; Klapetek, Ngo, & Spence, 2012; Marks, 1987; Martino & Marks, 1999; Melara, 1989; Mondloch & Maurer, 2004). Moreover, a rounded shape object is congruent with a low pitch tone, while an angular shape object is congruent with a high pitch tone (e.g., Bremner, Caparos, Davidoff, Fockert, Linnell, & Spence, 2013; Maurer, Pathman, & Mondloch, 2006; Parise & Spence, 2012). Finally, a big size object is congruent with a low pitch tone, while a small size object is congruent with a high pitch tone (e.g., Evans & Treisman, 2010; Gallace & Spence, 2006; Keetels & Vroomen, 2011; Mondloch & Maurer, 2004; Parise & Spence, 2008, 2009). The mismatching condition also based on the literature, according to which a white visual object is congruent with a low pitch tone, while a black visual object is congruent with a high pitch tone (e.g., Hubbard, 1996; Klapetek, Ngo, & Spence, 2012; Marks, 1987; Martino & Marks, 1999; Melara, 1989; Mondloch & Maurer, 2004). Moreover, a rounded shape object is congruent with a high pitch tone, while an

angular shape object is congruent with a low pitch tone (e.g., Bremner, Caparos, Davidoff, Fockert, Linnell, & Spence, 2013; Maurer, Pathman, & Mondloch, 2006; Parise & Spence, 2012). Finally, a big size object is congruent with a low pitch tone, while a small size object is congruent with a high pitch tone (e.g., Evans & Treisman, 2010; Gallace & Spence, 2006; Keetels & Vroomen, 2011; Mondloch & Maurer, 2004; Parise & Spence, 2008, 2009).

The experimental design we utilized is the TOJ-task where participants were asked to discriminate the temporal order of the presented audiovisual stimuli. Specifically, participants were asked to determine which of the audiovisual stimuli either the visual, or the auditory, had been presented first. According to the "unity assumption" participants expected to have difficulty in judging the order of appearance of the stimuli, assuming that referring to the same underlying multisensory event (i.e., matched condition), while on the contrary, participants expected to finding it more easy to judge the order of appearance of the stimuli, assuming that are not referring to the same underlying multisensory event (i.e., mismatched condition). Such an outcome would provide an empirical demonstration that the "unity assumption" can facilitate the crossmodal binding of audiovisual stimuli. It has to be mentioned that in the present study, by using the term of crossmodal correspondences, we refer exclusively to correspondences between simple sensory characteristics, which they constitute the point of interest of this study. In particular, we wanted to investigate the putative influence of these crossmodal correspondences on the "unity assumption", which previously has been studied by informationally and semantically rich crossmodal correspondences.

In the present study, we used this audiovisual matched-mismatched design together with a TOJ-task, in order to investigate the effect they would have on the "unity assumption", by eliminating any potential response bias due to the experimental design used. The main advantage of the specific design is the fact that participants are asked to discriminate the

order of appearance of the presented audiovisual stimuli (i.e., “vision first” or “sound first”) regardless of matching. In other words, according to the specific experimental design, participants obtain no knowledge in advance whether the presented audiovisual pairs constitute matched or mismatched crossmodal correspondences, due the fact that we elicit their assumptions by judging the order of appearance of the presented stimuli. Thus, we hypothesized that when participants assume the presented audiovisual stimuli to constitute a solid and tightly bind pair, then they would have difficulty in judging the order of the appearance of these stimuli, as a result from the fact that they could not split their bonding. On the contrary, had we used a simultaneity judgment task, we might have had evoked biased responses based on the simultaneously appearance of the audiovisual stimuli. In that case, the presentation of either matched or mismatched audiovisual pairs should not differentially affect the likelihood of participants making a “vision” or “sound” first response.

2. Methods

2.1 Participants

Eighty-four (11 males) naïve volunteers, aged 18-19 years (mean age= 18.5 years), took part in the experiment for course credit. All participants reported normal and corrected-to-normal vision and normal auditory perception. Twenty-six participants were excluded from further analysis, due to inappropriate completion of the task (i.e., the PSS and the JND, were larger or smaller to the SOA ±332).

2.2 Stimuli and apparatus

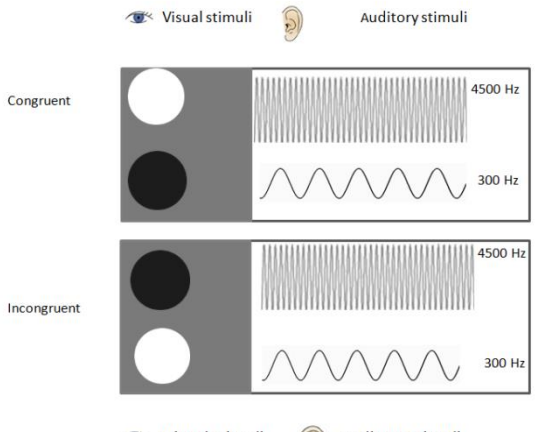
The experiment was composed of three blocks, included color-pitch, shape-pitch, size-pitch, of 363 trials. Nine possible SOAs between the auditory and visual stimuli were used. (SOAs: ±332, ±249, ±166, ±83, 0ms). Negative SOAs indicate that the auditory stream was presented first, while positive values indicate that the visual stream was presented first.

The bimodal audiovisual stimuli were presented equiprobably asynchronous in pairs either matched (i.e., white color-high pitch, black color-low pitch, rounded shape-low pitch, angular shape-high pitch, big size-low pitch, small size-high pitch) or mismatched (i.e. white color-low pitch, black color-high pitch, rounded shape-high pitch, angular shape-low pitch, small size- low pitch, big size-high pitch).

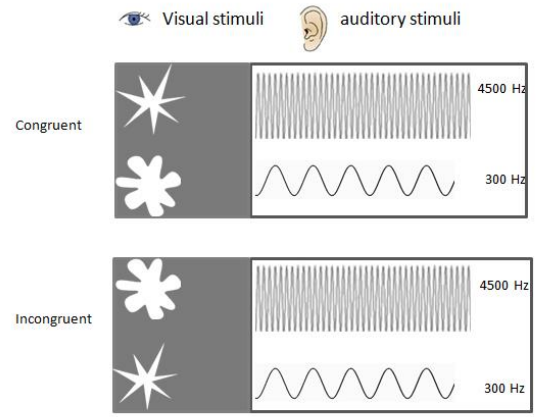
A total of 36 different conditions were presented with 10 repetitions for each. The experiment was divided in three blocks with breaks between them. The experiment was performed using Presentation (version 17.0, Neurobehavioral systems, Inc.). The visual stimuli were presented on a HP lap-top 17-in.; 60Hz refresh rate), while the auditory stimuli were presented by headphones (Sennheiser HD 380 pro). The order of stimuli presentation was randomized.

The stimuli used in the color – pitch condition, were pairs comprised of one auditory stimulus and one visual stimulus. The visual stimuli consisted of black and white cycles, presented on a gray background. The auditory stimuli consisted of high pitch (4500Hz) tone and low pitch (300Hz) tone, respectively (Figure 1A). The stimuli used in the shape – pitch condition, were pairs consisted of one auditory stimulus and one visual stimulus. The visual stimuli consisted of angular shape (i.e. kiki) and rounded shape (i.e. bouba). The auditory stimuli consisted of high pitch (4500Hz) tone and low pitch (300Hz) tone, respectively (Figure 1B). Finally, the stimuli used in the size – pitch condition, were pairs comprised of one auditory stimulus and one visual stimulus. The visual stimuli consisted of big and small cycles, while the auditory stimuli consisted of high pitch (3000Hz) tone and low pitch (1250Hz) tone, respectively (Figure 1C).

A



B



C

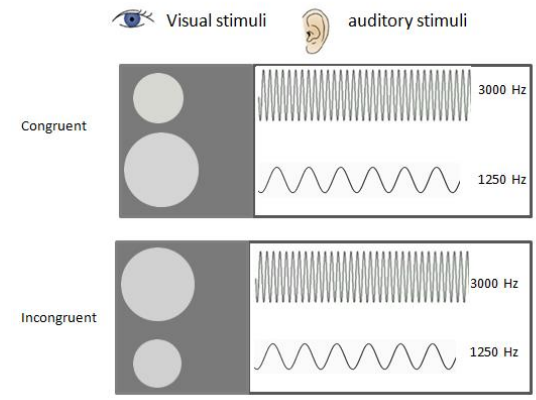


Figure 1. (A) White and black circles are congruent with 4500 Hz and 300 Hz respectively, while black and white circles are incongruent with 4500 Hz and 300 Hz respectively. (B) Angulated shape (kiki) and rounded shape (bouba) are congruent with 4500 Hz and 300 Hz respectively, while, angulated shape (kiki) and rounded shape (bouba) are incongruent with 300 Hz and 4500 Hz respectively. (C) Small and big circles are congruent with 3000 Hz and 1250 Hz respectively, while small and big circles are incongruent with 1250 Hz and 3000 Hz respectively.

2.3. Procedure

Participants were seated approximately 60 cm from the screen in a dimly-lit room. They were informed that they would be presented with a series of audiovisual stimuli pairs, where the auditory stimulus would come from the headphones they put on their ears, and the visual stimulus would be presented on the center of the laptop screen. They were asked to fixate on the fixation point (+) on the center of the screen and discriminate which of the stimulus comes first (the auditory or the visual). They made unspeeded temporal order judgements (TOJ) regarding which stimulus (visual or auditory) had been presented first. They were asked to reply by pressing on the keyboard either the button "A" for auditory stimulus, using the left hand, or the button "O" for the visual stimulus, using the right hand. After answering, they were presented with the next pair of audio-visual stimuli.

They were also informed that sometimes it would be easier to discriminate the temporal order of the presented stimulus (either visual or auditory) while other times it would be more difficult to discriminate the temporal order of the presented stimulus (either visual or auditory). In the latter case they were asked to answer by making an inform guess and they were also instructed to avoid the random answers and trying to be as accurate as possible.

Before the main experiment, a short practice test was given to the participants in order to familiarize them with the experimental procedure. The experimenter provided detailed verbal instructions and before the start of the experiment, written instructions were also provided on the lap-top screen.

3. Results

3.1 Analysis

The proportion from the "visual first" responses was calculated for all participants in each of the four audiovisual combinations (matched – mismatched) of the presented three crossmodal correspondences pairs. Z-scores were calculated from the proportion of the "visual first" responses assuming a cumulative normal distribution and then, a line of best fit was calculated from every participant derived from the nine SOAs (±332, ±249, ±166, ±83, 0ms) that he/she was exposed to during the experimental procedure (figure 2. A). The TOJ-Task provided us with two important measures which are the Just Noticeable Difference (JND) and the Point of Subjective Simultaneity (PSS), where the former constitutes a standardized measure concerning the participants' sensitivity referring to their Temporal Order Judgements of the presented stimuli and the latter, is referring to the point at which the two events are to be perceived as synchrony or occurring as equally often (Vatakis & Spence, 2007). Both JND (JND= 0.675/slope, given that ± 0.675 represents the 75% and 25% points on the cumulative normal distribution) and PSS (PSS= -intercept/slope) values were calculated from the slope and intercept values of each line. The data (JND & PSS) in all the audiovisual combinations (matched – mismatched) in the three crossmodal correspondences pairs were analyzed by repeated measures ANOVA and Bonferroni corrected t-tests (where $p < .05$ prior to correction) were used for all post hoc comparisons. In addition, Cluster analysis also executed for further investigation regarding the grouping of the participants in all three conditions (either in matched or mismatched group) based on their responses.

The average JNDs for the matched and the mismatched condition for the three crossmodal correspondences audiovisual pairs (i.e., color-pitch, shape-pitch and size-pitch)

are presented in Figure 3. Significant difference between groups ($p<0.05$) is highlighted by asterisk.

A

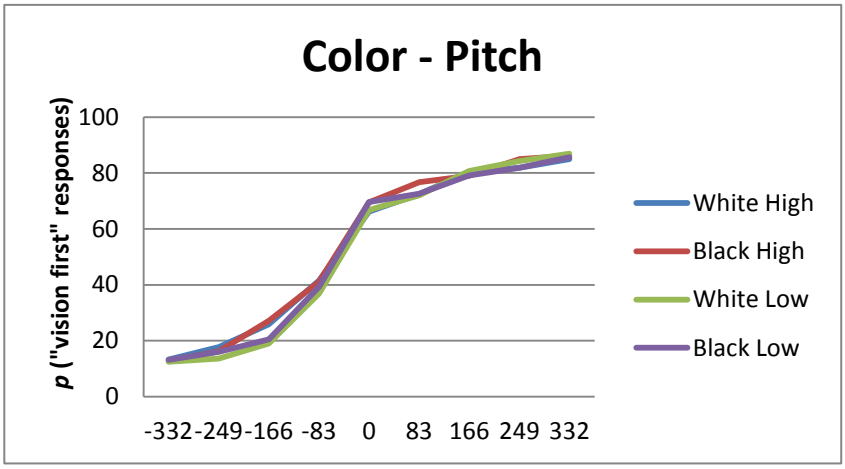


Figure 2. (A) Mean percentage of "vision first" responses plotted as a function of stimulus onset asynchrony (SOA) for each of the audiovisual condition. a) color-pitch: congruent (white-high (bleu)/black-low (purple), incongruent (white-low (green)/black-high (red).

B

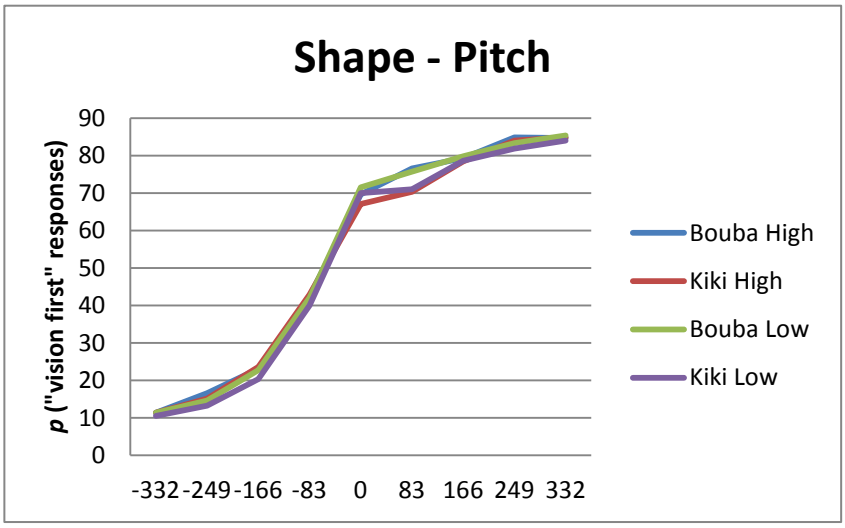


Figure 2. (B) shape-pitch: congruent (Kiki-high (red)/bouba-low (green), incongruent (Kiki-low (purple)/bouba-high (bleu).

C

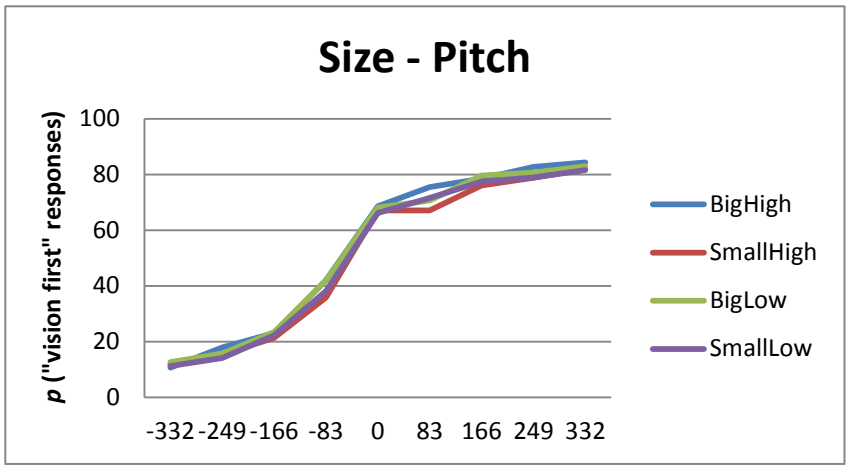


Figure 2. (C) size-pitch: congruent (small-high (red)/big-low (green), incongruent (small-low (purple)/big-high (bleu).

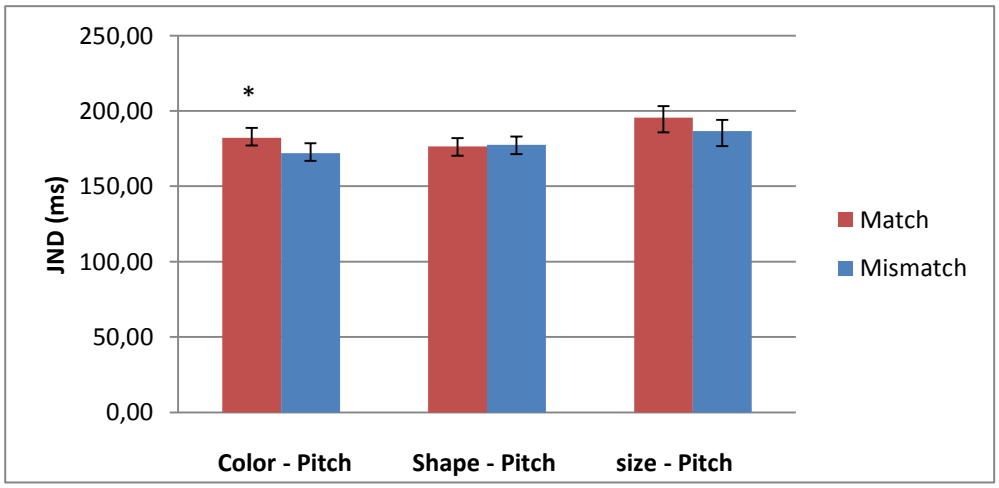


Figure 3. Average JNDs for the matched and the mismatched condition for the three audiovisual pairs, such as color-pitch, shape-pitch and size-pitch. Significant difference between groups ($p<0.05$) is highlighted by asterisk.

3.2. Color – Pitch condition

Analysis of the JND data for this crossmodal correspondence pair showed a significant main effect of color-pitch match [$F(1,57)=8.252$, $p< .05$, $\eta^2 = 0.126$], with participants finding it significantly more difficult to judge correctly the temporal order of auditory and visual stimuli when color and pitch were matched (M = 182,20 ms), that is when they perceived them as referring to the same underlying multisensory event. On the contrary, participants revealed improved performance when color and pitch were mismatched (M = 171,96 ms), that is when they perceived these audiovisual stimuli as not referring to the same underlying multisensory event.

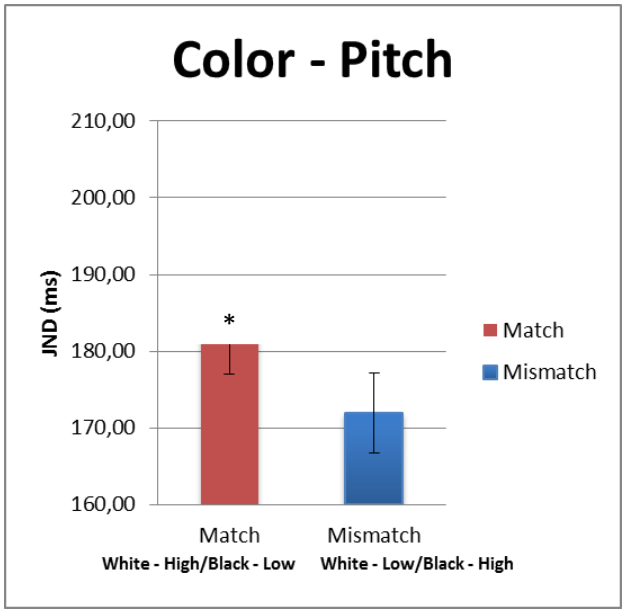


Figure 4. Average JNDs for the matched (i.e., white color-high pitch, and black color-low pitch) and mismatched (i.e., white color-low pitch and black color-high pitch) audiovisual stimuli.

Overall, by utilizing this specific crossmodal correspondence pair, it was revealed that participants had difficulty in discriminating which stimulus either the visual or the auditory, had been presented first, specifically, when these stimuli referred to the same underlying multisensory event. In other words it seems that when participants assumed that the presented audiovisual pair was tightly bind they could not judge which of the stimulus had been presented first. This outcome provides the first robust psychometrical evidence that the "unity assumption" can facilitate the crossmodal binding of multisensory information in audiovisual stimuli, based on their low level characteristics at a perceptual level.

This conclusion is warranted by the fact that no bias has affected participants' decision concerning the order of appearance of the stimuli, due to the experimental design we used, which is the TOJ-task. The main advantage of the specific design is the fact that participants are asked to discriminate the order of appearance of the presented audiovisual stimuli (i.e., "vision first" or "sound first") regardless of fitting with each other or not. On the contrary, had we used a simultaneity judgment task, we might have had evoked biased responses based on the simultaneously appearance of the audiovisual stimuli.

Besides, it has to be noted that it was eliminated any possibility of participants' having somehow "seen through" the experiment, and as a result to have performed more accurately on the TOJ trials in which the stimuli were mismatched, due the fact that it would have been unclear to them which response either audition or vision first, is the appropriate because of the multiple values of SOAs we have used. This is also one of the main advantages of the specific experimental design we have used, the TOJ task over the simultaneity judgment task when attempting to test the "unity assumption".

3.3 Shape – Pitch

Analysis of the JND data for this crossmodal correspondence pair showed none effect of matched [$F(1,57)=0.091$, $p=0.764$, $\eta^2 = 0.002$] comparatively to the mismatched condition, which means that participants in both conditions either matched or mismatched had approximately the same performance, as JND is about the same, for both matched (M = 176,41 ms), and mismatched (M = 177,42 ms) audiovisual stimuli.

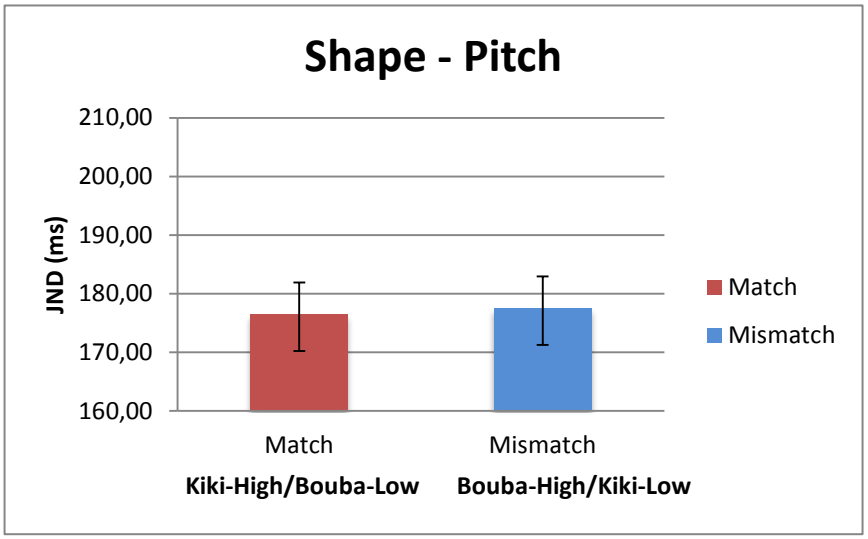


Figure 5. Average JNDs for the matched (i.e., kiki shape-high pitch and bouba shape-low pitch) and mismatched (bouba shape-high pitch and kiki shape-low pitch) audiovisual stimuli.

This means that participants had approximately the same difficulty in judging correctly the temporal order of the presented audiovisual stimuli (i.e., “visual” or “auditory” first) for both matched and mismatched condition. In other words, it seems that participants (in both

conditions) perceived these audiovisual stimuli as referring to the same underlying multisensory event. As a result, when they assumed that the presented audiovisual pair was tightly bind they could not judge which of the stimuli either the visual or the auditory had been presented first.

3.4 Size – Pitch

Analysis of the JND data for this crossmodal correspondence pair showed none effect of matched [$F(1,57)=1.702$, $p=0.197$, $\eta^2 = 0.029$] comparatively to the mismatched condition, which means that participants in both conditions either matched or mismatched had approximately the same performance, as JND is about the same, for both matched (*M*=195,54 ms), and mismatched (*M*=186,52 ms) audiovisual stimuli.

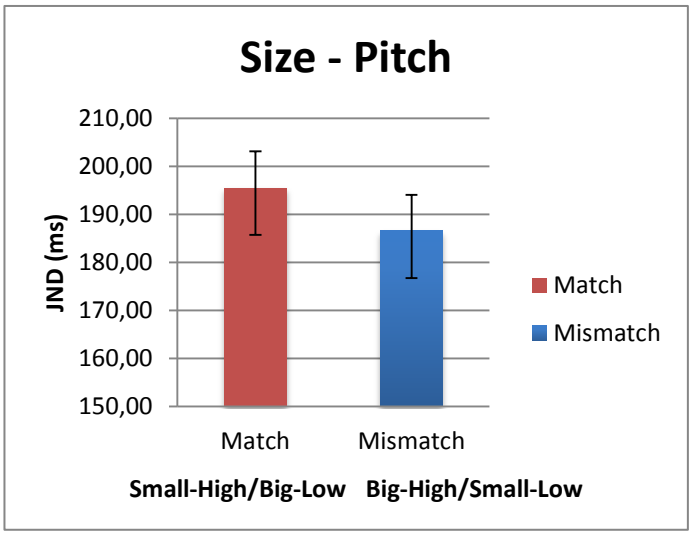


Figure 6. Average JNDs for the matched (i.e., small size-high pitch and big size-low pitch) and mismatched (big size-high pitch and small size low pitch) audiovisual stimuli.

This means that participants had approximately the same difficulty in judging correctly the temporal order of the presented audiovisual stimuli (i.e., “visual” or “auditory” first) for both matched and mismatched condition. In other words, it seems that participants (in both conditions) perceived these audiovisual stimuli as referring to the same underlying multisensory event. As a result, when they assumed that the presented audiovisual pair was tightly bind they could not judge which of the stimuli either the visual or the auditory had been presented first.

4. Discussion

In the present study, we utilized three classic and extensively used pairs of crossmodal correspondences consisted of color-pitch, shape-pitch, and size-pitch stimuli, in order to investigate the effect, if any, they would have to the “unity assumption”. For that purpose we used the experimental design of TOJ-task in which participants were asked to judge the order of appearance of the presented stimuli. We hypothesized that according to the “unity assumption” participants should have difficulty in discriminating the order of appearance of the presented stimuli when they perceive them as one unitary multisensory event. In other words, when participants assumed that the presented crossmodal correspondences constitute a tightly bind pair they would not been able to break the connection. As far it concerns the first audiovisual color-pitch pair we used, results are in aligned with our hypothesis. Specifically, we have found increased JNDs in matched, rather than mismatched condition, which means that participants had indeed difficulty in judging the order of appearance of the presented stimuli assuming that they constitute a solid unitary multisensory event. This denotes that when participants assumed that the presented audiovisual pair was tightly bind they could not judge which of the stimulus either the visual, or the auditory, had been presented first. This

outcome provides evidence that the "unity assumption" can facilitate the crossmodal binding of multisensory information in audiovisual stimuli, based on their low level characteristics at a perceptual level.

Contrary to our hypothesis, though, results from the other two conditions (i.e., shape-pitch and size-pitch) did not succeed in providing evidence according to which "unity assumption" has any influence on crossmodal binding, due to the fact that it was not found any significant difference in participants' performances in both matched and mismatched conditions. Our results revealed that the crossmodal correspondences consisted of visual (i.e., shape, size) and auditory (pitch) stimuli – the existence of which has been demonstrated by many cross-cultural (e.g., Bremner, Caparos, Davidoff, Fockert, Linnell, & Spence, 2013; Parise & Spence, 2012), and developmental (e.g., Fernandez-Prieto, Navarra, & Pons, 2015; Maurer, Pathman, & Mondloch, 2006) studies, together with variety of experimental designs (e.g., Evans & Treisman, 2010; Parise & Spence, 2009, 2012) – did not influence participants performance in our study, by utilizing the experimental design of TOJ-task, as participants had almost similar performance in both matched and mismatched conditions.

In order to approach the whole issue deeper, we accumulated all participants' responses for the three crossmodal correspondences as presented in Figure 7. The three groups for each audiovisual pair show participants performances regarding whether or not they are included in "unity" or "not unity" group or "none" for both matched and mismatched conditions. It has to be mentioned that for all the above crossmodal correspondences, the vast majority of the participants (total 58 participants) are concentrated in the two groups (i.e., "unity" – "not unity"), with the exception of only four participants in the color-pitch pair, three participants in the shape-pitch pair and four participants in the size-pitch pair. Furthermore, we can see that by the exception of shape-pitch crossmodal audiovisual pair, which gives raise to the

“not unity” group, the other two crossmodal audiovisual pairs (i.e., color-pitch and size-pitch) reveal almost the same pattern in participants’ classification – that is the “unity” group.

Specifically, in regard to the color-pitch audiovisual pair (figure 7 A.) we can see that participants in both groups (i.e., “unity” – “not unity”), seem to have similar tendencies while they are classified in both “unity” and “not-unity” group (with the former to be enhanced, especially in matched condition, comparatively to the latter). So one reasonable question is to investigate why some participants are influenced by the assumption they made that some correspondences refer to the some underlying multisensory event (i.e., top-down factors), while others are not (bottom-up, stimulus–driven factors). Some years ago Marks (1987) found strong crossmodal relation between colors and pitch audiovisual stimuli in a study where participants, exposed in RT experimental design, replied more quickly and accurately in matched rather than mismatched conditions. These results have led researcher to the conclusion that the cross-modal matches —probably determined by relative values of their attributes— may be defined from perceptual matching. Thus, he argued that the sensory cross-modal interactions in speed and accuracy of response take place at a sensory/perceptual stage of processing. The advantage to processing conferred by cross-modally matching stimuli may derive from activation of common perceptual rather than semantic units (Marks, 1987).

It has to be mentioned that, when participants are presented with stimuli on two corresponding perceptual dimensions of auditory pitch (high, low) and visual lightness (white, black), the classification of a tone as `high' is both quicker and more accurate when it is accompanied by a white (versus black) visual stimulus, and classification of a tone as `low' is quicker and more accurate when it is paired with a black stimulus (Marks 1987; Melara 1989). The semantic coding hypothesis is consistent with findings that congruence effects can occur not only when the two dimensions share verbal labels (e.g., pitch of a sound and

vertical position of a light), but also when they do not (e.g., pitch and lightness; Martino & Marks, 1999). Thus, the fact that participants responded quickly to congruent combinations of pitch and lightness (black-low pitch, white-high pitch) than to incongruent combinations (black-high pitch, white-low pitch; Melara, 1989) can be explained by the semantic coding hypothesis as the result of dimensional interactions that take place at a post-perceptual locus (Martino & Marks, 1999). Moreover, it has to be mentioned that multisensory integration can be attributed by both top-down (i.e., cognitive factors influence the assumption whether or not different stimuli refer to the same underlying multisensory event; Radeau & Bertelson, 1977) and bottom-up or stimulus driven factors (i.e., spatiotemporal factors or correlation of different stimuli features originate from different modalities).

Thus, taking into account the above, findings from this specific audiovisual color-pitch pair can be both attributed in bottom-up factors regarding the information processing, as also in the specific perceptual stage of processing. A simple working model assumes a series of stages for processing each perceptual dimension. Processing begins with sensory/perceptual encoding, is followed by comparison of internal representation to some reference, and ends with response selection. Semantic receding can intervene between encoding and comparison (Marks, 1987). Thus, it is possible that participants who belong to the "unity" group, simply to be influenced by cognitive factors regarding whether or not different stimuli refer to the same underlying multisensory event, while on the contrary others' who belong to the "not-unity" group to be influenced by stimulus-driven factors (i.e., spatiotemporal factors or congruency of different stimuli features originate from different modalities).

As it has already been mentioned, participants on the shape-pitch audiovisual pair, had approximately similar performance in both conditions, either matched or mismatched. Moreover, we can see (Figure 7 B.) that participants' in the "non-unity" group have the tendency of increased JNDs on the mismatched condition, rather than the matched one.

A

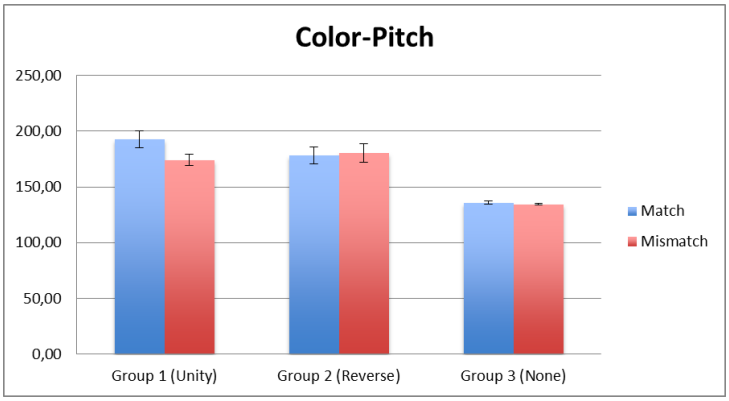
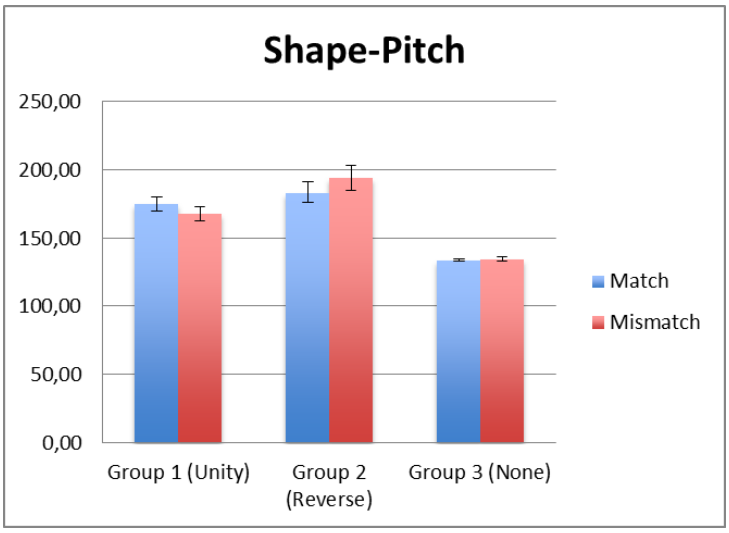


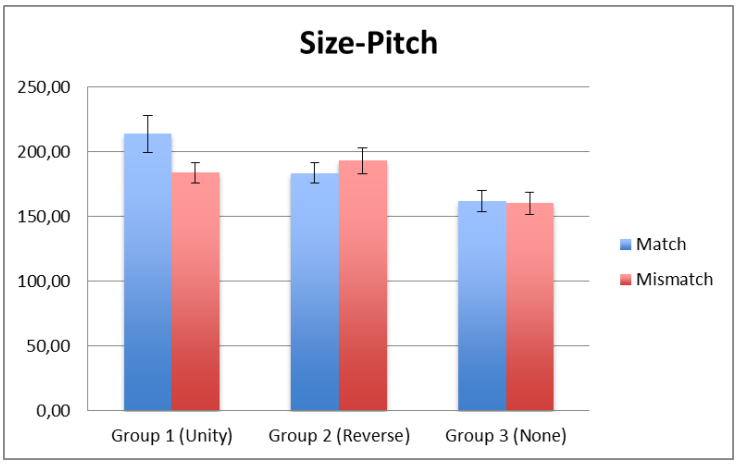
Figure 7. The three groups for each audiovisual pair show participants performances regarding whether or not they are included in "unity" or "not unity" or "none" group for both matched and mismatched conditions. The vast majority of the participants are concentrated in "unity" – "not unity" group. (A) Participants in both groups seem to have similar tendencies while they are classified in both "unity" and "not-unity" group (with the former to be enhanced, especially in matched condition, comparatively to the latter).

B



(B) Participants in the "non-unity" group have the tendency of increased JNDs on the mismatched condition, rather than the matched one.

C



(C) Participants that are accumulated in the "unity" group, show the tendency of increased JND on matched rather than mismatched condition - though, it is not significant difference, comparatively to the mismatched condition.

Finally, we have utilized the same size-pitch audiovisual pair that was used by Parise and Spence (2009). In their study it was found that participants had difficulty in judging the temporal order of presented stimuli when they were presented in matched (i.e., big or small visual circles combined with low, 300 Hz, or high, 4500 Hz, tone respectively) condition, rather than mismatched (i.e., big or small visual circles combined with high, 4500 Hz, or low, 300 Hz, tone, respectively) condition (Parise & Spence, 2009). The particular audiovisual pair is argued by the researchers to constitute a form of crossmodal (they call it synesthetic) association between stimuli presented by different modalities, thus synesthetic correspondence is depicted on participants' reliability on audiovisual TOJs and spatial localization judgments (Parise & Spence, 2009). More recently, it was found that participants responded extremely fast when a big visual cycle and a low pitch tone (300 Hz) were assigned to the same respond key as well as a small cycle and a high pitch tone (4500 Hz)

respectively (congruence condition), while on the contrary their reaction time was slower when a big cycle and high pitch (4500 Hz) tone were assigned to the same respond key as well as a small cycle and low tone (300 Hz), respectively (incongruence condition; Parise & Spence, 2012). On the contrary in our study this finding was not replicated. For the latter case, findings could be attributed to the experimental design that was used by the researchers, but in the former case, this finding is surprisingly strange, because we have used exactly the same experimental JND design. Furthermore, as we can see (Figure 7 C.) participants that are accumulated in the "unity" group, show the tendency of increased JND on matched rather than mismatched condition. Thus, taking together all the above it seems rather reasonable to claim that the different results on both studies, might bring into account the way that crossmodal correspondences can impact the efficacy of human information processing.

Therefore, at the present time, it would seem fruitful to assume that our findings regarding the influence of the these specific crossmodal correspondences pairs we utilized to the "unity assumption" reflect some combination of both top-down and bottom-up factors influencing multisensory integration. These two factors (i.e., bottom-up, and top-down) presumably operate simultaneously in order to facilitate the appropriate multisensory integration of environmental events under the majority of naturalistic conditions (e.g., Radeau & Bertelson, 1977). Additionally, more recently studies have investigated the role of automaticity in the crossmodal correspondences. As we have already been mentioned variety of studies, (e.g., Evans & Treisman, 2010; Klapetek et al., 2012; Parise & Spence, 2012), by utilizing different experimental designs (e.g., speeded classification; Evans & Treisman, 2010; visual search; Klapetek et al., 2012; Implicit Association Test; Parise & Spence, 2012), and crossmodal correspondences audiovisual pairs, have argued that crossmodal correspondences are automatic, thus, they are easily appeared. On the contrary, other researchers (i.e., Chiou, & Rich, 2012), have argued that crossmodal correspondences are not

automatic in the sense that they are 'primarily mediated by cognitive processes after initial sensory encoding' and occur at a 'relatively late stage of voluntary attention orienting' (Chiou & Rich, 2012). Likewise, others have suggested that the crossmodal correspondence between auditory pitch and visual brightness operates "at a more strategic (i.e., rather than at an automatic or involuntary) level" (Klapetek, Ngo, & Spence, 2012). Particularly, it was found that the crossmodal correspondence between auditory pitch and visual brightness isn't solely stimulus driven. However, this cannot provide evidence the process underlying the effect of crossmodal correspondence is conscious, controlled, or goal-dependent (Spence, & Deroy, 2013).

Very recently, it has been argued that the binding tendency (i.e., the brains' tendency to integrate or bind stimuli originate from different sensory modalities; Odegaard, & Shams, 2016) is stable over time, where spatial and temporal integration processes are not governed by a single, universal parameter in the brain (Odegaard, & Shams, 2016). In particular, researchers have utilized the Bayesian causal inference (BCI) which is a computational model in order to investigate the binding tendency in each observer, in a manner that did not confound binding tendency with the precision of unisensory encoding. They argued that spatial and temporal integration processes are not governed by single universal parameters (Odegaard, & Shams, 2016). If that is the case, the it probably can explained the fact that some people appear to be influenced by top-down factors, while others by bottom-up, stimulus-driven factors.

Summarizing all the above, as already has been mentioned, the multisensory integration can be modulated by both top-down or stimulus-driven factors and by bottom-up factors (Vatakis & Spence, 2007). Therefore, in our study contrary to other studies where informationally rich pairs of stimuli were used in order to investigate the "unity assumption" (i.e., Vatakis, Ghazanfar, & Spence, 2008; Vatakis & Spence; 2007, 2008) we were restricted

in utilizing crossmodal correspondences, based on the correspondences between simple characteristics. Though, our findings cannot rule out the possibility of being attributed to the top-down (i.e., cognitive) factors of multisensory integration. Thus, it seems that findings from our study can be interpreted as resulted from various levels of cognitive processing, depending on the task instructions and requirements, the strategy used by the participants, and their degree of awareness of the crossmodal correspondence (Klapetek, Ngo, & Spence, 2012).

References

Baart, M., Stekelenburg , J. J., & Vroomen, J. (2014). Electrophysiological evidence of speech-specific audiovisual integration. *Neuropsychologia, 53*, 115-121.

Bremner, A. J., Caparos, S., Davidoff, J., Fockert, J. d., Linnell, K. J., & Spence, C. (2013). "Bouba" and "Kiki" in Namibia? A remote culture make similar shape-sound matches, but different shape-taste matches to Westerners. *Cognition, 126*, 165-172.

Chiou, R., & Rich, A. N. (2012). Cross-modality correspondence between pitch and spatial location modulates attentional orienting. *Perception, 41*, 339-353.

Chuen, L., & Schutz, M. (2016). The unity assumption facilitates cross-modal binding of musical, non-speech stimuli: The role of spectral and amplitude envelope cues. *Atten Percept Psychophys*.

Corral, V. C. (2015). Cross-Cultural differences in crossmodal correspondences between western and xhosa children: Implications for design. Barcelona: Treball de Fi de Grau.

Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision, 10*(1), 1-12.

Fernandez-Prieto, I., Navarra, J., & Pons, F. (2015). How big is this sound? Crossmodal association between pitch and size in infants. *Infant Behavior & Develpment, 38*, 77-81.

Hubbard, T. (1999). Synesthasia-like mappings of lightness,pitch,and melodic interval. *The American Journal of Psychology, 109*(2), 219-238.

Jones, J. A., & Jarick, M. (2006). Multisensory integration of speech signals: the relationship between space and time. *Experimental Brain Research, 174*(3), 588-594.

Klapetek, A., Ngo, M. K., & Spence, C. (2012). Does crossmodal correspondence modulate the facilitatory effect of auditory cues on visual search? *Atten Percept Psychophys, 74*, 1154-1167.

Ludwig, V. U., Adachid, I., & Matsuzawad, T. (2011). Visuoauditory mappings between high luminance and high pitch are shared by chimpanzees (Pan troglodetes) and humans. *PNAS, 108*(51), 20661-20665.

Marks, L. E. (1987). On Cross-Modal similarity: Auditory-Visual Interactions in Speeded Discrimination. *Journal of Experimental Psychology: Human Perception and Performance, 13*(3), 384-394.

Martino, G., & Marks, L. E. (1999). Perceptual and liguistic interactions in speeded classification: tests of the semantic coding hypothesis. *Perception, 28,* 903-923.

Maurer, D., Pathman, T., & Mondloch, C. J. (2006). The shape of boubas: sound-shape correspondences in toddlers and adults. *Developmental Science, 9*(3), 316-322.

Melara, R. D. (1989). Dimentional Interaction Between color and pitch. *Journal of Experimental Psychology: Human Perception and Performance, 15*(1), 69-79.

Morein-Zamir, S., Soto-Faraco, S., & Kingstone, A. (2003). Auditory capture of vision: examining temporal ventriloqism. *Cognitive Brain Research, 17,* 154-163.

Oberman, L. M., & Ramachandran, V. S. (2008). Preliminary evidence for dificits in multisensory integration in autism spectrum disorders: The mirror neuron Hypothesis. *Social Neuroscience, 3*((3-4)), 248-355.

Odegaard, B., & Shams, L. (2016). The brain's tendency to bind audiovisual signals is stable but not general. *Psychological Science, 27*(4), 583-591.

Parise, C. V. (2015). Crossmodal Correspondences: Standing Issues and Experimental Guidelines. *Multisensory Research*. doi:10.1163/22134808-00002502

Parise, C. V., & Spence, C. (2009). 'When birds of a feather flock together': Synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLos ONE, 4*(5).

Parise, C. V., & Spence, C. (2012). Audiovisual crossmodal correspondences and sound symbolism: a study using the implicit association test. *Exp Brain Res, 220*, 319-333.

Parise, C., & Spence, C. (2008). Synesthetic congruency modulates the temporal ventriloquism effect. *Neuroscience Letters, 442*, 257-261.

Radeau, M., & Bertelson, P. (1977). Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations. *Perception & Psychophysics, 22*(2), 137-146.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Atten Percept Psychophys, 73*, 971-995.

Spence, C., & Deroy, O. (2013). How automatic are crossmodal correspondences. *Consciousness and Cognition, 22*, 245-260.

Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio-visual speech perception is special. *Cognition*, B13-22.

Vatakis, A., & Spence, C. (2007). Crossmodal Binding: Evaluating the 'unity assumption' using audiovisual speech stimuli. *Perception & Psychophysics, 69*(5), 744-756.

Vatakis, A., & Spence, C. (2008). Evaluating the influence of the "unity assumption" on the temporal perception of realistic audiovisual stimuli. *Acta Psychologica 127*, 12-23.

Vatakis, A., Ghazanfar, A. A., & Spence, C. (2008). Facilitation of multisensory integration by the "unity effect" reveals that speech is special. *Journal of Vision, 8*(9), :14,1-11.

Vroomen, J., & Keetels, M. (2006). The spatial constraint in intersensory pairing: No role in Temporal Ventriloquism. *Journal of Experimental Psychology: Human Perception and Performance, 32*(4), 1063-1071.

Vroomen, J., & Stekelenburg, J. J. (2010). Perception of intersensory synchrony in audiovisual speech: Not that special. *Cognition*.