New Design Ideas
Vol.9, No.1, 2025, pp.5-29
https://doi.org/10.62476/ndi.91.05

THE PERCEPTION OF A VISUAL BISTABLE DESIGN CAN BE SEMANTICALLY MODULATED BY AN INCOMPREHENSIBLE SPOKEN LANGUAGE

Guillermo Rodríguez-Martínez*

Faculty of Arts and Design, Universidad Jorge Tadeo Lozano, Bogotá, Colombia

Abstract. Semantic congruency is a phenomenon by which a stimulus is perceived, semantically speaking, in a way that perception is close related to contents provided by other sources. This way, when an observer is looking at an image, its perception can be influenced by the semantic load of auditory stimuli, if their content is related to the meaning of the image. When it happens, the semantic congruency effect arises. This study was aimed at establishing the modulating effect that tones of voice can exert on the perception of an ambiguous image, wanting to vindicate the semantic congruency effect. Thirty-two participants viewed the bistable image *My girlfriend or my mother-in-law*, while listening to two incomprehensible modulating tones of voice. An eye-tracker device was used to measure the time of visual percepts that were congruent with the semantic load of the audios. There were significant differences between the durations of the congruent and incongruent visual percepts with the modulating audio. Therefore, it is concluded that tones of voice can bring about an impact on the perception of ambiguous images. An incomprehensible spoken language modulates the understanding of bistable visual designs, where, by means of top-down perceptual modulating processes, the semantic congruency effect emerges.

Keywords: *Bistable images, visual perception, semantic congruency, visual bistable design.*

****Corresponding Author:*** *Guillermo Rodríguez-Martínez, Faculty of Arts and Design, Universidad Jorge Tadeo Lozano, Bogotá, Colombia, Tel.: 6012427030, e-mail: guillermo.rodriguez@utadeo.edu.co*

Received: *2 August 2024;* ***Accepted:*** *13 February 2025;* ***Published:*** *2 April 2025.*

1. Introduction

1.1. Background information

The semantic congruency effect is a perceptual psychological phenomenon by which information provided or stored in memory influences the perception of a visual stimulus due to a semantic match (Di Stefano & Spence, 2023; Spence, 2011; Hsiao *et al*., 2012). That match corresponds to the content (or the meaning) that the influential information has in relation to the semantic load of the image (Marroquín-Ciendúa *et al*., 2020; Rodríguez-Martínez *et al*., 2021). It is possible to produce the semantic congruency effect when auditory stimuli are being heard while observing a bistable image (Hsiao *et al*., 2012; Rodríguez-Martínez, 2023; Yeh *et al*., 2011). It occurs if the semantic load of the auditory input is closely related to the meaning of one of the recognizable percepts of the bistable image (Smith *et al*., 2007). Given the fact that bistable visual stimuli allow for two possible interpretations, they can also be called *ambiguous images* (Gijs & van Ee, 2006; Okazaki *et al*., 2008). When observers are

How to cite (APA):
Rodríguez-Martínez, G. (2025). The perception of a visual bistable design can be semantically modulated by an incomprehensible spoken language. *New Design Ideas*, *9*(1), 5-29 https://doi.org/10.62476/ndi.91.05

looking at this kind of images, the perception shifts between the two possible visual percepts that can be interpreted (Devia *et al.*, 2022; Leopold & Logothetis, 1999; Rodríguez-Martínez *et al.*, 2022). These leaps are known as *perceptual reversals* (Baker *et al*., 2015; Clément & Demel, 2012; Intaité *et al*., 2014; Koivisto & Pallaris, 2024; Sandberg *et al*., 2014; Rodríguez-Martínez *et al*., 2022). In Figure 1, a bistable image is shown, the one designed by Boring (1930), called *My girlfriend or my mother-in-law*. When viewing at this visual stimulus, observers can perceive two images: the figure of an old woman and the figure of a young lady. This bistable image has been classified as *In meaning-content reversals* bistable image, because the two percepts have a similar level of salience and each of them comes to be different in terms of shape and meaning (Rodríguez-Martínez, 2023). According to Carbon (2014), the original image, known as *The young-old-woman illusion*, was popular in Germany in the 19th century because it had been frequently depicted on postcards. Boring (1930) was the first who presented this ambiguous image within the scope of the science of psychology and visual perception (Carbon, 2014). It was until 1930 that Edwin Boring (1930) introduced the figure to psychologists in a paper titled *'A new ambiguous figure'*.

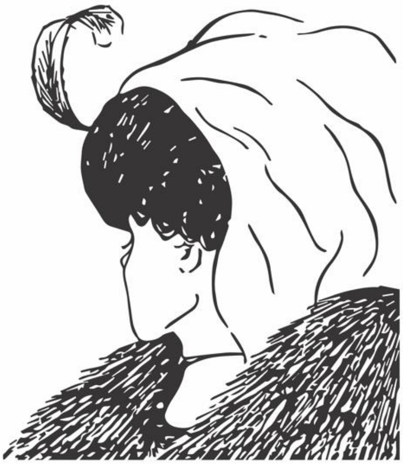


Figure 1. The ambiguous image *My girlfriend or my mother-in-law*, put forward by Boring (1930)
Source: Carbon (2014)

Regarding modulating processes that are involved in bistable perception, it has been stated that the physical characteristics of ambiguous images can modulate the resulting perception (Poom, 2024). It depends to some extent on the way the stimulus is being viewed and also on the areas of the image upon which the eyes are fixed (Gale & Findlay, 1983; García-Pérez, 1989; García-Pérez, 1992; Hsiao *et al*., 2012; Marroquín-Ciendúa *et al*., 2020; Rodríguez-Martínez, 2024). This type of influence on the visual perception is known as bottom-up modulating factor (Hsiao *et al*., 2012; Meng & Tong, 2004). Consequently, the perceptual understanding of a bistable image also depends on isolated information processed from stimulus' physical characteristics (Brouwer & van Ee, 2006). Nevertheless, concepts and also predispositions can take part in the perceptual process (Hsiao *et al*., 2012). It implies, in terms of basic psychological processes, what is called *top-down processing* (Barrera & Calderón, 2013). In this spirit and coming back to the question of the semantic congruency effect, when auditory stimuli are influencing the perception of an ambiguous image, top-down modulation is operating, if the semantic load of the acoustic information is related to the meaning of the percept perceived from de observation of the bistable visual stimulus (Kesoglou & Mikellidou, 2024).

Given the fact that there can be two sensorial modalities operating when emerging the semantic congruency effect, a crossmodal perceptual process also arises, because two different sensorial inputs (auditory and visual) are taking part in the parceptual experience (Hsiao *et al*., 2012; Roberts *et al*., 2024). When considering top-down perceptual mechanisms, what is implied is that the interpretation of the ambiguous image is defined by information stored in memory (Tarder-Stoll *et al*., 2020) or also by inputs that forays into the perceptual system (Kiefer, 2007). According to Intaité et al. (2013), there can be and interaction between both top-down and bottom-up perceptual processes, which leads to the involvment of different attentional mechanisms. In this sense, the shifts between the two possible interpretations of ambiguous visual stimuli can be both, involuntary or voluntary, which brings about the interaction of the two attentional modulating processes (Katsuki & Constantinidis, 2014; Kornmeier *et al*., 2009).

When considering the observation of the phenomenon of semantic congruency by means of crossmodal stimulation, experiment setups are designed in such a way that bistable perception paradigms are incorporated. It works this way: first of all, a bistable figure that accepts two different interpretations (in terms of meaning) is selected, recognizing the two different meanings of each single visual percept. In this regard, the images of the *In meaning-content reversals* kind come to be useful, if it is possible to find auditory stimuli that are linked to the two possible interpretations of the ambiguous image. In this spirit, the second relevant aspect (concerning sensory stimulation) emerges, which is the presentation of audios that have the capability of modulate (top-down modulating mechanism) the perception of the visual image. The semantic congruency effect will be observed if the semantic load of the audios influences the perception of the bistable image in a way that what is seen matches semantically with the semantic load of the auditory stimulation. There would be a crossmodal sensory stimulation because two different kinds of sensory stimulation are present, the visual and the auditory.

In view of the foregoing, the present study is related to this sort of scientific observation. Audios that can provide semantic load which, in turn, can influence the perception of a bistable image, were used so as to observe the occurrence of the semantic congruency effect. As it will be explained later, the auditory stimuli used were voices saying words in an incomprehensible language, but considering that their tone (voice of a female old woman and voice of a female young woman) had the semantic load that could match the two possible perceptions of the bistable image *My girlfriend or my mother-in-law*: an old woman or a young lady. By analyzing the congruences between the visual percepts reported and the audios displayed, the semantic congruency effect was observed. It was expected that the visual percept 'old woman' would be more dominant when participants listened to the voice of an old woman. On the other hand, the percept 'young woman' would be more dominant for the perceiver when the provided audio was the voice of a young lady.

1.2. Related work

As previously mentioned, several studies have reinforced the idea that both *top-down* and *bottom-up* processes imply an effect on perception (Intaité *et al*., 2013), in such a way that a perceptual integration can occur between both mechanisms (Schuman et al., 2021). Bottom-up perceptual processes are mediated by the physical characteristics of the stimuli, whereas top-down mechanisms imply the influence of information stored in memory or provided externally (Rodríguez & Castillo, 2028). As Hsiao et al. (2012) proposed, it is possible to use crossmodal stimulation to convey semantic content while perceiving ambiguous images. This way, a top-down modulating process would be emerging (Marroquín-Ciendúa *et al*., 2020). In this sense, semantic keys that are provided not from the visual sensory modality, but from other sensory sources such as sight, smell, hearing, etc., have the capability of conveying information whose semantic content can influence the perceptual outcome (Rantala *et al*., 2024).

Taking into account what was stated by Chen et al. (2011), the environment normally provides contextual information via several different sensory modalities rather than just one. In this sense, it is possible that a specific semantic content provided by auditory stimuli affects the interpretation of an ambiguous visual figure if the observer is seeing the image while listening to the auditory information itself (Zeljko *et al*., 2022). In view of the foregoing and taking into account that the study that is outlined in this article is related to the assessment of the semantic congruency effect based on the use of an incomprehensible spoken language (where incomprehensible spoken words operate as top-down modulating auditory stimuli), different studies are shown here in order to understand the scope of them in terms of the evaluation of the semantic congruency effect by using bistable images and tones of voice as modulators.

Within the framework of an experiment conducted by Smith et al. (2007), an androgynous face was exposed for observers to report whether it seemed masculine or feminine. The particularity of the experiment was that, while the observers looked at the ambiguous face, audios related to male and female tones of voice were exposed. Researchers aimed at establishing whether listening to one or another sound frequency could modulate the perception of the perceived gender of the face. Indeed, that study revealed that sound stimulation significantly influenced the interpretation that observers made of the image in relation to gender identification. Thus, when a sound stimulation (referred to a masculine voice tone) was provided, the identification of the masculine gender was reported to the exposed androgynous face. On the other hand, when seeing the same androgynous face, its gender was identified as female when the audio provided corresponded to a female tone of voice. In this sense, a disambiguation of the neutrality of the androgynous face (in terms of its gender) emerged, a fact that is related to the modulation mediated by the semantic loads of the two tonalities of voices used in the experimental task.

Similar to this study, Rodríguez-Martínez and Sojo (2022) conducted an experiment in which an androgynous face was displayed on a remote eye-tracker device. In this study, just one female human voice was used so as to determine if the phenomenon of semantic congruency could emerge if participants identified a female face when observing the androgynous stimulus. The results showed that there was a significant increase in the perception of a female face when hearing an auditory stmilulus with the semantic load of a female human voice. Once again, the semantic correspondance between the auditory stimulus and the perceived visual one vindicated the top-down modulating effect that the voice exerts on visual perceptual processes associated with the disambiguation of a bistable visual stimulus.

For their part, Frassinetti et al. (2002), after having considered different spatial positions for visual and auditory stimulation, were able to demonstrate the existence of an integrated spatial visuo-auditory system in normal subjects, with functional properties similar to that described at neuronal and behavioural level in animals. In this study, experiments were conducted taking into account both unimodal perceptual influences and crossmodal perceptual modulations. When being focused on the results concerning semantic correspondences (marked by the perceptual integration between visual stimuli and auditory ones), what was found is that multisensory interaction only takes place when periods of peak activity of unimodal discharge produce overlap. In other words, one factor that was considered here was the temporal rule of multisensory integration (Spence, 2011). According to this rule, a perceptual multisensory correspondence occurs in close temporal proximity. If the two different sensory stimuli are presented separated by long intervals, then they are processed as separate events (Frassinetti *et al*., 2002). In the latter case, an auditory stimulus would play the role of an alerting signal (Robertson *et al*., 1998). With this in mind, the temporal proximity, or even the simultaneous sensory acquisition of the two signals in question, are necessary to produce a perceptual integration, where the correspondence can be marked not only by semantic similarities, but also by a close time proximity (Cox *et al*., 2015; Spence, 2011). This has previously been vindicated in such a way that the definition of multisensory integration mentions that this phenomenon is more likely to occur the closer that the presentation of stimuli from different sensory modalities is done in time (Cox *et al*., 2015; Spence, 2011, 2013; Spence & Squire 2003). In order for the multisensory stimulation to be integrated, crossmodal correspondences require a sort of compatibility between attributes or dimensions so that the integration is produced by a kind of perceptual similarity (Di Stefano & Spence, 2023).

When reviewing studies that have used bistable images so as to vindicate the semantic congruency effect by means of providing crossmodal sensory stimulation (visual and auditory), several studies that used the image *My girlfriend or my mother-in-law* appear (Yeh *et al*., 2011; Hsiao *et al*., 2012; Marroquín-Ciendúa *et al*., 2020; Rodríguez-Martínez *et al*., 2021). As far as the study conducted by Yeh et al. (2011) is concerned, the main findings demonstrated a relevant perceptual top-down modulating process, specially when the participants of an experiment were instructed to maintain their attention when indentifying a percept of the bistable visual stimulus. In that study, researchers used tones of voice related to each possible percept of the image *My girlfriend or my mother-in-law*. In this sense, they displayed two tones, a voice of an old woman and a voice of a young lady. As expected, there would be a prominent perception of the percept that was linked to the semantic load of the sound provided. The version of the image selected was the one that is in complete black and white contrast (Table 1).

Following the same main idea of research, Hsiao et al. (2012) tried to prove the effect of semantic congruency by modulating the perception of the same version of the image designed by Boring (1930). The image in question was used in four different experiments this way: in the first one, the participants had to view the bistable image while listening to the voice of an old woman, the voice of a young lady and other sounds like beeps, or else to no sound (Hsiao *et al*., 2012). The voices, rather than the content of a monologue or speech, provided the expected crossmodal context to produce the semantic congruency effect. The participants were instructed to report continuously either the old woman or the young woman, as they identify one or another visual dominant percept. In this experiment, participants identified the old woman as their first percept for more of the trials when the sound displayed was the congruent auditory female human voice (old woman). The most predominant congruent visual percept (congruent with the auditory stimulus) was the old woman. Nevertheless, it was found that both tones of voice were able to modulate the perception of the congruent percept, vindicated, this way, the semantic congruency effect. As for the second experiment, what had to be reported was the predominance of one of the competing percepts (Hsiao *et al*., 2012). This time, the auditory stimulation presented was never congruent with the visual image they must report. The results showed a predominant duration for the old woman in comparison with the reports concerning the young lady. According to Hsiao et al. (2012), there would have been a sort of bias that conditioned the predominance of the recognition of the old woman. This bias was assumed on the basis that the face of the old woman 'extends over a larger area and exhibits more front angle than the face of the young lady' (Hsiao *et al*., 2012). In relation to the third experiment, Hsiao et al. (2012) tried to prove that when participants fixated at a particular area that could favor one of the visual percepts, they were more likely to identify that percept initially and for more of the time after the initial recognition (in this case, they also wanted to prove bottom-up modulating perceptual influences). In order to control the position of the ocular fixations made on the image, an eye-tracker Eye-Link 2000 was used, with a sampling rate of 1000 Hz. Once again, auditory stimuli were presented simultaneously with the image, two different ones, voices of an old woman and of a young lady. What was found was that the semantic congruency effect was clearly observed, regardless of the areas fixated (that were different for each visual congruent percept reported). Finally, the fourth experiment wanted to demonstrate if the crossmodal semantic modulation would occur when manipulating selective attention over the bistable image. In this spirit, two factors, auditory stimuli (old woman voice and young lady voice) and selective attention (maintain elderly percept, passive or maintain the percept of the young lady) were variables controlled by researchers (Hsiao *et al*., 2012). This time, the predominance measure for the ambiguous image decreased when participants tried willingly to maintain a particular percept, which led to revealing the meaningful interaction between the auditory stimuli and the attentional factors involved (Hsiao *et al*., 2012).

By taking into account the results obtained in the two studies previously described, Marroquín-Ciendúa et al. (2020) conducted a study in which participants were wanted to report each time they recognized the old woman percept or the young lady percept, in relation to the bistable image *My girlfriend or my mother-in-law*. As Hsiao et al. (2012) had proposed, two different tones of voice were displayed, a voice of an elderly woman and a voice of a young woman. These auditory stimuli were voices pronouncing words. Given the fact that all participants were native speakers of Spanish, it was necessary to record and display the voices in a different language: French. As is known, French has some similarities with Spanish in such a way that some words sound similar due to their Latin origins. That is why the researchers decided to use a scale through which measure the comprehension of the audios. On the other hand, apart from aiming at assessing the semantic congruency effect, they wanted to determine a possible bottom-up modulating effect, based on the hypothesis that viewing some areas of the image might favor the perception of the percepts. To confirm the hypothesis, the researchers used an eye-tracker device (reference Tobii®, T-120) to observe if some modulating areas were related to the identified percept and considering, at the same time, the semantic top-down modulating influence provided by the audios. To select the bottom-up modulating areas, they took into account the proposals of Gale and Findlay (1983), using a delineated version of the image (Table 1). Four bottom-up modulating areas were established, whereby one of them favored the perception of the young lady and the other one the perception of the old woman. The results show that the area in which the eye of the young lady is placed favors the perception of the young lady. No matter this bottom-up modulating effect, the semantic congruency effect was observed. When the variable 'familiarity of the words' (measured by a scale) was analyzed, there was no a significant effect of the measurements of the understanding of the audios on the reports that were congruent with the visual percept recognized.

Another study was conducted to assess the semantic congruency effect using the image *My girlfriend or my mother-in-law*: this time, a new different variable was considered, the position of the observer in front of the image. Rodríguez-Martínez et al. (2021), authors of that study, displayed audios in French (old woman voice and young lady voice) so as to identify congruences between the visual percepts reported and the semantic load of the audios. All the participants were native speakers of Spanish. The image used was the delineated version put forward by Gale and Findlay (1983). They also used a Tobii® T-120 eye-tracker device, placing participants in two positions: up-right and up-side down. The results showed that the effect of semantic congruency emerged, but just in the group of participants that were placed in the up-right position. Another relevant finding was the fact that the area in which the eye of the young lady is located is a zone that strongly favors the recognition of the young lady.

When reviewing all these studies based on bistable perception paradigms, several issues come up within the context of evaluating the effect of semantic congruency. First of all, the use of crossmodal sensory stimulation, specifically, the combination of auditory stimuli with a bistable image that accepts two different interpretations, semantically speaking. Secondly, the concern about the correspondence that should exist between the semantic loads of the auditory stimuli and the semantic content of the percepts of the bistable stimulus. Thirdly, the fact that the effect of semantic congruency arises a perceptual disambiguation of the bistable image. As has been mentioned before, semantic congruence is marked by the relation established in terms of the semantic load between the content of the auditory stimulus and what is depicted by the image (Di Stefano & Spence, 2023; Feist & Gentner, 2007; Goolkasian & Woodberry, 2010; Hsiao *et al*., 2012; Smith *et al*., 2007). Given the fact that it is likely to use short stories (auditory stimulation) in such a way that they (their content) influence the perception of an ambiguous visual stimulus, the semantic content of the story will have to be related to the semantic content of one of the possible bistable image interpretations (Balcetis & Dale, 2007; Hsiao *et al*., 2012) if wanting to observe the semantic congruency effect. The foregoing implies that when observers are looking at an ambiguous image like *My girlfriend or my mother-in-law*, they can perceive easily the young woman if there is a simultaneous modulating audio, which could be the voice of a young woman (*e.g*. Hsiao *et al*., 2012; Rodríguez-Martínez *et al*., 2021). Conversely, if the modulating audio is the voice of an old woman, the observer may note the presence of an elderly woman in the image due to the semantic congruence effect (Hsiao *et al*., 2012).

In line with these ideas, there is something to consider in relation to the use of human voices as modulators: the recognition of the content of the words themselves. Every word has a meaning and a semantic load, which are related to a context and patterns of word co-occurrence (Diveica *et al*., 2024). If reviewing the related work put here, it is clear that all the studies used, as auditory modulators, tones or voices, but always understanding that the content of the words should not interfere with the perceptual process. The study that used pure tones (but not voices articulating words) assured that there were not going to be biases in relation to external semantic loads (Smith *et al*., 2007). The other studies used tones of female voices articulating words, trying to avoid the interference of the semantic load of the words pronounced. In order to reach that goal, the variable 'understanding of the content of the words' was controlled, by means of producing a speech in a different language (different from the mother tongue of participants) or also by using scales to measure the possible understanding of what the voices said.

1.3. Research gap

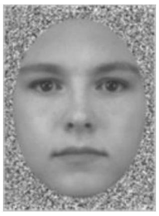
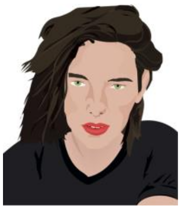
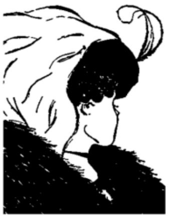
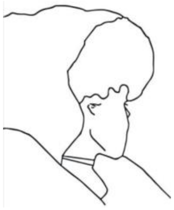
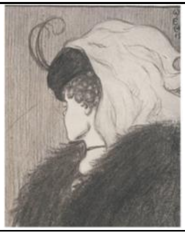
When using human voices as top-down modulators that can trigger the semantic congruency effect while viewing a bistable image, a fact that must be assured is that the semantic load provided should have the exact semantic content implicated in both of the possible visual percepts to be perceived visually. Although it is possible to use one word or sentences that mean the same in relation to the percepts of the ambiguous image, it is preferable to use just a type of voice if this kind of sound represents what is going to be interpreted when observing the bistable visual stimulus. In the case of the image *My girlfriend or my mother-in-law*, what researchers have done is selecting speeches by considering the tone of voice (voice of a young lady or voice of an old woman). In this regard, the words that are constituent parts of the speech does not have to produce a semantic cognitive load if the sensory integration that is going to be observed is linked to the semantic units 'old woman' and 'young lady'. As such, researchers have tried to use female voices when analyzing the semantic congruency effect with the image *My girlfriend or my mother-in-law,* but being aware of the fact that participants were not able to understand the content of what was being said. Instead, the participants would have to recognize the source of the voice and old woman or a young one. In view of the foregoing, it is common to use unfamiliar languages, just as Yeh et al. (2011), Hsiao et al. (2012), Marroquín-Ciendúa et al. (2020) and Rodríguez-Martínez et al. (2021) did.

Considering that there are not enough studies through which assess the semantic congruency effect by using spoken human voices whose language is completely different from the one used by native speakers of Spanish, a first research gap is found in this sense. When assuming a pure different language, what is being considered is the fact that there are not similarities between the unfamiliar words and the words used in the native language (Surayyo, 2022). Yeh et al. (2011) and Hsiao et al. (2012) used tones of voices assuring that an explicit auditory recognition of the type of voice would emerge, rather than an understanding of a sort of monologue. Similarly, Smith et al. (2007) decided to use pure tones that match with the genders implicated in the androgynous image presented. For their part, Marroquín-Ciendúa et al. (2020) and Rodríguez-Martínez et al. (2021) both used monologues spoken in French, controlling that the participants (native speakers of Spanish) did not understand the words and their meaning. Nevertheless, it has been stated that French and Spanish have some commonalities (Ganfi *et al*., 2023), including phonological characterizations such as syllable structure and morphological and syntactic conditioning (Ingram & Babatsouli, 2024). Due to these commonalities, the aural reception implied when Spanish native speakers (who do not know anything about French) are listening to people talking in

French is not fully constrained. A complete constraint for a listener to understand a new language is the variance in the symbols, their prosody, syllable structure, phonotactics and grammatical interactions (Ingram & Babatsouli, 2024). In this respect, aural reception of Chinese, for instance, is one of the most difficulties that western foreigners who do not speak that language find when trying to learn or understand words and contents (Gabbianelli & Formica, 2017). Unless native speakers of Spanish have learnt some Chinese, it is very unlikely that they can understand a sequence of words spoken in the eastern language (Gabbianelli & Formica, 2017). The research gap that arises in relation to the possible bias concerning the understanding of some words spoken in French should be solved, if the effect of semantic congruency based on bistable perception paradigms was going to be observed in native speakers of Spanish in Latin America.

On the other hand, the version of the bistable image used is a variable, which can exert an effect on the observation of the semantic congruency effect. When reviewing the original ambiguous image of *My girlfriend or my mother-in-law,* it is different in relation to the ones used in the experiments and studies previously reported. Although some variations of the image have been made for psychological purposes, there is no one study in which the original image *My girlfriend or my mother-in-law* (the version first proposed in 1915) have been used. In Table 1, the images used in the studies previously described are presented, as well as the original version of the ambiguous image *My girlfriend or my mother-in-law.*

Table 1. Comparison of the bistable images used so as to observe the semantic congruency effect

				
Androgynous image	Androgynous image	Boring's image (fully contrasted)	Boring's image (delineated)	Boring's image (original version)
Smith et al. (2007).	Rodríguez-Martínez and Sojo (2022).	Yeh et al. (2011); Hsiao et al. (2012).	Marroquín-Ciendúa et al. (2020); Rodríguez-Martínez et al. (2021).	The present study.

Note: Authors (researchers) can be seen in the third row. In the second one, below the images, the type of images is described, where there are two androgynous images used and three different versions of the Boring's image *My girlfriend or my mother-in-law*. On the right, the very first original version of this bistable image is shown

1.4. Research question, aims and objectives

After having reviewed the state of art in relation to the assessment made on the semantic congruency effect by means of providing crossmodal sensory stimulation and taking into account paradigms based on visual bistable perception, the research question that arose was arranged this way: would there be a significant top-down modulating effect on the perception of the bistable image *My girlfriend or my mother-in-law* when providing monologues spoken by two different modulating female voices in an utterly incomprehensible language? Is it possible to observe the semantic congruency effect if native speakers of Spanish listen to female voices that produce words in an incomprehensible foreign language whilst perceiving the original version of the bistable image *My girlfriend or my mother-in-law*?

Regarding the research questions outlined before, the present study was proposed to achieve the aim of observing whether the effect of semantic congruency occurs, providing that auditory semantic-congruent information (monologues in an incomprehensible language) is heard by native speakers of Spanish whilst recognizing the two percepts of the first version of the bistable image *My girlfriend or my mother-in-law*. As has previously been mentioned, these two visual percepts have congruent semantic meanings in relation to the semantic load of each auditory stimulus, as several studies that used the mentioned bistable image also considered (Yeh *et al*., 2011; Hsiao *et al*., 2012; Marroquín-Ciendúa *et al*., 2020; Rodríguez-Martínez *et al*., 2021). As for the present study, it was assumed that there was going to be an influence of the semantic content provided by the auditory stimuli on the perception of an ambiguous visual stimulus, taking into account that both, audios and the image itself, would have a related semantic content, as mentioned before. This way, the hypothesis underlying the present study was that the visual semantically-congruent percept (congruent with the auditory stimuli) should be dominant for a larger duration than the other, during an experimental visual task regarding the observation of the bistable visual stimulus. It was hypothesized that the reported visual percepts that were congruent with the semantic content of the auditory stimulus would be longer than the percepts unrelated to the sound. If so, the semantic congruency effect would emerge as a consequence of a top-down modulating effect (Marroquín-Ciendúa *et al*., 2020; Rodríguez-Martínez *et al*., 2021).

2. Participants, materials and methods

Thirty-two healthy volunteers participated in this study (50% = female; 50% = male; age ranging between 18 and 25 years old; mean age, *M* = 21.656; *SD*= 2.009). All of them were naive of the purpose of the experiment and all had normal or corrected-to-normal vision and also normal hearing. They all were Spanish native speakers. In order to select the participants, several inclusion criteria were considered as follows: firstly, they had to report not having had clinical records regarding cerebral damages and cognitive disorders. Secondly, all the participants had to have perfect vision. If not, a vision corrected by contact lenses would be the accepted condition to be selected. On the other hand, the participants had to be young people, not being over 28 years of age, so that their cognitive and perceptual performances were in a good level, according to psychological statements put forward in relation to human cognitive development (Eagleman, 2015). Given the fact that the experimental tasks were done in Colombia, where to be of legal age people must be 18 years old, the accepted age range was 18-28.

An experiment was conducted in a dimly-lit experimental chamber where variables like external noise and temperature were completely controlled. Participants had to view the bistable image *My wife or my mother-in-law* while listening to the tone of voices of two women, similar to previous studies (Hsiao *et al*., 2012; Rodríguez-Martínez *et al*., 2021). Each participant was able to view the visual bistable design twice, at a viewing distance of 60 cm., just like other researchers did (Marroquín-Ciendúa *et al*., 2020; Rodríguez-Martínez, 2024). A red fixation point was presented centered located, at the bottom area of the image (Figure 2). This was done in order to control the bottom-up factor involved in the perception of the bistable image at the beginning of the observation. Thus, the fixation point was presented in such a way that it was unable to favor the percept of the elderly woman as well as that of the young woman, as stated in previous related studies (Gale & Findlay, 1983; Hsiao *et al*., 2012; Rodríguez-Martínez *et al*., 2021).

On the first occasion, when participants viewed the image, they were able to hear the tone of the voice of an elderly woman while viewing the image. The second time, they viewed the bistable image whilst the tone of a voice of a young woman was being played. Both voices were presented randomly, so as to obtain counterbalanced exposure. These auditory stimuli were presented at the same volume.

The audios themselves lacked the spectral components that characterize human vocalization. They were created based on tones of Chinese voices taken from Chinese native speakers (the sounds were recorded in a professional recording studio). Chinese sounds were the ones used so as to be sure that the participants would not understand any acoustic syllable as part of words (all the participants reported that they did not understand Chinese language). In order to provide this auditory stimulation, *Creative HN-900* headphones were used. Taking into account the owner's manual, these headphones can reduce environmental noise up to 18 dB. The selection of this device was based on the criterium of being closed-ear headphones, in order to control external noise that could affect the task. As Hsiao et al. (2012) did, the auditory stimuli were presented at 52 dB SPL, approximately. The visual stimulus was displayed on the monitor of an eye-tracking device, with a refresh rate of 120 Hz. (remote eye-tracker, reference Tobii®, T-120). As has been mentioned, one of the tones corresponded to one of the possible bistable image's percepts. The other was linked to the second percept. Participants pressed keys of a keyboard so as to report which percept was being perceived by them during image presentation (each exposure lasted 24 seconds). Thus, each key was marked with a specific letter and color in such a way that participants were able to report the corresponding percept.

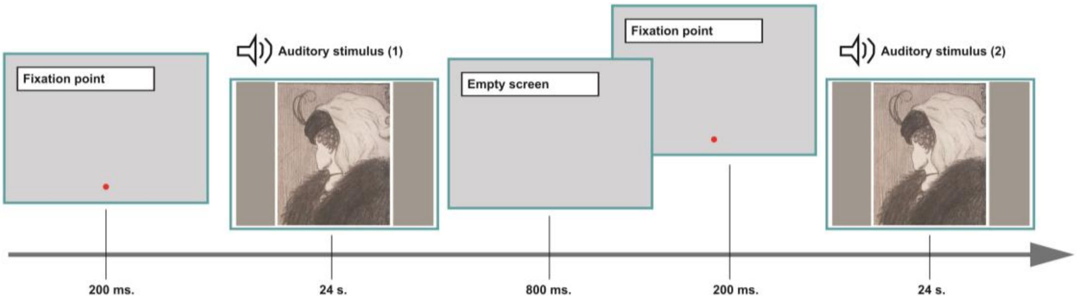


Figure 2. The sequence of the experimental task. The time-line presented here (from left to right) shows the setup used by means of the software Tobbi® Studio (related to the eye-tracker device, Tobii T-120) At the beginning, a fixation point appeared (200 milliseconds). After that, the original version of *My girlfriend or my mother-in-law* was presented (for 24 seconds) with a simultaneous auditory stimulus (the voice of an old woman saying things in Chinese). Before a new exhibition of the fixation point, there was a gap that lasted 800 milliseconds. To finish the task, the same bistable image appeared (for 24 seconds), this time with a simultaneous display of a second voice (the voice of a young lady) saying things in Chinese. The order of the auditory stimuli was randomly shifted by programming it through the software Tobbi® Studio
Source: Own design

During each image presentation, auditory modulators were displayed. In other words, each tone of voice lasted 24 seconds. These auditory stimuli were exposed concurrently with the image. The version of the image that was used was different from the one that was displayed in the previous studies conducted by Gale and Findlay (1983), Hsiao et al. (2012), Marroquín-Ciendúa et al. (2020) and also by Rodríguez-Martínez et al. (2021). This time, it was used the original bistable visual design of the image *My girlfriend or my mother-in-law*, as can be seen within the Figure 2, on the right. In other words, it was used one of the oldest known form of this image (the version first released in 1915). In Figure 2, the procedure utilized in the present study is shown and also the version of the bistable image used, as mentioned before. Informed consent was obtained from all subjects involved in the research project. The present study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of the University Jorge Tadeo Lozano.

3. Results

First of all, the Shapiro-Wilk test was done so as to determine that the data was normally distributed. It was found that data concerning durations for correspondent visual percepts with auditory stimulus was normal: as for time for old-woman percept, *W* = .933, *p* = .47; as for the duration of young-woman percept, *W* = .909, *p* = .10. The duration of each visual percept congruent with the audio was the estimated measurement for observer’s visual performance. Each time that a participant reported a visual percept which was semantically congruent to the modulating audio, the time was recorded from that moment until a report was made for the incongruent percept. Thus, durations with congruent and incongruent percepts with audio modulators were obtained.

The results show that there is a significant difference between the durations of the congruent percepts in relation to the time concerning incongruous percepts. As can be seen in table 2, while listening to the tone of the young female voice, the time concerning young woman visual percept is longer in comparison with the duration of the old woman percept (for young woman, *M* = 13051.36, *S.D.* = 7452.503; for old woman, *M* = 698.466, *S.D.* = 2384.66; *t* (*31*) = -8.648, *p* < .001). On the other hand, the time measured with respect to old woman percept is longer than the time for the young woman percept when analyzing the reports corresponding the tone of the old female voice (for old woman, *M* = 12434,561, *S.D.* = 7363.836; for young woman, *M* = 2518,477, *S.D.* = 5354.865; *t* (*31*) = 4.964, *p* < .001). Units of time are expressed in millisecods.

Table 2. Durations for each reported visual percept

Modulating audio	Reported percept	*M*	*S.D.*
Young woman tone	Old woman	698.466	2384.66
	Young woman	13051.336	7452.503
Old woman tone	Old woman	12434.561	7363.836
	Young woman	2518.477	5354.865

As can be seen in Figure 3, the time of the percept corresponding to the semantically-congruent auditory stimulus is longer than the duration of the visual percept that is not congruent with the non-consistent audio. Thus, the duration of the percept 'young woman' is longer while listening to the tone of voice related to a young voice and in turn, the time for the percept 'old percept' is longer when what is being heard is the tone that corresponds to an old voice.

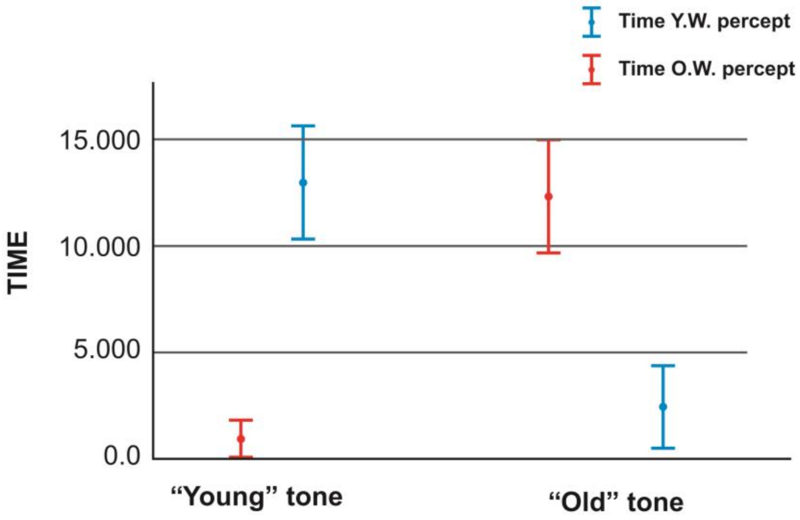


Figure 3. Time differences between visual percepts considering the two modulating auditory stimuli
Source: Own design

According to the results, there is an effect depending on the audio displayed simultaneously with the image. In other words, as is shown in Figure 4, the duration of the visual percept 'old woman' decreases when the cross-modal stimulation implies the non-corresponding tone of voice. As far as the young-woman percept is concerned, it is shown how the duration of it has an increase if the voice used as an auditory stimulus is consistent with that interpretation.

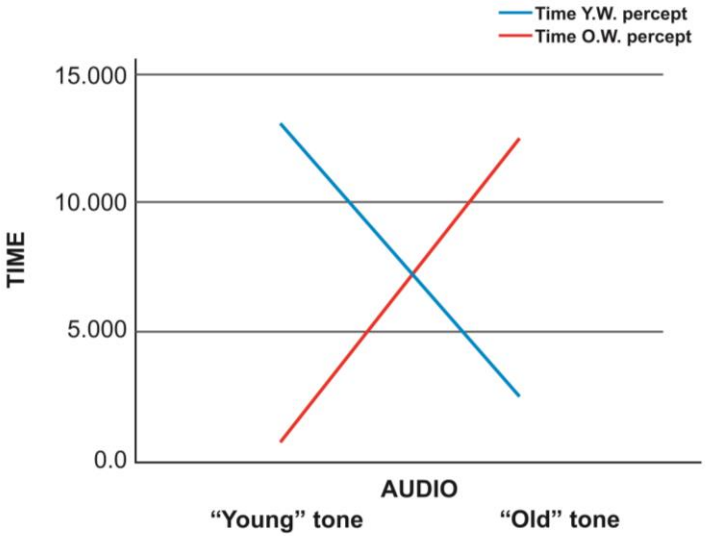


Figure 4. Time-duration variations of each visual percept considering modulating audios used
Source: Own design

When the participants viewed twice the bistable image, each time with a different modulating auditory stimulus (old woman voice and young woman voice), there was a predominant visual percept. This predominance was established, as mentioned before, by the difference between the durations of the perception of the two possible percepts. In this regard, there were durations for each percept (the old woman and the young lady) in the two auditory modulating conditions.

When considering all the durations for the reported image identified, the difference in time in favor of the correspondences (the occurrence of the semantic congruency effect), most of the participants recognized much more the visual percept related to the congruent modulating auditory stimulus. Just as can be seen in table number 3, after having subtracted the time of the modulated percept from the time of the unmodulated one, most of the results came to be positive, which means that the duration of the modulated visual percept was larger than the duration of the other visual percept. This way, the effect of semantic congruency was vindicated, as was mentioned previously at the beginning of this section. It happened in both cases (perception with the voice of the old woman and perception with the voice of a young lady), but the results were better for the case of the modulating process that implied the voice of the young lady (Figure 5). Three cases (in relation to the trial concerning the audio of the young woman) were found in which there were no a recognition of either of the visual percepts. As for the other trial (presentation of the bistable image with the audio of the old woman), just one participant reported no durations for any of the two visual percepts. Although their values were equivalent to zero (0), it was decided to take them into account, due to the fact that, within the context of bistable perception, it is possible that observers are not able to identify either of the percepts of an ambiguous image (Hsiao *et al*., 2012). Besides, after having reviewed if the participants who did not recognize the visual percepts had or not a condition that could have affected the recording of data and after having collated these specific records in order to find any bias concerning the recording itself, there were no anomalies, so, what was decided was to include these values within the statistical analyzes.

Table 3. Positive and negative semantic congruences observed based on differences between the reports concerning the visual percepts for each modulating audio

Difference favorable to the percept congruent (young lady voice)	Occurrence of the semantic congruency effect for the Y.W. percept.	Difference favorable to the percept congruent (old woman voice)	Occurrence of the semantic congruency effect for the O.W. percept.
12031,603	Positive	-22324,2	Negative
21032,492	Positive	14694,13	Positive
4983,134	Positive	12816,15	Positive
20515,846	Positive	-6384,85	Negative
17432,636	Positive	-5547,78	Negative
11616,202	Positive	13016,15	Positive
183,326	Positive	17699,29	Positive
16649,334	Positive	-13294,7	Negative
22332,440	Positive	-8349,67	Negative
0,000	No difference	12766,16	Positive
-1216,618	Negative	14528,42	Positive

5566,444	Positive	9046,38	Positive
21815,794	Positive	4883,14	Positive
16216,018	Positive	449,98	Positive
22982,414	Positive	12582,83	Positive
1387,778	Positive	9132,97	Positive
10599,576	Positive	0	No difference
0,000	Positive	2265,76	Positive
21715,798	Positive	18249,27	Positive
15532,712	Positive	17682,63	Positive
20732,504	Positive	12999,48	Positive
11466,208	Positive	15327,2	Positive
0,000	No difference	10716,24	Positive
20449,182	Positive	19815,87	Positive
0,000	No difference	22815,75	Positive
11066,224	Positive	21032,49	Positive
18482,594	Positive	16982,65	Positive
10599,576	Positive	23765,72	Positive
9616,282	Positive	21149,15	Positive
17499,300	Positive	9299,63	Positive
17326,400	Positive	17315,97	Positive
16676,660	Positive	22182,45	Positive

Note: The term ‘positive’ is used for the positive difference in time in favor of the congruent visual percepts (congruent with their correspondent audio). The term ‘negative’ is linked to a negative difference, that is to say, when the visual percept identified was not correspondent with the semantic load of the auditory stimulus. It was necessary to use the label ‘no difference’ for the cases in which there was no a difference between the durations reported in relation to the identification of the two possible visual percepts

As mentioned before, when comparing the means of the differences between the expected favorable durations for each modulating audio (Figure 5), what is found is that there was a better result (in terms of the occurrence of the semantic congruency effect) for the female young lady voice: mean of the positive difference in favor of the congruent perception with the young voice, *M*= 12.352, 87, *S.D.*= 8080,73; mean of the positive difference in favor of the congruent perception with the old woman voice, *M*= 9916, 08, *S.D.*= 11300,11. Units of time are expressed in milliseconds.

On the other hand, when organizing visually the reports for both cases of auditory modulation, it is noticeable how the occurrence of the semantic congruency effect was better for the young female voice stimulus, as previously mentioned (Figure 6). More negative differences (visual reports incongruent with the semantic load of the audio) were found when participants viewed the bistable image *My girlfriend or my mother-in-law*, listening, at the same time, to the voice of the old woman (Table 3).

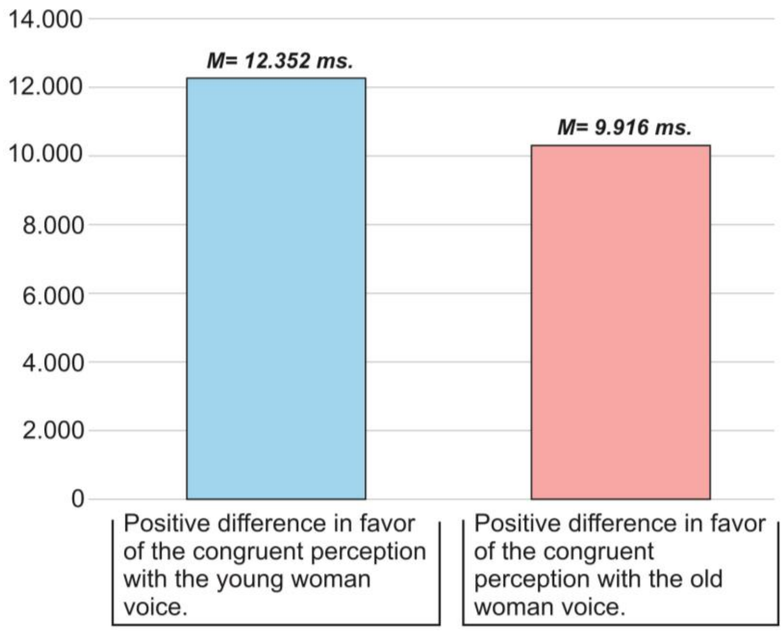


Figure 5. Differences in favor of the durations of the congruent visual percept, considering each modulating audio

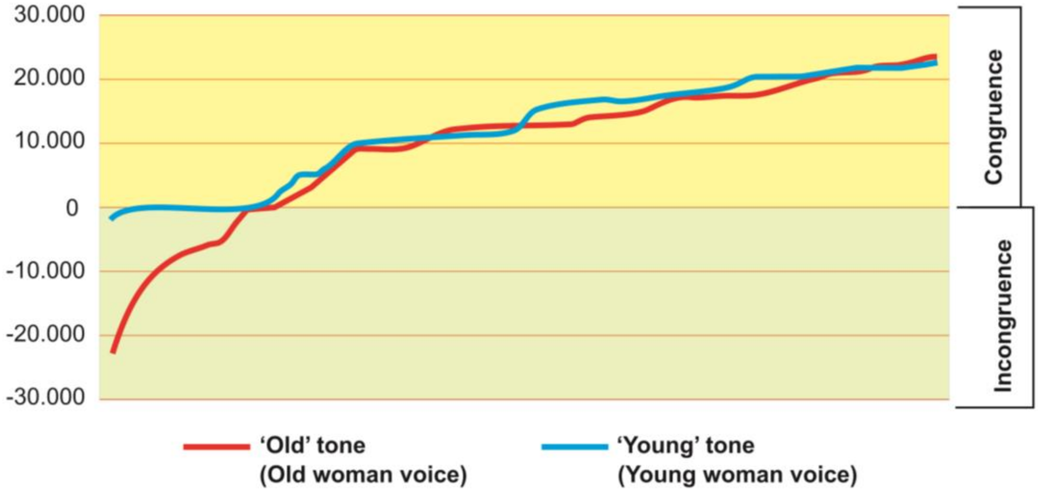


Figure 6. Visual comparison between the differences found in favor of the durations of the congruent visual percept, considering both auditory modulating stimuli. Positive values implied congruent visual percepts reported (zone marked in yellow). Units of time are expressed in milliseconds

4. Discussion

According to the results, the modulating auditory stimuli influenced significantly the interpretation of the ambiguous image. A significant difference was found in favor of the percepts which correspond to each semantically congruent tone of voice. As the statistical tests showed, for both congruent interpretations with each auditory stimulus (young woman and old woman), evidence was found to support the hypothesis that the use of tones of voice affects the decoding of bistable visual stimuli, just as found in previous studies (Hsiao *et al*., 2012; Marroquín-Ciendúa *et al*., 2020; Rodríguez & Sojo, 2022; Smith *et al*., 2007). As was stated when defining the aims of the present study, what had been assumed was that there was going to be an influence of the semantic content provided by the auditory stimuli on the perception of an ambiguous visual stimulus, taking into account that both, audios and the version of the image used, would have a related semantic content. When mapping the findings obtained, relevant issues contribute to the confirmation of the hypotheses proposed: an expected result was that the visual semantically-congruent percept (congruent with the auditory stimuli) should be dominant for a larger duration than the other, a fact that was confirmed when analysing the durations for congruent and incongruent visual percepts reported. As a matter of fact, and in line with the hypothesis that proposed that the reported visual percepts that were congruent with the semantic content of the auditory stimulus would be longer than the percepts unrelated to the sound, the differences found between durations for congruent audio-visual integrations and durations of non-congruent reports were positive. This way, the occurrence of the congruent crossmodal integration was observed, vindicating the semantic congruency effect, where the identification of the figures (old woman and young lady) must have been influenced as a consequence of a top-down modulating effect, as also found in several studies (Hsiao *et al*., 2012; Marroquín-Ciendúa *et al*., 2020; Rodríguez-Martínez *et al*., 2021; Yeh *et al*., 2011).

Several issues have to be taken into consideration: firstly, the fact that both stimuli (visual and auditory) were exposed simultaneously, which implied a crossmodal sensory stimulation (Chen *et al*., 2011; Hsiao *et al*., 2012). When assessing the semantic congruency effect, the use of crossmodal stimuli is common because the observer is supposed to receive external information by means of using two different channels so that each one provides information that is likely to be linked in terms of semantic content (Delong & Noppeney, 2021). What is implied here is the fact that some information provided by a channel can affect the perception of the information given by a second sensory channel (Hartcher-O'Brien *et al*., 2017). If the channels that were selected are visual and auditory, the phenomenon that can emerge in relation to the semantic congruency effect is the so-called *unity assumption*, provided the auditory information is integrated to the visual stimulus (Feenders & Klump, 2018). It would imply the occurrence of multisensory integration, a phenomenon that is likely to happen the closer that the presentation of stimuli from different sensory modalities is done in time (Spence, 2011; Spence, 2013; Spence & Squire 2003). In this regard, semantic-based audio-visual integration requires, not only semantic similarities, but also temporal proximity or simultaneity (Cox *et al*., 2015).

On the other hand, semantic congruency often refers to stimulating observers by using pairs of auditory and visual stimuli that convey a similar semantic content (Hsiao *et al*., 2012; Marroquín-Ciendúa *et al*., 2020). In this regard, semantic congruency is commonly evaluated by measuring outputs while presenting matching and mismatching images and sounds (Di Stefano & Spence, 2023; Spence, 2011). Some examples of assessing semantic congruency effects have used bistable perception paradigms due to the unique characteristics of bistable or ambiguous images (Hsiao *et al*., 2012; Kesoglou & Mikellidou, 2024; Rodríguez-Martínez *et al*., 2021; Smith *et al*., 2007; Zeljko *et al*., 2022).

Considering the results found in the present study, crossmodal semantic correspondance can imply the overlap of the two dimensions of the stimuli based on the similarity of what is conveyed semantically speaking by the stimuli themselves (Di

Stefano & Spence, 2023; Zeljko *et al*., 2022). It means that semantic content is, necessarily, implied, given the fact that the correspondance is marked by the semantic charge of each stimulus. The correspondance can emerge when it can be assumed that both, the auditory stimulus and the visual one, constitute a perceptual unity, also called unity assumption (Vatakis & Spence, 2008), as mentioned before. This audio-visual perception has been stated as a type of crossmodal integration where auditory and visual modalities converge (Delong & Noppeney, 2021; Spence, 2011; Smith *et al*., 2007), so as to create a single coherent integrated representation (Lalanne & Lorenceau, 2004). The integration may occur based on top-down modulating effects (Rodríguez-Martínez *et al*., 2021). Besides, the semantic congruence effect marks the occurrence of the perceptual integration (Chen *et al*., 2011).

Previous findings obtained in several studies regarding semantic modulation of visual perception, which underpinned the visual task based on the bistable perception paradigm, indicated that, if it is possible to use a tone of voice as a modulator (Hsiao *et al*., 2012; Smith *et al*., 2007; Rodríguez-Martínez & Sojo, 2022), the semantic congruence effect can occur. Nevertheless, bottom-up modulating factors could impact the final perception, generating an effect on the semantic congruence reading. Given that the present study does not take into account ocular fixation areas observed during verbal reports referring to the two possible percepts of the visual stimulus, it is likely that the bottom-up modulating factors associated with ocular fixation areas played a decisive role in the occurrence of the semantic congruence effect. This is a limitation of the present study.

On the other hand, in the dominion of bistable perception, it has been assumed that conditions, particular to the viewer and alluding to their learning and also to their visual tracking capacities, may be factors in the way that perceptual reversals occur, that is to say, the changes between one percept to the other one (Novicky *et al*., 2024; Rodríguez-Martínez *et al*., 2021). Thus, top-down and bottom-up processes can be implicated in the observation of a bistable image. In the moment in which an auditory modulator with semantic content is presented as equivalent to the semantic contents which belong to the possible percepts of the ambiguous image, the viewer can search for something that can be associated with the audio. This searching would involve ocular movements control, that is to say, voluntary visual tracking trajectories (van Dam & van Ee, 2006).

Gale and Findlay’s study (1983) shows the relation that exists between areas observed in the image used in the present study (*My girlfriend or my mother-in-law*). That study was vindicated by Rodríguez-Martínez et al. (2021) and also by Rodríguez-Martínez (2025). However, this time what was used was the image released in 1915 instead of the simplified version used by Gale and Findlay in 1983. Despite the fact that top-down perceptual processing appears in the review of bistable images such as *My girlfriend or my mother-in-law*, it is necessary to consider that characteristics belonging to the visual stimulus (lines, contours, contrasts, textures) exert an impact on perception, so that certain interaction between bottom-up and top-down processes can also emerge and in turn, can affect perceptual outcomes. As such, studies regarding semantic modulation of bistable perception should consider both factors, or, if desired, should control them during experimental tasks. Although the version of the bistable image *My girlfriend or my-mother-in-law* used in the present study was proved to be useful to measure the occurrence of audio-visual perceptual integration, it is not clear what visual modulating areas could have exerted an influence on the perceptual outcomes, a fact that also constitutes a limitation for the present study.

For their part, when reviewing the usefulness of the monologues spoken in Chinese (as top-down modulating stimuli), their effect came to be effective, showing that tones of voices selected to provide specific semantic loads can operate as perceptual modulators, which, in turn, constitutes a way to observe the semantic congruency effect by means of bistable perception paradigms, just as other researchers proved (Smith *et al*., 2007; Rodríguez-Martínez & Sojo, 2022; Zeljko *et al*., 2022). It should be taken into account that voices themselves were the auditory stimuli, rather than the content of words. This is not a minor factor if considering that monologues imply the combination of words (Abassi & Zatorre, 2024). Adult people expect that there are going to be words while listening to human voices articulating symbols defined to convey messages (Keshishian *et al*., 2023). In this spirit, the fact that the participants who took part in the visual tasks of the present study knew that some messages were being conveyed in Chinese (regardless of not being able to understand any content) could imply a perceptual interference. As a matter of fact, it has been demonstrated that languages can affect perception in the way that top-down perceptual modulating processes emerge when sensory processing inputs turned to be decoded with an influence provided by the cognitive reasoning that imply a mental use of lexical codes (Slivac & Flecken, 2023). In this regard, if what is wanted is to observe the semantic congruency effect by means of using human voices, there are, definitely, two ways to do so, that is to say, by selecting words whose meaning could influence (modulate) the recognition of an intrinsic visual percept that is contained in a bistable image and also by basing it on semantic cues that characteristics of human voices can provide, like their mood, tone, intensity, among others (Zhao *et al*., 2024). But possible interferences should be considered previously. Even by using neuroscientific instruments (such as EEG, fMRI, etc.), it is possible to understand the integration of the processes that are implicated when observing multisensory audio-visual perception, when visual sensory inputs are perceived as a unity with correspondent auditory stimuli (Marly *et al*., 2023).

To close this section, it should be taken into consideration that when using audio-visual crossmodal stimulation, the ocular fixations can be a crucial variable, as has been mentioned. Nevertheless, the semantic congruency effect can be assessed just by means of considering visual percepts in relation to different auditory stimuli semantically-congruent with what can be visually perceived. As a matter of fact, the results found here lend support to the occurrence of modulating cross-modal perceptual processes, which, in turn, lead to draw the conclusion that visual perception of bistable images can be influenced by means of top-down perceptual mechanisms. The semantic congruency effect emerges when stimulating with cross-modal semantically congruent audiovisual stimuli, as stated by Hsiao et al. (2012). These findings should be taken into consideration while trying to understand the complexity of the semantic congruency effect. When graphic designers conceive a bistable image, they have to take into account that its perception can be influenced by means of semantic top-down perceptual modulating stimuli.

5. Recommendations for future research

The present study just focused on observing the modulating effect of tones of voice, where each tone corresponded to each possible percept of the bistable image, but without comparing or considering bottom-up modulating factors (although a fixation point was considered, previous to the presentation of the bistable image so as to direct the first glimpse to a point located in specific coordinates of the image which do not favor any of the percepts, according to the findings of preliminary research projects). With this in mind, future research projects could analyze the areas in which ocular fixations are made when recognizing visual percepts that are intended to be integrated to auditory stimulation. Similar to what Hsiao et al. (2012) and Rodríguez-Martínez et al. (2021) did, in the sense that participant's ocular fixations were estimated as a co-variable within the context of the implication of bottom-up modulating processes regarding the areas of the bistable figure that might exert an influence on its perception. In order to do so, another factor that should be considered is the quality of the sample concerning oculomotor activity. In this regard, eye-tracker devices that have 1000 Hz., 1200 Hz, or, if possible 2000 Hz., would be advisable if using one or more bistable figures with which observe the effect of semantic congruency. As has been stated, the sample quality of eye-trackers (the oculomotor activity that can be recorded per second) is an issue that must be taken into account when trying to study areas of interest (AOIs) of a visual stimulus (Rodríguez-Martínez, 2025). The more hertz the eye-tracker device has, the higher the quality of the ocular fixations recorded (Leube & Rifai, 2017).

It has to be said here that, apart from ocular fixations, attention plays a leading role in modulating visual perceptual processes (Gale & Findlay, 1983; Hsiao *et al*., 2012; Rodríguez-Martínez, 2024). In this regard and considering that the present study did not analyze ocular fixation areas gazed during the reports referring to the two possible percepts of the ambiguous image used, it is likely that, as mentioned before, some bottom-up modulating factors associated with ocular fixation points could have played an important role in the emergence of the semantic congruency effect, understanding that attentional modulating issues converge with bottom-up perceptual modulating processes (Rodríguez & Castillo, 2018).

On the other hand, new experiments can be conducted, attempting to establish differences depending on the version of the bistable visual stimulus. It is likely to happen that several differences can be discovered when comparing three or four versions of the same ambiguous image. Besides, factorial experiments could be done in such a way that different categories of variables are considered, such as version of the image, kind of bistable image (in perspective reversals, in meaning-content reversals and in figure-ground reversals), attentional fixation points that condition (or control) bottom-up modulating aspects, among others. For their part, if controlling types of auditory stimuli that can be perceptually integrated with a visual percept, another possible important influence would be detected. Although several studies have used pure tones or acoustic stimuli that do not imply human sounds (Smith *et al*., 2007; Hsiao *et al*., 2012; Zeljko *et al*., 2022), the human voice can operate as a top-down modulating factor, as well as words clearly understood by participants who take part in an experiment (Balcetis & Dale, 2007; Feist & Gentner, 2007; Goolkasian & Woodberry, 2010; Silva & Bellini-Leite, 2020). In this spirit, factorial experiments can be arranged so as to observe the effect of different types of auditory modulating stimuli. By doing so and broadening the number of participants, the observation of the semantic congruency effect could be explained deeply, even contemplating transcultural studies through which it is possible to find differences due to cultural factors.

6. Conclusions

Auditory stimulation provides information that can exert an influence on the perception of one of the oldest versions of the bistable image *My girlfriend or my mother-in-law* (the version released in 1915). When using pure tones of voice as semantic modulators, the semantic congruency effect can emerge so that a modulating perceptual top-down process is implied, affecting the perception of the ambiguous image in question. An incomprehensible spoken language can operate as a top-down modulating factor by which to modulate bistable perception. Subsequently, it is concluded that tones of voice used as modulators can bring about a disambiguation of the image in terms of its interpretation whereby the visual percepts that are mostly recognized are the ones that are consistent with semantically-congruent auditory stimuli. This fact vindicates the effect of semantic congruency, an effect that should be considered when designing visual bistable images.

References

Abassi, E., Zatorre, R. (2024). Influence of social and semantic contexts in processing speech in noise. *bioRxiv*, 2024-01. https://doi.org/10.1101/2024.01.10.575068

Baker, D.H., Karapanagiotidis, T., Coggan, D.D., Wailes-Newson, K. & Smallwood, J. (2015). Brain networks underlying bistable perception. *NeuroImage*, 119, 229-234. https://doi.org/10.1016/j.neuroimage.2015.06.053

Balcetis, E., Dale, R. (2007). Conceptual set as top-down constraint on visual object identification. *Perception*, 36, 581-595. https://doi.org/10.1068/p5678

Barrera, M., Calderón, L. (2013). Notes for supporting an epistemological neuropsychology: Contributions from three perspectives. *International Journal of Psychological Research*, *6*(2), 107-118. https://doi.org.10.21500/20112084.692

Boring, E. (1930). A new ambiguous figure. *American Journal of Psychology*, 42, 444-445. https://doi.org/10.2307/1415447

Brouwer, G.J., van Ee, R. (2006). Endogenous influences on perceptual bistability depend on exogenous stimulus characteristics. *Vision Research*, 46, 3393-3402. https://doi.org/10.1016/j.visres.2006.03.016

Carbon, C.C. (2014). Understanding human perception by human-made illusions. *Frontiers in Human Neuroscience*, 8, 566. https://doi.org/10.3389/fnhum.2014.00566

Chen, Y.C., Yeh, S.L. & Spence, C. (2011). Crossmodal constraints on human perceptual awareness: Auditory semantic modulation of binocular rivalry. *Frontiers in Psychology*, 2, 212. https://doi.org/10.3389/fpsyg.2011.00212

Clément, G., Demel, M. (2012). Perceptual reversal of bi-stable figures in microgravity and hypergravity during parabolic flight. *Neuroscience Letters*, 507, 143-146. https://doi.org/10.1016/j.neulet.2011.12.006

Cox, D., Hong, S.W. (2015). Semantic-based crossmodal processing during visual suppression. *Frontiers in Psychology*, 6, 722. https://doi.org/10.3389/fpsyg.2015.00722

Delong, P., Noppeney, U. (2021). Semantic and spatial congruency mould audiovisual integration depending on perceptual awareness. *Scientific Reports*, *11*(1), 1-14. https://doi.org/10.1038/s41598-021-90183-w

Devia, C., Concha-Miranda, M. & Rodríguez, E. (2022). Bi-stable perception: Self-coordinating brain regions to make-up the mind. *Frontiers in Neuroscience*, 15, 805690. https://doi.org/10.3389/fnins.2021.805690

Di Stefano, N., Spence, C. (2023). Perceptual similarity: Insights from crossmodal correspondences. *Review of Philosophy and Psychology*, 1-30. https://doi.org/10.1007/s13164-023-00692-y

Diveica, V., Muraki, E.J., Binney, R.J. & Pexman, P.M. (2024). Mapping semantic space: Exploring the higher-order structure of word meaning. *Cognition*, 248, 105794. https://doi.org/10.1016/j.cognition.2024.105794

Eagleman, D. (2015). *The Brain: The Story of You*. Great Britain by Canongate Books Ltd.

Feenders, G., Klump, G.M. (2018). Violation of the unity assumption disrupts temporal ventriloquism effect in starlings. *Frontiers in Psychology*, 9, 1386. https://doi.org/10.3389/fpsyg.2018.01386

Feist, M., Gentner, D. (2007). Spatial language influences memory for spatial scenes. *Memory and Cognition*, 35, 283-296. https://doi.org/10.3758/BF03193449

Frassinetti, F., Bolognini, N. & Làdavas, E. (2002). Enhancement of visual perception by crossmodal visuo-auditory interaction. *Experimental Brain Research*, *147*(3), 332-343. https://doi.org/10.1007/s00221-002-1262-y

Gabbianelli, G., Formica, A. (2017). Difficulties and expectations of first level chinese second language learners. In *Explorations into Chinese as a Second Language*. *Educational Linguistics*, 31, 183-206. https://doi.org/10.1007/978-3-319-54027-6_8

Gale, A., Findlay, J. (1983). Eye-movement patterns in viewing ambiguous figures. In *Eye Movements and Psychological Functions: International Views*, 145-168. https://doi.org/10.4324/9781003165538

Ganfi, V., Piunno, V. & Mereu, L. (2023). Body part metaphors in phraseological expressions: A comparative survey of Italian, Spanish, French and English. *Languages in Contrast*, *23*(1), 1-33. https://doi.org/10.1075/lic.21006.gan

García-Pérez, M. (1989). Visual inhomogeneity and eye movements in multistable perception. *Perception & Psychophysics,* 46, 397-400. https://doi.org/10.3758/BF03204995

García-Pérez, M.A. (1992). Eye movements and perceptual multistability. *Advances in Psychology*, 88, 73-109. https://doi.org/10.1016/S0166-4115(08)61743-4

Gijs, B., van Ee, R. (2006). Endogenous influences on perceptual bistability depend on exogenous stimulus characteristics. *Visual Research*, 46, 3393-3402. https://doi.org/10.1016/j.visres.2006.03.016

Goolkasian, P., Woodberry, C. (2010). Priming effects with ambiguous figures. *Attention, Perception & Psychophysics*, 72, 168-178. https://doi.org/10.3758/APP.72.1.168

Hartcher-O'Brien, J., Soto-Faraco, S. & Adam, R. (2017). A matter of bottom-up or top-down processes: The role of attention in multisensory integration. *Frontiers in Integrative Neuroscience*, 11, 5. https://doi.org/10.3389/fnint.2017.00005

Hsiao, J., Chen, Y., Spence, C. & Yeh, S. (2012). Assessing the effects of audiovisual semantic congruency on the perception of a biestable figure. *Consciousness and Cognition*, 21, 775-787. https://doi.org/10.1016/j.concog.2012.02.001

Ingram, D., Babatsouli, E. (2024). Cross-linguistic phonological acquisition. *The Handbook of Clinical Linguistics*, 2nd edition, 407-419. https://doi.org/10.1002/9781119875949.ch29

Intaité, M., Kovisto, M. & Castelo-Branco, M. (2014). Event-related potential responses to perceptual reversals are modulated by working memory load. *Neuropsychologia*, 56, 428-438. https://doi.org/10.1016/j.neuropsychologia.2014.02.016

Intaité, M., Noreika, V., Šoliūnas, A. & Falter, C.M. (2013). Interaction of bottom-up and top-down processes in the perception of ambiguous figures. *Vision Research*, 89, 24-31. https://doi.org/10.1016/j.visres.2013.06.011

Katsuki, F., Constantinidis, C. (2014). Bottom-up and top-down attention: Different processes and overlapping neural systems. *The Neuroscientist, 20*(5), 509-521. https://doi.org/10.1177/1073858413514136

Keshishian, M., Akkol, S., Herrero, J., Bickel, S., Mehta, A.D. & Mesgarani, N. (2023). Joint, distributed and hierarchically organized encoding of linguistic features in the human auditory cortex. *Nature Human Behaviour*, *7*(5), 740-753. https://doi.org/10.1038/s41562-023-01520-0

Kesoglou, A.M., Mikellidou, K. (2024). The effect of semantic content on the perception of audiovisual movieclips. *bioRxiv*, 2024-01. https://doi.org/10.1101/2024.01.24.576956

Kiefer, M. (2007). Top-down modulation of unconscious' automatic'processes: A gating framework. *Advances in Cognitive Psychology*, *3*(1-2), 289. https://doi.org/10.2478/v10053-008-0032-2

Koivisto, M., Pallaris, C. (2024). Cognitive flexibility moderates the relationship between openness-to-experience and perceptual reversals of Necker cube. *Consciousness and Cognition*, 122, 103698. https://doi.org/10.1016/j.concog.2024.103698

Kornmeier, J., Hein, C.M. & Bach, M. (2009). Multistable perception: When bottom-up and top-down coincide. *Brain and Cognition,* 69, 138-147. https://doi.org/10.1016/j.bandc.2008.06.005

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

Leopold, D.A., Logothetis, N.K. (1999). Multistable phenomena: Changing views in perception. *Trends in Cognitive Sciences*, *3*(7), 254-264. https://doi.org/10.1016/S1364-6613(99)01332-7

Leube, A., Rifai, K. (2017). Sampling rate influences saccade detection in mobile eye tracking of a reading task. *Journal of Eye Movement Research*, *10*(3). http://orcid.org/0000-0002-9182-5408

Marly, A., Yazdjian, A. & Soto-Faraco, S. (2023). The role of conflict processing in multisensory perception: behavioural and electroencephalography evidence. *Philosophical Transactions of the Royal Society B*, *378*(1886), 20220346. https://doi.org/10.1098/rstb.2022.0346

Marroquín-Ciendúa, F., Rodríguez-Martínez, G. & Rodríguez-Celis, H.G. (2020). Modulación de la percepción biestable: Un estudio basado en estimulación multimodal y registros de actividad oculomotora. *Tesis Psicológica*, *15*(1), 106-124. https://doi.org/10.37511/tesis.v15n1a6

Meng, M., Tong, F. (2004). Can attention selectively bias bistable perception? Differences between binocular rivalry and ambiguous figures. *Journal of Vision*, 4, 539 - 551. https://doi.org/10.1167/2.7.447

Novicky, F., Parr, T., Friston, K., Mirza, M.B. & Sajid, N. (2024). Bistable perception, precision and neuromodulation. *Cerebral Cortex*, *34*(1), bhad401. https://doi.org/10.1093/cercor/bhad401

Okazaki, M., Kaneko, Y., Yumoto, M. & Arima, K. (2008). Perceptual change in response to a bistable picture increases neuromagnetic beta-band activities. *Neuroscience Research*, 61, 319-328. https://doi.org/10.1016/j.neures.2008.03.010

Poom, L. (2024). Divergent mechanisms of perceptual reversals in spinning and wobbling structure-from-motion stimuli. *Plos One*, *19*(2), e0297963. https://doi.org/10.1371/journal.pone.0297963

Rantala, J., Salminen, K., Isokoski, P., Nieminen, V., Karjalainen, M., Väliaho, J., … & Surakka, V. (2024). Recall of odorous objects in virtual reality. *Multimodal Technologies and Interaction*, *8*(6), 42. https://doi.org/10.3390/mti8060042

Roberts, K., Jentzsch, I. & Otto, T. U. (2024). Semantic congruency modulates the speed-up of multisensory responses. *Scientific Reports*, *14*(1), 567. https://doi.org/10.1038/s41598-023-50674-4

Robertson, I.H., Mattingley, J.B., Rorden, C. & Driver, J. (1998). Phasic alerting of neglect patients overcomes their spatial deficit in visual awareness. *Nature*, *395*(6698), 169-172. https://doi.org/10.1038/25993

Rodríguez, G., Castillo, H. (2018). Bistable perception: Neural bases and usefulness in psychological research. *International Journal of Psychological Research*, *11*(2), 63-76. https://doi.org/10.21500/20112084.3375

Rodríguez-Martínez, G. (2023). Perceptual reversals and creativity: Is it possible to develop divergent thinking by modulating bistable perception? *Revista de Investigación, Desarrollo e Innovación*, *13*(1), 129-144. https://doi.org/10.19053/20278306.v13.n1.2023.16064

Rodríguez-Martínez, G. (2024). Can ocular fixations modulate the perception of a bistable logo? An eye-tracking study. *Gráfica*, *13*(25), 103-111. https://doi.org/10.5565/rev/grafica.328

Rodríguez-Martínez, G. (2025). Impact of face inversion on eye-tracking data quality: A study using the Tobii T-120. In *International Conference on Applied Informatics*, 68-82. https://doi.org/10.1007/978-3-031-75147-9_5

Rodríguez-Martínez, G., Castillo-Parra, H., Rosa, P.J. & Marroquín-Ciendúa, F. (2021). Ocular fixations modulate audiovisual semantic congruency when standing in an upright position. *Suma Psicológica*, *28*(1), 43-51.https://doi.org/10.14349/sumapsi.2021.v28.n1.6

Rodríguez-Martínez, G., Marroquín-Ciendúa, F., Rosa, P.J. & Castillo-Parra, H. (2022). Perceptual reversals and time-response analyses within the scope of decoding a bistable image. *Interdisciplinaria*, *39*(1), 257-273. https://doi.org/10.16888/interd.2022.39.1.16

Rodríguez-Martínez, G., Sojo, J. (2022). Biestabilidad perceptual en rostros andróginos: Análisis del efecto de congruencia semántica considerando registros de actividad oculomotora. *Revista Ibérica de Sistemas e Tecnologias de Informação*, E52, 133-147. https://www.proquest.com/openview/aa17500f4faeaf59d9d99670f1b5d763/1?pq-origsite=gscholar&cbl=1006393

Sandberg, K., Barnes, G.R., Bahrami, B., Kanai, R., Overgaard, M. & Rees, G. (2014). Distinct MEG correlates of conscious experience, perceptual reversals and stabilization during binocular rivalry. *Neuroimage*, 100, 161-175. https://doi.org/10.1016/j.neuroimage.2014.06.023

Schuman, B., Dellal, S., Prönneke, A., Machold, R. & Rudy, B. (2021). Neocortical layer 1: An elegant solution to top-down and bottom-up integration. *Annual Review of Neuroscience*, 44, 221-252. https://doi.org/10.1146/annurev-neuro-100520-012117

Silva, D.M.R., Bellini-Leite, S.C. (2020). Cross-modal correspondences in sine wave: Speech versus non-speech modes. *Attention, Perception & Psychophysics*, *82*(3), 944-953. https://doi.org/10.3758/s13414-019-01835-z

Slivac, K., Flecken, M. (2023). Linguistic priors for perception. *Topics in Cognitive Science*, *15*(4), 657-661. https://doi.org/10.1111/tops.12672

Smith, E., Grabowecky, M. & Susuki, S. (2007). Auditory-visual crossmodal integration in perception of face gender. *Current Biology*, 17, 1680-1685. https://doi.org/10.1016/j.cub.2007.08.043

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception & Psychophysics*, *73*(4), 971-995. https://doi.org/10.3758/s13414-010-0073-7

Spence, C. (2013). Just how important is spatial coincidence to multisensory integration? Evaluating the spatial rule. *Annals of the New York Academy of Sciences*, *1296*(1), 31-49. https://doi.org/10.1111/nyas.12121

Spence, C., Squire, S. (2003). Multisensory integration: Maintaining the perception of synchrony. *Current Biology*, 13, R519-R521. https://doi.org/10.1016/S0960-9822(03)00445-7

Surayyo, A. (2022). Similarities and differences between first and second language acquisition. *European Scholar Journal*, *3*(2), 100-106. https://scholarzest.com/index.php/esj/article/view/1789

Tarder-Stoll, H., Jayakumar, M., Dimsdale-Zucker, H.R., Günseli, E. & Aly, M. (2020). Dynamic internal states shape memory retrieval. *Neuropsychologia*, 138, 107328. https://doi.org/10.1016/j.neuropsychologia.2019.107328

van Dam, L.C., van Ee, R. (2006). The role of saccades in exerting voluntary control in perceptual and binocular rivalry. *Vision Research*, *46*(6-7), 787-799. https://doi.org/10.1016/j.visres.2005.10.011

Vatakis, A., Spence, C. (2008). Evaluating the influence of the 'unity assumption'on the temporal perception of realistic audiovisual stimuli. *Acta Psychologica*, *127*(1), 12-23. https://doi.org/10.1016/j.actpsy.2006.12.002

Yeh, S., Hsiao, J., Chen, Y. & Spence, C. (2011). Interplay of multisensory processing, attention and consciousness as revealed by bistable figures. *i-Perception*, *2*(8), 910. https://doi.org/10.1068/ic910

Zeljko, M., Grove, P.M. & Kritikos, A. (2022). Implicit expectation modulates multisensory perception. *Attention*, *Perception & Psychophysics*, *84*(*3*), 915-925. https://doi.org/10.3758/s13414-022-02460-z

Zhao, L., Sloggett, S. & Chodroff, E. (2022). Top-down and bottom-up processing of familiar and unfamiliar Mandarin dialect tone systems. In *Proceedings of the Speech Prosody*, 842-846. https://www.isca-archive.org/speechprosody_2022/zhao22_speechprosody.pdf