Highly Proficient L2 Speakers Still Need to Attend to a Talker's Mouth When Processing L2 Speech

Joan Birulés[a] and Laura Bosch[a], Ferran Pons[a], David J. Lewkowicz[b]

[a]*Department of Cognition, Development and Educational Psychology, Universitat de Barcelona, Pg. Vall d'Hebron 171, 08035 Barcelona, Spain.*

[b]*Haskins Laboratories, New Haven, CT 06511, USA*

Corresponding author: Joan Birulés.

Highly Proficient L2 Speakers Still Need to Attend to a Talker's Mouth When Processing L2 Speech

Adults attend to a talker's mouth whenever confronted with challenging speech processing situations. We investigated whether L2 speakers also attend more to the mouth and whether their proficiency level modulates such attention. First, in Experiment 1, we presented native speakers of English and Spanish with videos of a talker speaking in their native and non-native language while measuring eye-gaze to the talker's face. As predicted, participants attended more to the talker's mouth in response to non-native than native speech. Then, Experiment 2 explored whether language proficiency affects attention to the talker's eyes and mouth when perceiving non-native, second-language speech. Results indicated that non-native speakers attended more to the mouth than native speakers, regardless of their level of L2 expertise. These results not only confirm that attention to a talker's mouth increases whenever speech-processing becomes more challenging, but crucially, they show that this is also true in highly competent L2 speakers.

Keywords: audiovisual speech perception, lip-reading, selective attention, face perception, second-language perception, non-native speech processing

Introduction

During most social interactions, we not only hear our interlocutors but we also see them. Seeing our interlocutors' faces gives us access to their mouth and, thus, to the source of speech consisting of spatiotemporally congruent visual and auditory speech cues (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009; Yehia, Rubin, & Vatikiotis-Bateson, 1998). The advantage of having access to such concurrent and congruent cues is that when they are processed together and integrated, they give rise to perceptually more salient communicative signals than do auditory-only speech cues (McGurk & MacDonald, 1976; Meredith & Stein, 1986; Risberg & Lubker, 1978; Summerfield, 1979). Evidence that this is the case comes from studies showing that speech comprehension is enhanced by concurrent visual speech cues when auditory speech is presented in noise (Cotton, 1935; Sumby & Pollack, 1954), when auditory speech is filtered (Sanders & Goodrich, 1971), or when auditory speech is presented in competition with other and irrelevant speech (Reisberg, 1978).

Importantly, in addition to increasing the perceptual salience of auditory speech, concurrent visual speech can enhance the processing of clear auditory speech. Three studies have provided evidence of the enhancing effects of visual speech. Reisberg and colleagues (1987) observed an 8% performance increase in an audiovisual condition when participants were presented with clear but syntactically and semantically complex speech and a 15% increase when they were presented with speech uttered in an unfamiliar accent or language. Similarly, Arnold and Hill (2001) found that concurrent visual speech cues enhanced the processing of intact auditory speech signals presented in other accents, languages, and tasks. Finally, Navarra and Soto-Faraco (2007) found that concurrent speech cues enhance second language (L2) perception at the phonological level. In sum, evidence to date indicates that redundantly specified audiovisual speech is more salient

and comprehensible than auditory-only speech.

If redundantly specified audiovisual speech is more salient and, if this facilitates processing, then it is reasonable to postulate that perceivers are likely to deploy their attentional resources to its source, namely the talker's mouth. This should be especially the case during speech and language acquisition as well as when processing conditions become challenging. Indeed, these theoretical possibilities are supported by findings from studies of infants, young children, and adults. In the aggregate, these findings indicate that attention to a talker's mouth emerges early in development, that it is affected by early linguistic experience, and that its magnitude depends on the specific task at hand.

The first study to explicitly link selective attention to a talker's mouth in infancy and speech and language acquisition was by Lewkowicz and Hansen-Tift (2012). These researchers exposed 4-, 6-, 8-, 10-, and 12-month-old monolingual, English-learning infants to a talking face speaking either in their native language or in a non-native language (Spanish). Findings indicated that, regardless of whether the speech was native or non-native, the 4-month-old infants attended more to the talker's eyes, the 6-month-old infants attended equally to the eyes and mouth, and that the 8- and 10-month-old infants attended more to the talker's mouth. In addition, the findings showed that the 12-month-olds also attended more to the talker's mouth but that they did so only when the talker spoke in the non-native language. Lewkowicz and Hansen-Tift pointed out that the attentional shift to the talker's mouth by 8 months of age happens to correspond with the onset of endogenous attention as well as the start of canonical babbling. Given this, the authors proposed that the emergence of endogenous attention allows infants to voluntarily direct their selective attention to the talker's mouth and that, by doing so, infants maximize their acquisition of their native phonology through access to the highly salient audiovisual speech cues located in the mouth. Furthermore, Lewkowicz & Hansen-Tift

presumed that the emergence of canonical babbling reflects infants' new interest in speech production and, thus, interpreted the shift in attention to a talker's mouth as reflecting infants' discovery that access to the salient audiovisual speech cues located in the mouth can facilitate their imitation of human speech. This last conclusion is in line with recent evidence by Imafuku and colleagues (2019) showing that increased attention to a talker's mouth is, indeed, related to higher vocal imitation at 6 months of age.

Lewkowicz and Hansen-Tift's (2012) finding that 12-month-old infants no longer attended more to a talker's mouth when exposed to native audiovisual speech but that they attended more to it when exposed to non-native audiovisual speech is important because it provides direct evidence that early language experience plays a key role in infants' selective attention to a talker's mouth. Infants attain their expertise with their native phonology by 12 months of age (Maurer & Werker, 2014). This means that the 12-month-olds' declining reliance on redundantly specified audiovisual cues is consistent with the idea that they no longer need to augment their processing when the speech is native because they are now familiar with it.

Overall, Lewkowicz & Hansen-Tift (2012) concluded that their findings of developmental changes in the relative amount of selective attention that infants deploy to a talker's eyes and mouth reflects speech processing *per se*. Findings from subsequent studies have been consistent with this conclusion. They have not only replicated the original findings but also shown that infants who are cognitively challenged during their early linguistic experience by having to master two closely related languages exhibit greater attention to a talker's mouth than their monolingual counterparts (Birulés, Bosch, Brieke, Pons, & Lewkowicz, 2018; Pons, Bosch, & Lewkowicz, 2015). Other studies also have shown that attention to a talker's mouth predicts later language acquisition (Tenenbaum et al., 2015; Tsang, Atagi, & Johnson, 2018; Young, Merin, Rogers, &

Ozonoff, 2009) and that failure to attend to a talker's mouth is associated with language learning disorders (Pons, Sanz-Torrent, Ferinu, Birulés, & Andreu, 2018).

Importantly, Lewkowicz & Hansen-Tift (2012) also tested adults by presenting the same videos as those presented to infants and asked the adults to just watch and listen. Results showed that unlike infants, adults deployed more attention to the talker's eyes. This finding was interpreted as reflecting the fact that adults normally focus on their interlocutors' eyes during typical social interactions (Yarbus, 1967). By focusing on the eyes, adults gain access to the various deictic social cues that are available there (for a review see: Birmingham & Kingstone, 2009). The Lewkowicz & Hansen-Tift (2012) adult findings are interesting in the context of findings from studies in which adults have been explicitly asked to process and/or disambiguate audiovisual speech as opposed to just watch and listen to it. These studies have found that, indeed, adults increase their attention to a talker's mouth when the speech processing task becomes more challenging. For example, studies have found that adults increase their attention to a talker's mouth as noise level increases (Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998) and as sound intensity decreases (Lansing & McConkie, 2003). Similarly, attention to the mouth increases when a silent face starts talking (Võ, Smith, Mital, & Henderson, 2012) or when a new speaker is presented (Buchan, Pare, & Munhall, 2008). Finally, studies have found that adults attend more to a talker's mouth when their task is to segment artificial speech (Lusk & Mitchel, 2016), report the words they hear (as opposed to judge faces' emotions, Buchan, Paré, & Munhall, 2007), or when they are asked to compare and identify specific speech utterances (Barenholtz, Mavica, & Lewkowicz, 2016). Overall, evidence to date has shown that information-seeking and specific task requirements play an important role in adults' relative distribution of selective attention to a talker's eyes and mouth.

If speech processing *per se* elicits greater attention to a talker's mouth, then this

raises an interesting question: Might adults rely more on the audiovisual cues located in a talker's mouth when they need to process non-native as opposed to native speech? Barenholtz et al. (2016) investigated this question and found that adults who were given an explicit speech-processing task, which required them to compare and identify 3 s-long audiovisual speech utterances, not only attended more to a talker's mouth when exposed to native audiovisual speech but even more when exposed to non-native audiovisual speech. This finding was interpreted as reflecting the greater difficulty of processing non-native speech and adults' greater reliance on audiovisual speech cues to overcome this challenge.

The fact that Barenholtz et al. (2016) assigned participants a specific task raises two interesting questions. First, do adults rely on the greater perceptual salience of audiovisual speech cues in a talker's mouth when they are exposed to non-native speech in the absence of a specific experimental task? Second, might L2 proficiency modulate the degree to which L2 learners/speakers attend to a talker's mouth? Put differently, might more experienced L2 learners/speakers of a non-native language rely less on attention to a talker's mouth to process audiovisual speech than those who are less experienced?

The purpose of the present study was to investigate the two questions posed above. To do so, we conducted two experiments. In Experiment 1, we investigated selective attention to talkers speaking in native and non-native fluent speech in adults whose knowledge of a non-native language was negligible. Crucially, here, we did not give the participants any specific task besides informing them that they would be asked some questions at the end of the testing session. In Experiment 2, we investigated whether relative L2 expertise modulates selective attention to a talker's mouth by testing L2 adult speakers who had varying degrees of proficiency in their second, non-native language. For Experiment 1, one plausible prediction was that the greater attention to the mouth

when perceving a non-native language would still be present in the absence of a specific speech-processing task. For Experiment 2, one plausible prediction was that highly proficient L2 speakers may attend less to the mouth than less proficient speakers and, hence, that highly proficient L2 speakers might exhibit a pattern of selective attention to a talker's face that is similar to that usually found in native speakers. Despite the plausibility of our second prediction, an equally plausible but alternative prediction is that highly proficient L2 speakers may attend more to a talker's mouth than do native speakers. This alternative prediction is based on evidence that highly competent non-native speakers do not generally reach the level of performance found in native speakers (Hyltenstam & Abrahamsson, 2000; Lecumberri, Cooke, & Cutler, 2010). Given this finding, it is possible that even expert L2 speakers rely on and profit from the greater perceptual salience of audiovisual speech cues in a talker's mouth.

Experiment 1

As noted earlier, Barenholtz et al. (2016) found that adults attended more to a talker's mouth when they were asked to identify a relatively brief (3 s) snippet of non-native as opposed to native audiovisual speech. One possible reason for this outcome is that the task of having to rapidly identify a speech utterance from relatively sparse information modulated adults' performance. If, however, greater attention to the talker's mouth was not due to the characteristics of the task in that study, then it is possible that adults might still exhibit greater attention to the mouth of a talker who can be seen and heard producing longer and more naturalistic non-native speech utterances, and that they will do so even in the absence of a specific speech-processing task.

The current experiment tested the possibility raised above by investigating selective attention to the eyes and mouth of a talker who could be seen and heard recounting segments of a story, rather than the types of 3 s clips of audiovisual speech

presented in the Barenholtz et al. (2016) study. We chose to present relatively extended, fluent speech utterances (60s long) to better capture a type of speech that we can encounter in our daily social interactions with our interlocutors (e.g., listening to a friend telling a story). The stories were presented in the participants' native and non-native languages. In addition, we counterbalanced the participants' native language by conducting the experiment in Spain and in the US. This enabled us to explore the effect of a non-native language on the deployment of selective attention to a talker's eyes and mouth independent of the specific language in which the speech was uttered.

Materials and Methods

Participants. We recruited 45 adults who had no or very little knowledge of the non-native language. Of these, 22 were native Spanish and Catalan bilingual speakers who were students at the University of Barcelona (mean age = 20.3 years, sd = 1.9; 4 male) and 23 were native, monolingual, English speakers who were students at Northeastern University in Boston (mean age = 23.6 years, sd = 2; 4 male). The students participated in the study for course credit. All participants answered a short questionnaire[1] whose purpose was to ascertain their knowledge, use, and formal training in their native and in the non-native language (Spanish for the American group and English for the Spanish-Catalan group). Participant inclusion criteria were that they had exclusive exposure to their native language/s while growing up and that they received a score of 2 or less (out of 5) in the self-reported competence of their basic skills in the non-native language (i.e. speaking and understanding, and a global self-report of the non-native language). Crucially, all participants reported having no or very little knowledge of the non-native language (in no case above an A2 Level, Common European Framework of Reference for Languages).

Stimuli. The stimulus materials consisted of video clips of a Catalan-Spanish-English trilingual female actor who was filmed from her shoulders up and who spoke in a natural voice while she kept her head still. The actor was recorded speaking a set of three 60 s long children's stories in Catalan, Spanish and English, respectively. It should be noted that the population in Barcelona is bilingual, meaning that people are native speakers of both Catalan and Spanish. Consequently, these two languages were presented in the experiment as native for the Spanish group and non-native for the English group.

Apparatus and procedure. Participants were tested in a quiet laboratory either at the University of Barcelona or at Northeastern University. In both laboratories, selective attention was measured with a REDn SensoMotoric Instruments (SMI, Teltow, Germany) eye tracker running at a sampling rate of 60 Hz. The participants sat at a table with a Dell Precision m4800 laptop computer in front of them at a distance of 60 cm from their eyes. The eye tracker camera was attached to the bottom of the computer screen and SMI's iViewRed software controlled the camera and processed eye gaze data. SMI's Experiment Center software controlled the stimulus presentation and data acquisition. The video clips were presented on the computer's 11 x 13 in screen and the soundtrack corresponding to the videos was presented through a pair of Sony headphones which participants wore throughout the experiment. We used a 9-point calibration routine to calibrate eye gaze by presenting a small yellow star in the centre of the screen as well as in the 4 corners of the screen and the 4 midpoints between the corners and the centre of the screen.

Once calibration was completed, we presented three videos in which the actor could be seen and heard speaking in Catalan, Spanish, or English. Participants were given the following instructions: "You are going to watch a woman telling you three different short stories, in three different languages. Please listen carefully because I will ask you

some questions about the stories you heard". These instructions were only given to ensure that participants were fully engaged in the experiment. The videos and the specific stories in them were assigned in random order and counterbalanced across participants. Crucially, it should be noted that our crossed design ensured that the stories spoken in Catalan and Spanish were in the Spanish participants' native languages and the stories spoken in English were in their non-native language while the reverse was true for the American participants. As a result, we were able to control for language-specific effects while examining the effects of language familiarity per se.

Results and Discussion

We defined three areas of interest (AOIs): the mouth, the eyes, and the face (see Figure 1) and measured the total amount of looking to each AOI. Using these data, and consistent with previous studies (Barenholtz et al., 2016; Birulés et al., 2018; Lewkowicz & Hansen-Tift, 2012), we calculated the proportion of total looking time (PTLT) deployed to the eyes and mouth, respectively, by dividing the total amount of time spent looking at each of these AOIs by the total amount of time spent looking at the face.

(Figure 1 about here)

As a first step, we averaged responsiveness to the Catalan and Spanish stories[2]. This allowed us to reduce the design to a native vs. non-native language comparison and, thus, permit us to relate our findings to those from the two most relevant previous studies (Barenholtz et al., 2016; Lewkowicz & Hansen-Tift, 2012). In addition, this enabled us to make a balanced comparison of responsiveness in the Spanish and American participants. Next, we analyzed the data from the native and non-native language conditions for both groups of participants as defined above. To do so, we used a mixed, repeated-measures ANOVA, with Language Group (Spanish, English) as a between-

subjects factor and Language Condition (native and non-native) and AOI (eyes, mouth) as within-subject's factors. Results revealed a main effect of AOI [$F(1, 43) = 8.34$, $p = .006$, $\eta p2 = .16$] and an AOI x Language Condition interaction [$F(1, 43) = 55.08$, $p < .001$, $\eta p2 = .56$]. The Language Group main effect was not significant [$F(1, 43) = 0.38$, $p = .539$, $\eta p2 < .01$], nor did it interact with AOI [$F(1, 43) = 1.29$, $p = .262$, $\eta p2 = .03$].

Figure 2 shows the two statistically significant findings. As can be seen, even though participants exhibited an overall preference for the eyes, they deployed their selective attention to the eyes and mouth differently depending on whether the actor spoke in a native or non-native language. Follow-up t-tests, comparing the PTLT to the eyes and mouth, respectively, across the native and non-native language conditions revealed that participants attended less to the eyes and more to the mouth in the non-native language than in the native one [eyes: $t(44) = 6.76$, $p < .001$, $d = 1.01$; mouth: $t(44) = 7.07$, $p < .001$, $d = 1.05$]. Paired t-tests comparing PTLT to the eyes and mouth within each of the language conditions, respectively, indicated a preference for the eyes in the native condition [$t(44) = 5.43$, $p < .001$, $d = 0.81$] and equal attention to the eyes and mouth in the non-native condition [$t(44) = 0.49$, $p = .624$, $d = 0.07$][3].

(Figure 2 about here)

The results from this experiment indicate that when adults are exposed to an extended audiovisual monologue and are asked to pay attention to its content, they exhibit differential patterns of selective attention to the talker's eyes and mouth as a function of their familiarity with the language spoken. Specifically, when the speech is in their native language, adults attend more to the talker's eyes than mouth. When, however, the speech is not in their native language, adults attend more to the talker's mouth, resulting in equal attention to the eyes and mouth. This pattern of findings is consistent with evidence from speech-in-noise experiments where adults have been found to attend more to a talker's

eyes in a silent context but equally to the talker's eyes and mouth in a noisy context (Buchan et al., 2007; Lansing & McConkie, 2003; Vatikiotis-Bateson et al., 1998). The current findings add to this evidence by showing that adults' strategy of deploying greater attention to a talker's mouth under challenging conditions includes the processing of non-native audiovisual speech. That is, our findings indicate that adults' selective attention to different parts of a talker's face is modulated by their prior familiarity with a specific language. When the audiovisual speech is in a familiar language, adults direct most of their attention to the talker's eyes presumably because they do not need to direct greater cognitive resources to processing the audiovisual speech information *per se* and can, instead, focus on the social cues available in a social partner's eyes. In contrast, when audiovisual speech is in an unfamiliar language, adults attend more to the talker's mouth. This is presumably because this permits them to take advantage of the greater perceptual salience of audiovisual speech which, in turn, is presumed to augment their ability to extract the semantic information inherent in an utterance spoken in an unfamiliar language.

Importantly, the absence of a Language Group x AOI interaction indicates that the American and the Spanish participants exhibited a similar pattern of selective attention to the eyes and mouth in their response to native and non-native audiovisual speech. This suggests that these effects are not language-specific but rather that they reflect a general feature of responsiveness to an unfamiliar language. Moreover, the absence of a Language Group x AOI interaction indicates that language background (i.e., bilingual vs. monolingual) did not affect the relative deployment of selective attention to the eyes and mouth. These results indicate that, in the absence of specific processing requirements, bilingual adults do not take greater advantage of the greater salience of redundantly specified audiovisual speech. This finding contrasts with findings from

studies comparing selective attention to the eyes and mouth in monolingual vs. bilingual infants and children (Catalan and/or Spanish) showing that relative attention to the eyes and mouth differs as a function of language background (Birulés et al., 2018; Pons et al., 2015). The most likely explanation for the adult-infant/child difference is that, compared to adults, bilingual infants and children are compelled to rely more on the greater perceptual salience of audiovisual speech in speech processing tasks because they are cognitively and linguistically more naïve than adults.

Experiment 2

The results from Experiment 1 showed that adults attend more to a talker's mouth when processing non-native fluent audiovisual speech. This suggests that the difficulty of a speech-processing task affects the degree to which adults rely on the audiovisual redundancy cues available in a talker's mouth. If adults do, indeed, allocate their selective attention to a talker's eyes and mouth as a function of processing demands, it is possible that the degree of proficiency in another language might affect the relative distribution of attention to a talker's eyes and mouth. Put differently, might L2 adults who are highly proficient in a non-native language and, therefore presumably find the processing of non-native audiovisual speech easier, exhibit the same pattern of selective attention to a talker's eyes and mouth found in adults' response to native speech? If language proficiency is an index of speech-processing expertise then one plausible prediction is that less proficient L2 learners might attend more to a talker's mouth than eyes, whereas highly proficient L2 speakers may attend more to a talker's eyes than mouth when they are exposed to a talker speaking in a non-native language. As noted earlier, however, given the fact that L2 speakers rarely attain native-like levels of expertise for non-native

speech (Lecumberri et al., 2010), an equally plausible prediction is that even highly proficient L2 learners may attend more to a talker's mouth than do native speakers.

The present experiment was designed to test these predictions. To examine them, we presented a video of a talker speaking in English to Spanish-Catalan speakers differing in the degree of language proficiency in a non-native language (i.e., English) and to native speakers of English and recorded their selective attention to the talker's eyes and mouth.

Materials and Method

Participants. We tested 76 adult participants. These were classified into four different groups based on their knowledge of the English Language: native, high-, intermediate- and low-level of proficiency. The participants from the native group were undergraduate students from Northeastern University in Boston who were native English speakers (mean age = 23 years, sd = 1.3, 3 male). The participants from the three non-native groups were undergraduate students at the University of Barcelona who were native Catalan and Spanish bilingual speakers. From those, 19 were highly proficient in English (high B2 to a C2 levels of the Common European Framework of Reference for Languages; mean age = 21.2 years, sd = 3.2, 4 male), 19 had an intermediate-level of proficiency (high A2 to a B1 levels; mean age = 19 years, sd = 1.8, 3 male), and 19 had a low level of English proficiency (A1 to A2 levels; mean age = 20.7 years, sd = 1.9, 3 male)[4]. Spanish participants were asked to self-report their level of English based on their previous official exams (i.e. Cambridge English tests, TOEFL, IELTS, etc.). Once the participants completed the experiment, their English proficiency level was re-evaluated by administering the "Cambridge General English Placement Test" which consists of 25 multi-choice questions. Three participants were excluded from the sample because their self-reported proficiency level was higher than the level obtained with the English test.

Stimuli. New stimulus videos were created because we were concerned that the children's tales used in Experiment 1 may not reveal differences within the proficiency levels due to comprehension ceiling effects. As a result, we recorded three new videos that consisted of an American female speaker reciting 20s English monologues (these consisted of anecdotes and opinion pieces on social topics). Together, the three videos presented participants with 60s of fluent speech as in Experiment 1. As in Experiment 1, the actor was recorded from her shoulders up, her eyes and mouth size and position were similar to those in the videos presented in Experiment 1, and she held her head still while speaking in a natural tone of voice.

Apparatus and procedure. The apparatus and procedure were the same as in Experiment 1 except that here we administered a post-test questionnaire to the non-native speakers. This questionnaire consisted of nine multi-choice questions about the content of the stories. The current experiment was conducted at the University of Barcelona and at Northeastern University. The laboratories in both locations were dimly lit and sound-attenuated.

Results and Discussion

We used a mixed, repeated-measures ANOVA, with Proficiency (low, intermediate, high and native) as a between-subjects factor and AOI (eyes and mouth) as a within-subjects factor to determine whether the four English proficiency groups differed in their selective attention to the talker's eyes and mouth. Results yielded a significant interaction between Proficiency and AOI [$F(3, 72) = 7.04$, $p < .001$, $\eta p2 = .23$] and no significant main effects [Proficiency: $F(3, 72) = 1.40$, $p = .250$, $\eta p2 = .06$; AOI: $F(1, 72) = 1.64$, $p = .205$, $\eta p2 = .02$]. The lack of an AOI main effect reflects an overall balanced distribution between the eyes and mouth, while the significant AOI x Proficiency

interaction indicates that the distribution of selective attention depended on participants' proficiency level. Figure 3 shows the PTLT scores for the eyes and mouth in each of the proficiency groups.

(Figure 3 about here)

To identify the source of the Proficiency x AOI interaction, we used paired t-tests to compare the PTLT scores for the eyes and mouth in each group, respectively. Results revealed that whereas the three non-native groups looked equally to the two AOIs [low: $t(18) = 1.33$, $p = .201$, $d = .30$; intermediate: $t(18) = 0.70$, $p = .491$, $d = .16$; high: $t(18) = 0.24$, $p = .817$, $d = .05$], the native group looked more to the eyes than to the mouth [$t(18) = 7.93$, $p < .001$, $d = 1.82$]. To further identify the source of the interaction, we used independent t-tests to compare attention to the mouth and eyes, respectively, across the four groups. The t-tests confirmed that the three non-native groups looked less to the eyes than the native group [low: $t(36) = 4.98$, $p < .001$, $d = 1.62$; intermediate: $t(36) = 3.51$, $p = .001$, $d = 1.14$; high: $t(36) = 3.51$, $p = .001$, $d = 1.14$] and that they looked more to the mouth than the native group [low: $t(36) = 4.33$, $p < .001$, $d = 1.40$; intermediate: $t(36) = 4.10$, $p < .001$, $d = 1.33$; high: $t(36) = 2.82$, $p = .009$, $d = .92$]. In addition, the t-tests across the three non-native groups yielded no significant differences.

These results indicate that the three proficiency groups distributed their selective attention to the talker's eyes and mouth in a similar way. Nevertheless, visual inspection of the data seen in Figure 3 suggests that attention to the mouth was slightly lower in the higher proficiency groups. Therefore, we extracted the proportion of correct responses of each participant's (1) English Test and (2) Post-viewing Comprehension Test and tested the correlation between these scores (number of correct responses divided by the total) and their PTLT difference scores (PTLTeyes - PTLTmouth). The Pearson Product

Moment correlation yielded null results [r = .068, n = 57, p = .615; r = .10, n = 57, p = .444, respectively] and, thus, confirmed our previous conclusions (see Figure 4).

(Figure 4 about here)

The results from Experiment 2 indicate that the degree of non-native language proficiency does not affect the relative deployment of selective attention to a talker's eyes versus mouth in Catalan-Spanish speakers tested with fluent English audiovisual speech utterances. Interestingly, however, and in line with the findings from Experiment 1, native English speakers attended more to the talker's eyes than mouth, whereas Spanish speakers attended equally to the talker's eyes and mouth regardless of their proficiency in English. Follow-up comparisons showed that the Catalan-Spanish speakers attended less to the talker's eyes and more to the talker's mouth than did the English speakers.

Discussion

Studies have found that adults attend more to the mouth of a talking face when they have to process speech in noise or non-native, as opposed to native, audiovisual speech (Barenholtz et al., 2016; Buchan et al., 2007; Lansing & McConkie, 2003; Vatikiotis-Bateson et al., 1998). The current study investigated the theoretically reasonable proposition that the degree of L2 proficiency reflects speech-processing expertise and that this factor also might have an effect on the relative amount of selective attention that L2 speakers deploy to a talker's eyes and mouth when exposed to non-native audiovisual speech. We conducted two experiments to test this proposition. First, we wanted to establish that adults also exhibit greater attention to a talker's mouth when exposed to non-native as opposed to native audiovisual speech when the speech is a relatively long utterance and when they are not given a specific task. Thus, in Experiment 1, we tested native adult speakers of English and Spanish with videos of a talker

producing fluent native and non-native audiovisual speech and only told them that they would be asked some questions at the end of the testing session. Consistent with the findings from previous studies, we found that participants attended more to the talker's mouth when they were exposed to non-native than native audiovisual speech. Having established that relatively long, non-native audiovisual speech utterances elicit greater attention to the mouth under minimal instruction conditions, we then put our primary hypothesis to test by examining selective attention to a talker's face in native Spanish-Catalan speakers who differed in their level of English-language proficiency to fluent English audiovisual speech and compared their responsiveness to native English speakers. Findings showed that level of non-native language proficiency did not have differential effects on selective attention to a talker's face and that L2 learners deployed equal amounts of attention to the talker's eyes and mouth. Crucially, however, as a group, L2 learners attended more to the talker's mouth than did native speakers of English who attended more to the talker's eyes.

The present findings provide new insights when considered in the context of previous findings. These have shown that adults look at the eyes of talking faces when not given a specific speech-processing task (e.g. Lewkowicz & Hansen-Tift, 2012; Yarbus, 1967) but that they attend more to the mouth when they are asked to identify short native audiovisual speech utterances and that they attend even more to the mouth when asked to identify non-native utterances (Barenholtz et al., 2016). Like in the Barenholtz et al. (2016) study, we also found in Experiment 1 that adults attend more to a talker's mouth when they spontaneously process relatively long non-native audiovisual speech utterances (i.e., when they are not given an explicit processing task). In contrast to Barenholtz et al. (2016), however, we also found that overall, adults attended more to a talker's eyes than mouth. This difference is most likely due to the fact that Barenholtz

et al. (2016) presented very short speech segments whereas we presented much longer ones (60 s). The short speech segments, together with an explicit identification task, most likely compelled participants to quickly focus their attention on the critical information in a talker's mouth. In contrast, the longer speech segments, and the absence of any explicit processing task, most likely enabled participants to more fully explore the talker's face.

The current results are also interesting in light of findings from previous studies showing that adults shift their attention from the eyes to the mouth when auditory-only cues become compromised by factors such as noise (Buchan et al., 2007; Lansing & McConkie, 2003; Vatikiotis-Bateson et al., 1998), the participants' older age (Thompson & Malloy, 2004), or the relevance of speech processing (Buchan et al., 2007; Lusk & Mitchel, 2016). Overall, these findings, along with the present ones, suggest that the greater attention accorded to a talker's mouth provides access to the redundant and, thus, highly salient audiovisual speech cues. Such cues are known to increase comprehension (Macleod & Summerfield, 1987; Sumby & Pollack, 1954; Summerfield, 1979), including the perception of non-native speech (Arnold & Hill, 2001; Navarra & Soto-Faraco, 2007; Reisberg et al., 1987). Moreover, our results are interesting in light of findings from previous studies showing that the processing of non-native speech is cognitively more effortful than the processing of native speech (Borghini & Hazan, 2018). Once again, this suggests that an attentional shift to a talker's mouth provides non-native speakers with greater access to audiovisual speech cues which presumably helps them overcome the greater challenge of processing unfamiliar linguistic input.

If adults deploy greater attention to the mouth under challenging processing conditions, including the processing of non-native speech, it follows that the difficulty of the processing task also might modulate the amount of attention directed to the mouth.

Indeed, Vatikiotis-Bateson et al. (1998) found that adults' attention to the mouth increased continuously with the amount of noise (i.e. none, low, medium and high). Similarly, in an audiovisual speech segmentation task, Lusk & Mitchel (2016) found that attention to the mouth decreased as familiarization progressed and as adults learned new artificial word boundaries. Based on such findings, we expected that participants' level of non-native language proficiency would modulate the amount of attention directed to the mouth. In other words, we expected that highly proficient L2 speakers of English would not need to rely on the audiovisual speech cues to the same extent as speakers with lower proficiency. Accordingly, we made two opposite, but theoretically plausible predictions. One was that highly proficient L2 speakers might exhibit a selective attention pattern similar to that found in native speakers. The other was that highly proficient L2 speakers may, nonetheless, attend more to a talker's mouth because studies have found that even highly proficient L2 speakers differ from native ones in some crucial aspects of language perception such as phonology (McClelland, Fiez, & McCandliss, 2002).

Remarkably, the results of Experiment 2 were consistent with the latter prediction. They showed that despite the fact that the L2 speakers differed significantly in their level of English competence, all of them exhibited a similar pattern of selective attention in that they attended more to the mouth than did the native-language group. In addition, as in Experiment 1, the L2 group exhibited equal attention to the eyes and mouth whereas the native-language group exhibited a clear preference for the eyes.

Although our results are in line with the fact that increased processing difficulty is correlated with increased attention to a talker's mouth, they also suggest that this relationship is a non-linear one. That is, at least in the case of L2 speakers differing in their level of non-native language expertise, increasing expertise does not appear to be correlated with decreasing attention to a talker's mouth. This finding is consistent with

evidence that adults' selective attention to a talking face cannot be attributed to single attentional shifts to the mouth to disambiguate an ambiguous phoneme or a word that is difficult to understand (Vatikiotis-Bateson et al., 1998; Võ et al., 2012). Given this, it may be that participants' specific patterns of selective attention to a talker's eyes and mouth, as measured by us and in all previous studies, are a relatively crude measure of dynamic changes in speech processing. It may be that more sensitive measures of selective attention are required to reveal a relationship between L2 expertise and differential selective attention to a talker's mouth.

Although our findings were not consistent with the possibility that highly expert L2 speakers can dispense with access to redundant audiovisual cues, the fact that the responsiveness of the highly proficient L2 speakers differed from that observed in the native speakers is consistent with findings from second-language learning studies. These studies have found that the production and perception of L2 phonology is quite an arduous task for L2 learners. They have also found that learners' plasticity is limited and that highly proficient L2 speakers rarely attain the ultimate phonological competence of native speakers (McClelland et al., 2002; Pallier, Bosch, & Sebastián-Gallés, 1997). Even when their speech recognition performance appears to be native-like, the addition of noise renders competent non-native listeners less accurate than native speakers (Cutler, Garcia Lecumberri, & Cooke, 2008) and they require more cognitive effort when processing non-native speech because they rely on strategies that tend to be less efficient than those of native speakers (Borghini & Hazan, 2018). For example, in phoneme discrimination, highly proficient L2 speakers sometimes focus on different and less informative formants than native speakers do (Iverson et al., 2003). Moreover, they rely less on contextual plausibility (Mattys, Carroll, Li, & Chan, 2010) due to the fact that their lexical and semantic knowledge is not as easily accessed (Bradlow & Alexander, 2007).

All in all, when the findings discussed above are considered together with those from Experiment 2 it becomes apparent that even highly proficient L2 speakers find second language speech perception challenging and, hence, they do not process speech in the same automatic fashion as native speakers do. Rather, L2 speakers seem to rely on access to the redundant audiovisual speech cues located in the talker's mouth to augment their L2 comprehension.

In conclusion, the results from the current study corroborate findings from other studies (Barenholtz et al., 2016; Lansing & McConkie, 2003; Lusk & Mitchel, 2016; Vatikiotis-Bateson et al., 1998). They demonstrate that greater speech-processing difficulty elicits greater reliance on the highly salient audiovisual perceptual cues available in a talker's mouth both in native speakers processing non-native audiovisual speech and in all L2 speakers, regardless of their expertise, in processing an L2 language. If redundant audiovisual cues facilitate speech processing, then it is possible that L2 learning could be maximized by training with audiovisual, rather than auditory-only, non-native speech materials (Bernstein, Auer, Eberhardt, & Jiang, 2013; Heikkilä et al., 2018). Future studies might consider exploring this possibility.

Acknowledgements

This work was supported by the Spanish Ministerio de Ciencia e Innovación, Grant PSI2014-55105-P and PGC2018-097487-B-100 and by the National Science Foundation, Grant BCS-0751888.

Disclosure statement

The authors report no conflict of interest.

Data availability statement

The data that support the findings of this study are available from the corresponding author, [J. B.], upon reasonable request.

References

Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology*, *92*(2), 339–355. https://doi.org/10.1348/000712601162220

Barenholtz, E., Mavica, L., & Lewkowicz, D. J. (2016). Language familiarity modulates relative attention to the eyes and mouth of a talker. *Cognition*, *147*, 100–105. https://doi.org/10.1016/j.cognition.2015.11.013

Bernstein, L. E., Auer, E. T., Eberhardt, S. P., & Jiang, J. (2013). Auditory perceptual learning for speech perception can be enhanced by audiovisual training. *Frontiers in Neuroscience*, *7*(7 MAR), 1–16. https://doi.org/10.3389/fnins.2013.00034

Birulés, J., Bosch, L., Brieke, R., Pons, F., & Lewkowicz, D. J. (2018). Inside Bilingualism: Language Background Modulates Selective Attention to a Talker's Mouth. *Developmental Science*. https://doi.org/10.1111/desc.12755

Borghini, G., & Hazan, V. (2018). Listening effort during sentence processing is increased for non-native listeners: A pupillometry study. *Frontiers in Neuroscience*, *12*(MAR), 1–13. https://doi.org/10.3389/fnins.2018.00152

Bradlow, A. R., & Alexander, J. A. (2007). Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *The Journal of the Acoustical Society of America*, *121*(4), 2339–2349. https://doi.org/10.1121/1.2642103

Buchan, J. N., Pare, M., & Munhall, K. G. (2008). The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception. *Brain Research*, *1242*, 162–171. https://doi.org/10.1016/j.brainres.2008.06.083

Buchan, J. N., Paré, M., & Munhall, K. G. (2007). Spatial statistics of gaze fixations during dynamic face processing. *Social Neuroscience*, *2*(1), 1–13.

https://doi.org/10.1080/17470910601043644

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*(7). https://doi.org/10.1371/journal.pcbi.1000436

Cotton, J. C. (1935). Normal "Visual Hearing." *Science*, *82*(2138), 592–593.

Cutler, A., Garcia Lecumberri, M. L., & Cooke, M. (2008). Consonant identification in noise by native and non-native listeners: Effects of local context. *The Journal of the Acoustical Society of America*, *124*(2), 1264–1268. https://doi.org/10.1121/1.2946707

Heikkilä, J., Lonka, E., Meronen, A., Tuovinen, S., Eronen, R., Leppänen, P. H., … Tiippana, K. (2018). The effect of audiovisual speech training on the phonological skills of children with specific language impairment (SLI). *Child Language Teaching and Therapy*, (September), 026565901879369. https://doi.org/10.1177/0265659018793697

Hyltenstam, K., & Abrahamsson, N. (2000). Who can become native-like in a second language? All, some, or none?: On the maturational constraints controversy in second language acquisition. *Studia Linguistica*, *54*(2), 150–166. https://doi.org/10.1111/1467-9582.00056

Imafuku, M., Kanakogi, Y., Butler, D., & Myowa, M. (2019). Demystifying infant vocal imitation: The roles of mouth looking and speaker's gaze. *Developmental Science*, (March), e12825. https://doi.org/10.1111/desc.12825

Iverson, P., Kuhl, P. K., Akahane-Yamadac, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *38*, 361–363. https://doi.org/10.1016/S0

Lansing, C. R., & McConkie, G. W. (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Perception & Psychophysics*, *65*(4), 536–552. https://doi.org/10.3758/BF03194581

Lecumberri, M. L. G., Cooke, M., & Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech Communication*, *52*(11–12), 864–886. https://doi.org/10.1016/j.specom.2010.08.014

Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(5), 1431–1436. https://doi.org/10.1073/pnas.1114783109

Lusk, L. G., & Mitchel, A. D. (2016). Differential Gaze Patterns on Eyes and Mouth During Audiovisual Speech Segmentation. *Frontiers in Psychology*, *7*(February), 52. https://doi.org/10.3389/fpsyg.2016.00052

Macleod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, *21*(2), 131–141. https://doi.org/10.3109/03005368709077786

Mattys, S. L., Carroll, L. M., Li, C. K. W., & Chan, S. L. Y. (2010). Effects of energetic and informational masking on speech segmentation by native and non-native speakers. *Speech Communication*, *52*(11–12), 887–899. https://doi.org/10.1016/j.specom.2010.01.005

Maurer, D., & Werker, J. F. (2014). Perceptual narrowing during infancy: A comparison of language and faces. *Developmental Psychobiology*, *56*(2), 154–178. https://doi.org/10.1002/dev.21177

McClelland, J. L., Fiez, J. A., & McCandliss, B. D. (2002). Teaching the /r/-/l/ discrimination to Japanese adults: Behavioral and neural aspects. *Physiology and*

Behavior, *77*(4–5), 657–662. https://doi.org/10.1016/S0031-9384(02)00916-2

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 764. https://doi.org/10.1038/260170a0

Meredith, M. A., & Stein, B. E. (1986). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of Neurophysiology*, *56*(3), 640–662. https://doi.org/10.1152/jn.1986.56.3.640

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: Visual articulatory information enables the perception of second language sounds. *Psychological Research*, *71*(1), 4–12. https://doi.org/10.1007/s00426-005-0031-5

Pallier, C., Bosch, L., & Sebastián-Gallés, N. (1997). A limit on behavioral plasticity in speech perception. *Cognition*, *64*, B9–B17. https://doi.org/10.1016/S0010-0277(97)00030-9

Pons, F., Bosch, L., & Lewkowicz, D. J. (2015). Bilingualism Modulates Infants' Selective Attention to the Mouth of a Talking Face. *Psychological Science*, *26*(4), 490–498. https://doi.org/10.1177/0956797614568320

Pons, F., Sanz-Torrent, M., Ferinu, L., Birulés, J., & Andreu, L. (2018). Children With SLI Can Exhibit Reduced Attention to a Talker's Mouth. *Language Learning*, (68), 180–192. https://doi.org/10.1111/lang.12276

Reisberg, D. (1978). Looking where you listen: visual cues and auditory attention. *Acta Psychologica*, *42*(4), 331–341. https://doi.org/10.1016/0001-6918(78)90007-0

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by Eye: The Psychology of Lip-reading* (pp. 97–113). New Jersey, US: Lawrence Erlbaum Associates, Inc.

Risberg, A., & Lubker, J. (1978). Prosody and speechreading. *Quarterly Progress and Status Report*, *4*, 1–16. Retrieved from http://www.speech.kth.se/prod/publications/files/qpsr/1978/1978_19_4_001-016.pdf

Sanders, D. A., & Goodrich, S. J. (1971). The Relative Contribution of Visual and Auditory Components of Speech to Speech Intelligibility under Varying Conditions of Frequency Distortion. *Journal of Speech Language and Hearing Research*, *14*(1), 154–159. https://doi.org/10.1121/1.2143572

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, *26*(2), 212–215. https://doi.org/10.1121/1.1907309

Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetica*, *36*(4–5), 314–331. https://doi.org/10.1159/000259969

Tenenbaum, E. J., Sobel, D. M., Sheinkopf, S. J., Shah, R. J., Malle, B. F., & Morgan, J. L. (2015). Attention to the mouth and gaze following in infancy predict language development. *Journal of Child Language*, *42*(6), 1173–1190. https://doi.org/10.1017/S0305000914000725

Thompson, L. A., & Malloy, D. (2004). Attention resources and visible speech encoding in older and younger adults. *Experimental Aging Research*, *30*(3), 241–252. https://doi.org/10.1080/03610730490447877

Tsang, T., Atagi, N., & Johnson, S. P. (2018). Selective attention to the mouth is associated with expressive language skills in monolingual and bilingual infants. *Journal of Experimental Child Psychology*, *169*, 93–109. https://doi.org/10.1016/j.jecp.2018.01.002

Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S., & Munhall, K. G. (1998). Eye movement

of perceivers during audiovisual speech perception. *Perception & Psychophysics*, *60*(6), 926–940. https://doi.org/10.3758/BF03211929

Võ, M. L.-H., Smith, T. J., Mital, P. K., & Henderson, J. M. (2012). Do the eyes really have it? Dynamic allocation of attention when viewing moving faces. *Journal of Vision*, *12*(13), 3–3. https://doi.org/10.1167/12.13.3

Yarbus, A. L. (1967). *Eye movements and vision* (Translated). New York, New York, USA: Plenum Press. https://doi.org/10.1016/0028-3932(68)90012-2

Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, *26*(1–2), 23–43. https://doi.org/10.1016/S0167-6393(98)00048-X

Young, G. S., Merin, N., Rogers, S. J., & Ozonoff, S. (2009). Gaze Behaviour and Affect at 6 Months: Predicting Clinical Outcomes and Language Development in Typically Developing Infants and Infants At-Risk for Autism. *Developmental Science*, *12*(5), 798–814. https://doi.org/10.1111/j.1467-7687.2009.00833.x.Gaze

Footnotes:

1 Link to the questionnaire used in the USA: https://forms.gle/raCZpBtXbL4fCT1K6
Link to the questionnaire used in Spain: https://forms.gle/mot9W3faCp4RdzWGA

2 The same pattern of results is obtained when using only the Catalan, only the Spanish or an average of both languages.

3 Although the results are reported in PTLT Scores, the whole analysis was repeated using the raw scores, and the results yielded the same significant effects and the conclusions remained the same.

4 As a reference of the English level of the students, the CEFRL B1 (Intermediate) level is defined as someone who can understand the main points of clear standard input on familiar matters, can deal with most travelling situations in that language, and can produce simple connected text on familiar topics. The CEFRL C2 (highly proficient) level is defined as someone who can understand with ease virtually everything heard or read, can summarize information from different sources in a coherent presentation, and can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.

Figure 1. Still photo of the talker's face showing the eyes, mouth, and face AOIs.

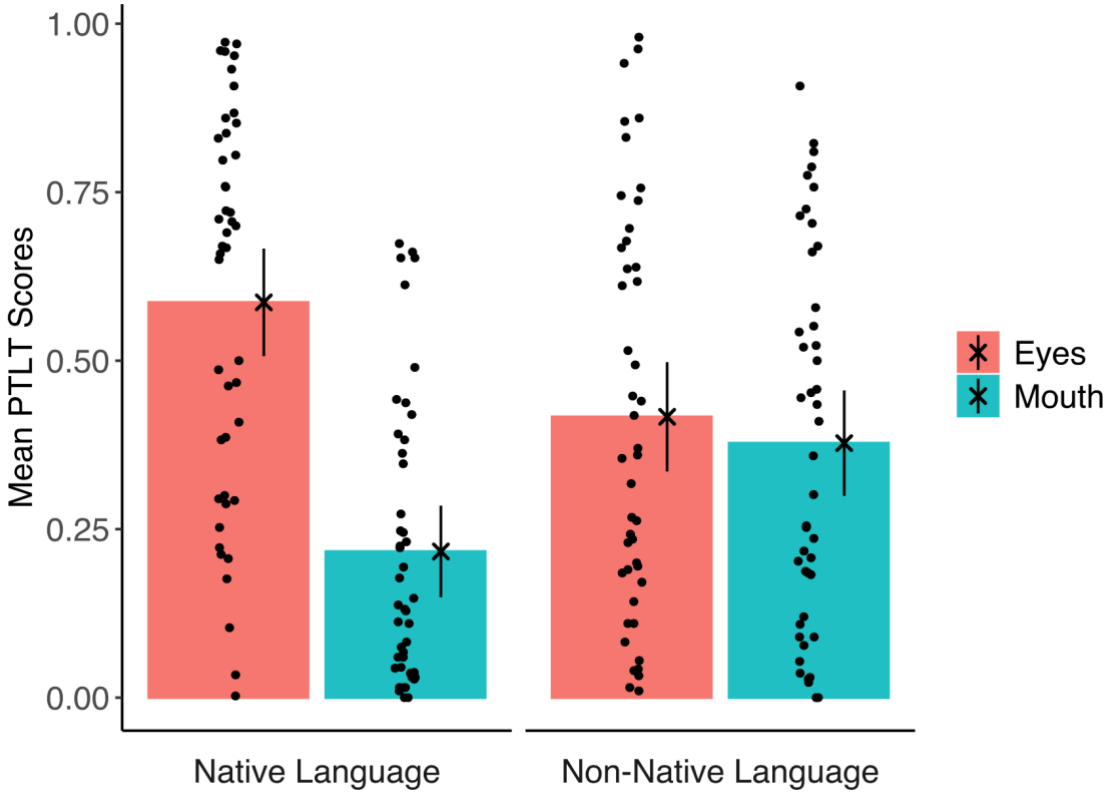

Figure 2. Average PTLT scores for the eyes and mouth AOIs, respectively, in the native and non-native language conditions. Error bars represent the standard errors of the mean.

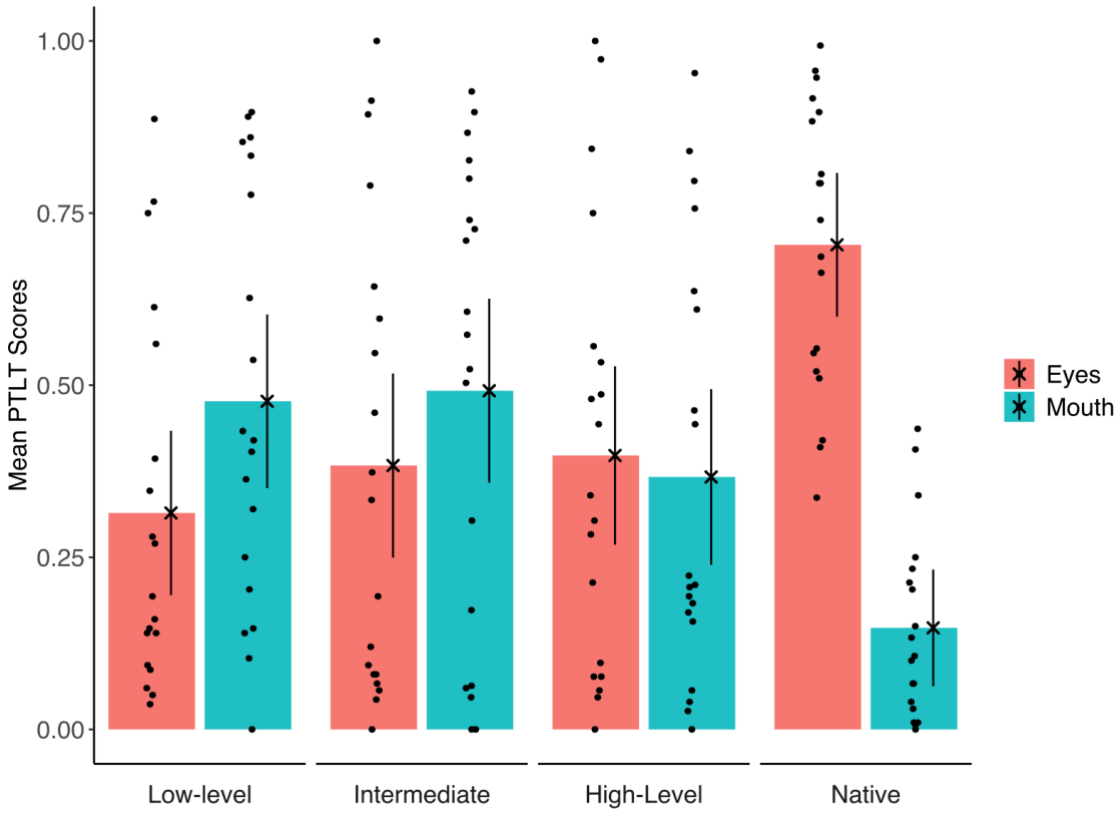

Figure 3. Mean PTLT scores to the eyes and mouth for the non-native (Low-, Intermediate-, high-level) and native language conditions. Error bars represent the standard errors of the mean.

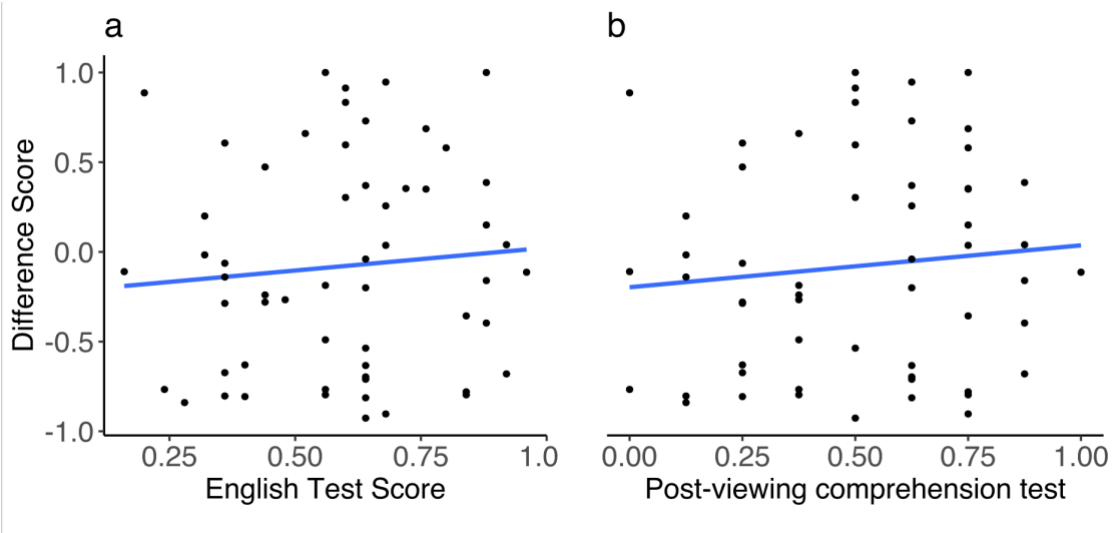

Figure 4. Correlation between the Difference Score (PTLTeyes - PTLTmouth) and the proportion scores of (a) the English Test (n° of correct questions divided by the total), and (b) the Post-viewing comprehension test (n° of correct questions divided by the total) of non-native participants.