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, joanbirules@gmail.com.

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 a great deal of information. From a language perspective, an interlocutor's mouth is an especially rich source of information because it provides spatiotemporally congruent visual and auditory speech cues (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009; Yehia, Rubin, & Vatikiotis-Bateson, 1998). When the visual and auditory cues are processed together, they provide a perceptually new and more salient speech signal than the one provided by auditory-only information (McGurk & MacDonald, 1976; Meredith & Stein, 1986; Risberg & Lubker, 1978; Summerfield, 1979). The greater perceptual salience of combined visual and auditory speech cues is illustrated by findings showing that speech comprehension is enhanced by concurrent visual speech cues when auditory-only speech is presented in noise (Cotton, 1935; Sumby & Pollack, 1954), is filtered (Sanders & Goodrich, 1971), or when it is presented in competition with other, irrelevant, speech (Reisberg, 1978).

Importantly, studies also have found that concurrent visual speech cues provide more than a "back-up system" to be employed in the context of environmental noise (Johnstone, 1996). This is illustrated by the fact that concurrent visual speech cues can even enhance processing of clear auditory speech. For example, 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, studies have found that concurrent and redundant visual

speech gestures can enhance second language (L2) perception at the phonological level (Navarra & Soto-Faraco, 2007). In sum, it is clear that combined visual and auditory speech cues provide a perceptually more salient and comprehensible linguistic signal than auditory-only cues.

If the greater perceptual salience of audiovisually redundant speech signals facilitates processing then it is likely that perceivers will deploy their attentional resources to the source of audiovisual redundancy, namely a talker's mouth. Indeed, this is supported by findings from studies of infants, young children, and adults. These studies show that perceivers deploy their attention to a talker's mouth and that the degree to which they do so is modulated by early linguistic experience and the specific task at hand. For example, in the first study to demonstrate the developmental onset of selective attention to a talker's mouth specifically related to speech processing, Lewkowicz and Hansen-Tift (2012) exposed monolingual, English-learning infants to a talking face speaking either in the participants' native language or in a non-native language (Spanish) and examined their selective attention to the talker's eyes and mouth. Results indicated that 4-month-old infants deployed more of their attention to the talker's eyes but that 8- and 10-month-old infants deployed more of their attention to the talker's mouth and that they did so regardless of whether the talker spoke in their native or non-native language. Lewkowicz and Hansen-Tift noted that the attentional shift to the talker's mouth by around 8 months of age corresponds with the start of canonical babbling and the onset of endogenous attention in infancy. Given this, they interpreted the attentional shift to the talker's mouth as evidence that, by this age, infants have become interested in speech perception and production and thus begin directing their attention to the talker's mouth to obtain maximally salient speech information which, in turn, enables them to acquire their native phonology. This conclusion is in line with recent evidence by Imafuku and colleagues

(2019) showing that increased attention to a talker's mouth is related to higher vocal imitation at 6 months of age. Finally, Lewkowicz and Hansen-Tift obtained evidence that early language experience affects selective attention to a talker's mouth. They found that 12-month-old infants look equally to a talker's eyes and mouth when they are exposed to audiovisual speech in their native language but that they continue to show a preference for the mouth when they are exposed to audiovisual speech in a non-native language. This last set of findings is consistent with the idea that by 12 months of age, infants are becoming more familiar with the phonology of their native language (Maurer & Werker, 2014). Presumably, once infants have tuned to their native phonology, they are less likely to rely on a talker's mouth to augment their processing of speech information.

Recent studies have supported Lewkowicz & Hansen-Tift's (2012) conclusion that the attentional shift from a talker's eyes to the mouth reflects infants' emerging interest in audiovisual speech. These studies have replicated the original findings and have shown that early linguistic experience (e.g., the learning of two closely related languages) can modulate these attentional patterns (Birules, Bosch, Brieke, Pons, & Lewkowicz, 2018; Pons, Bosch, & Lewkowicz, 2015). In addition, these studies have shown that greater attention to a talker's mouth is associated with language acquisition (Tenenbaum et al., 2015; Tsang, Atagi, & Johnson, 2018; Young, Merin, Rogers, & Ozonoff, 2009).

Importantly, Lewkowicz & Hansen-Tift (2012) also tested adults to obtain a baseline measure of selective attention to talking faces in the mature state. To replicate their infant study as closely as possible, they presented the same videos that they presented to the infants and asked them to just watch and listen. Results showed that unlike the 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, especially when they are engaged in social interaction rather than speech processing per se (Yarbus, 1967). By focusing on a social partner's 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 were explicitly asked to process and/or disambiguate audiovisual speech. These studies have found that adults deploy more attention to a talker's mouth when speech is masked by noise (Lansing & McConkie, 2003; Vatikiotis-Bateson, Eigsti, Yanoyi, & Munhall, 1998) or when a silent face starts talking (Võ, Smith, Mital, & Henderson, 2012). These studies also 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 (Buchan, Paré, & Munhall, 2007), or identify speech utterances (Barenholtz, Mavica, & Lewkowicz, 2016). Together, these studies reveal 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 participants who were given an explicit speech-processing task (i.e. to identify 3 s-long audiovisual speech utterances) not only deployed more attention to the talker's mouth when exposed to talkers speaking in their native language but that they deployed even more attention to the mouth when exposed to talkers speaking in a non-native language. This finding was interpreted as reflecting the greater difficulty of processing speech in a non-native language and adults' greater reliance on audiovisual speech cues to overcome this

challenge. One of the interesting questions that these findings raise is whether L2 proficiency also might modulate the degree of attention to the mouth. It is theoretically possible that L2 learners/speakers who are more experienced in a non-native language may rely less on attention to a talker's mouth to process speech than those who are less experienced.

The purpose of the present study was to investigate whether the degree of language proficiency with a non-native language might modulate the amount of selective attention to a talker's mouth when exposed to non-native versus native speech. To investigate this question, we hypothesized that highly proficient L2 speakers may deploy less attention to the mouth than less proficient ones 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. Nonetheless, when considering previous evidence showing 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) we also recognize that highly proficient L2 speakers may still attend more to a talker's mouth than do native speakers.

To test these predictions, we conducted two experiments. First, in Experiment 1, we investigated selective attention to talkers speaking in native and non-native fluent speech in adults whose knowledge of the non-native language was negligible and did not vary. Then, in Experiment 2 we investigated whether relative expertise in a second language modulates selective attention to a talker's mouth by testing L2 adult speakers with varying degrees of proficiency in their second, non-native language.

Experiment 1

Barenholtz, Mavica, and Lewkowicz (2016) found that adults deployed more selective attention to a talker's mouth when their task was to identify an utterance in a non-native language. Participants performed a simple match-to-sample task where they had to determine whether a sound track that they heard in a post-test trial corresponded to the first or second video of a talker speaking two different utterances presented during a prior encoding phase. Crucially, the utterances presented in that study were relatively brief (3 s) snippets of speech. This raises the possibility that the pattern of selective attention found in that study was specifically due to the relatively demanding task of having to rapidly identify a speech utterance from relatively sparse information. In other words, it may be that the fairly high degree of selective attention deployed to the talker's mouth reflected the need to focus maximally on the audiovisual cues to enhance rapid identification and that longer utterances, which enable participants to explore a talker's face more freely, may elicit less attention to the mouth. This possibility is supported by Lewkowicz & Hansen-Tift's (2012) findings showing that when adults were exposed to longer fluent speech utterances and asked to just watch and listen, they attended more to the talker's eyes than mouth regardless of whether the speech was native or not.

Given that the Barenholtz et al. results may reflect a relatively demanding speech-processing task, here we investigated differential allocation of selective attention to a talker's eyes and mouth in response to the more usual types of utterances that we are normally exposed to in our daily lives. Thus, we presented relatively extended, fluent speech utterances (60s long) in the participants' native and non-native languages. Moreover, we counterbalanced subjects' 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. Finally, even though our participants were not given a specific speech-processing task, they were told that they would first see and hear some audiovisual speech utterances and that they would then be given some questions related to these utterances at the end of the experiment.

Materials and Methods

Participants. *A* total of 45 adults participated in this study. Of these, 22 were native Spanish and Catalan bilingual speakers who were students at the University of Barcelona and 23 were native, monolingual, English speakers who were students at Northeastern University in Boston. The students participated in the study for course credit. Crucially, all participants self-described as 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 3 short 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. Additionally, using a crossed design between the Spanish and the American controlled for language-specific effects while examining the effects of language familiarity per se.

Results and Discussion

Consistent with previous studies on selective attention to talking faces, 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, 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 AOI by the total amount of looking at the face.

(Figure 1 about here)

First, to ensure that the Spanish and American participants did not respond differently to the Catalan and Spanish videos, we used a repeated-measures analysis of variance (ANOVA), with Language Condition (Catalan and Spanish) and AOI (eyes and mouth) as within-subjects' factors, to analyse the PTLT scores in each participant group, respectively. The ANOVA of the Spanish participants' data yielded an AOI main effect [$F(1, 21) = 5.98$, $p = .023$, $\eta p2 = .222$], indicating greater overall looking at the eyes. Crucially, the Language Condition x AOI interaction was not significant [$F(1, 21) = 1.75$, $p = .200$, $\eta p2 = .077$], indicating that the Spanish participants looked more at the eyes in both language conditions. The ANOVA of the American participants' data did not yield a significant AOI effect [$F(1, 22) = .78$, $p = .386$, $\eta p2 = .034$], indicating that the American participants looked equally to the two AOIs. Also, like the Spanish participants, the American participants exhibited the same pattern of selective attention to the eyes and mouth across the two language conditions (Language Condition x AOI interaction [$F(1, 22) = 2.18$, $p = .154$, $\eta p2 = .090$]). Given that responsiveness to the Spanish and Catalan videos did not differ in either group, we only used the data from the Spanish video condition for the main analysis (a supplementary analysis of responsiveness in the Catalan video condition yielded results that were identical to those from the Spanish video condition). Overall, the native-language condition was Spanish for the Spanish participants and English for the American participants while the non-native language condition was English for the Spanish participants and Spanish for the American participants. This enabled us to both simplify the design to one native and one non-native language condition– similar to the design in the two previous studies (Barenholtz et al.,

2016; Lewkowicz & Hansen-Tift, 2012)–and to then make a balanced comparison between the Spanish and American participants.

Next, we analysed 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 [$F(1, 43) = 9.27$, $p < .001$, $\eta p2 = .177$] and an AOI x Language Condition interaction [$F(1, 43) = 46.19$, $p < .001$, $\eta p2 = .518$]. Figure 2 shows these 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 in the non-native language condition [$t(44) = 6.35$, $p < .01$, $d = .95$] and that they attended more to the mouth in the non-native condition [$t(44) = 6.41$, $p < .01$, $d = 1.07$]. 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 [$t(44) = 5.63$, $p < .01$, $d = 2.00$] and equal attention to the eyes and mouth in the non-native condition [$t(44) = .70$, $p = .49$ $d = .277$].

(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 contents, 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 deploy more of their attention 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 showing that adults usually attend more to a talker's eyes except in the context of noise when they attend equally to the talker's eyes and mouth (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 speech in an unfamiliar language. Specifically, our findings indicate that adults' selective attention to different parts of a talker's face is modulated by their familiarity and, thus, prior experience with a specific language. When the speech was in a familiar language, adults directed most of their attention to the talker's eyes. This is presumably because their familiarity with their native language enables them to engage in relatively 'automatic' speech processing. In contrast, when the speech was in an unfamiliar language, adults deployed more of their selective attention to the talker's mouth. Presumably, this enables them to take advantage of the greater perceptual salience of audiovisual speech and helps them overcome the greater challenge of trying to extract some of the content inherent in an utterance spoken in an unfamiliar language.

Importantly, the fact that the American and the Spanish participants exhibited the same pattern of attention in response to native and non-native speech suggests that these effects are not specific to English or Spanish but rather that they reflect a general feature of responsiveness to an unfamiliar language. Moreover, the lack of differences also indicates that participants' language background (i.e. bilingual vs. monolingual) did not affect their relative deployment of selective attention to a talker's eyes and mouth.

Experiment 2

Barenholtz et al. (2016) proposed an active-processing hypothesis to account for increased attention to a talker's mouth. According to this hypothesis, adults attend more to a talker's mouth when a task requires them to actively process the information inhere nt in an utterance and attend less to it in the absence of an explicit processing task. Cons istent with this hypothesis, the results from Experiment 1 showed that when adults were asked to actively process a non-native speech monologue, they attended more to the talk er's mouth. As noted earlier, however, the participants in the Barenholtz et al. study and in Experiment 1 had minimal knowledge of the non-native language and hence they cou ld not comprehend the content of the speech. This contrasts with the more usual L2 soci al circumstance where interlocutors often have some previous working knowledge of the non-native language and attempt to use it to their best ability to understand as much cont ent of a speech utterance as possible. Given this, it is possible that participants' L2 profi ciency may modulate their ability to comprehend the utterance.

If adults do, indeed, allocate their selective attention to a talker's eyes and mouth as a function of processing demands, this raises an interesting question with respect to th e results from Experiment 1. The question is whether the degree of proficiency in anothe r language also may affect the relative distribution of attention to a talker's eyes and mo uth. Put differently, might L2 adults who are highly proficient in a non-native language e xhibit the same pattern of selective attention to a talker's eyes and mouth found in mono lingual adults' response to native speech? If language proficiency affects selective attent ion to a talker's eyes and mouth, then one plausible prediction is that highly proficient L 2 speakers might spend most of their time attending to a talker's eyes when the talker sp eaks in their second language. A second and equally plausible prediction is that L2 adult who possess low or intermediate proficiency in a non-native language may attend more t

o a talker's mouth. As noted earlier, however, if the fact that L2 speakers rarely attain na tive-like levels of expertise for non-native speech is taken into account (Lecumberri et al ., 2010), it may be that the highly proficient L2 learners may still attend more to the mou th than native speakers do.

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 bilinguals differ ing in the degree of language proficiency in a non-native language (i.e., English) and to monolingual native speakers of English and recorded their selective attention to the talke r's eyes and mouth.

Materials and Method

Participants. We tested a total of 76 participants. The majority of the participants (n = 57) were undergraduate students at the University of Barcelona. All of these students were native Catalan and Spanish bilingual speakers. The remainder of the participants were 19 undergraduate students from Northeastern University in Boston who were native English speakers. The Spanish participants were subsequently classified into three groups: 19 who were highly proficient in English (high B2 to a C2 levels of the Common European Framework of Reference for Languages), 19 who had an intermediate-level of proficiency (high A2 to a B1 levels), and 19 who had a low level of English proficiency (A1 to A2 levels)[1]. When we first recruited the participants, we asked them 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". Three participants were excluded from the sample because their self-reported proficiency level and the level obtained with the English test did not match.

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 every-day life monologues (including anecdotes and opinion pieces on social topics, 60s in total as in Experiment 1). The video characteristics were comparable to those presented in Experiment 1. That is, the actor was recorded from her shoulders up, her eyes and mouth size and position were similar to that in the videos presented in Experiment 1, and she spoke in a natural voice while she held her head still.

Apparatus and procedure. The apparatus and procedure were identical to that in Experiment 1. 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

First, to ensure that the three pre-selected non-native English groups actually comprehended the stories according to their English level, we conducted an ANOVA on the post-test questionnaire scores to determine if they differed as a function of English proficiency level (low, intermediate, high). As expected, the results showed that the three groups differed in their performance [Low: $M = .20$, $SD = .14$; Intermediate: $M = .54$, $SD = .15$; High: $M = .80$, $SD = .08$, $F(56) = 98.92$, $p < .001$].

We then conducted the principal analysis whose purpose was to determine whether the three English proficiency groups differed in terms of their selective attention to the talker's eyes and mouth. We used a mixed, repeated-measures ANOVA, with Proficiency (low, intermediate, high) as a between-subjects factor and AOI (eyes and

mouth) as a within-subjects factor to analyse the data. Contrary to expectations, the ANOVA yielded no significant effects, indicating that the three proficiency groups distributed their selective attention to the talker's eyes and mouth in similar ways (see Figure 3). Inspection of the data (Figure 3) revealed that attention to the mouth was slightly lower in the higher proficiency groups. Therefore, we extracted each participant's (1) English Test Scores and (2) Post-viewing Comprehension Scores and tested the correlation between these scores and their PTLT difference scores. The Pearson Product Moment correlation yielded null results [$r = .068$, $n = 57$, $p = .615$, $r = .10$, $n = 57$, $p = .444$] and, thus, confirmed the results of the ANOVA (see Figure 4).

Finally, we collapsed the data for the three proficiency Spanish groups and compared their data to the data from the American group of participants for whom the talker spoke in their native language. For this comparison, we used a mixed, repeated-measures ANOVA, with Group (Spanish, American) as a between-subjects factor and AOI (eyes and mouth) as a within-subjects factor. Results yielded a significant AOI main effect [$F (1, 74) = 11.21$, $p = .001$, $\eta p2 = .132$] and a significant AOI x Group interaction [$F (1, 74) = 20.00$, $p < .001$, $\eta p2 = .213$]. The AOI main effect reflects an overall preference for the eyes while the significant interaction indicates that the distribution of selective attention depended on whether the language spoken was the participants' native language or a non-native one. To identify the source of the AOI x Group interaction, we first used paired t-tests to compare the PTLT eye versus mouth scores in each of the groups, respectively. Results revealed that the Spanish group looked equivalently to the two AOIs [$t (57) = 1.02$, $p = .31$] but that the American group looked more to the eyes than to the mouth [$t (18) = 7.93$, $p < .001$]. Finally, we used independent t-tests to compare attention to the mouth and eyes, respectively, across the two groups. Results confirmed

that the non-native group looked less to the eyes [t (74) = 4.46, p < .001] and more to the mouth [t (74) = 3.96, p < .001] than the native group.

(Figure 3 about here)

(Figure 4 about here)

The results from Experiment 2 are consistent with our alternative prediction. That is, they 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 Spanish bilingual speakers tested with English audiovisual 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 Spanish speakers attended less to the talker's eyes than the English speakers and that they attended more to the talker's mouth than did the English speakers.

Discussion

Studies have shown that adults attend more to the mouth of a talking face when presented with speech in noise and with non-native as opposed to native speech (Barenholtz et al., 2016; Lansing & McConkie, 2003). The current study investigated the theoretically reasonable proposition that the degree of second-language proficiency might also modulate the amount of attention that adults deploy to a talker's mouth when they are exposed to non-native audiovisual speech. To test this proposition, first we tested native speakers of English and Spanish with videos of a talker producing fluent speech in

their native and in a non-native, unfamiliar language. As in previous studies, we found that fluent non-native speech elicited more attention to the talker's mouth than did native speech. Then, we tested native bilingual Spanish/Catalan speakers, who differed in their level of English-language proficiency, with a video of a talker producing fluent audiovisual utterances in English and compared their responsiveness to that in native English speakers. Interestingly, we found that level of non-native language proficiency did not have differential effects on selective attention. That is, regardless of English proficiency level, L2 learners with different levels of English knowledge deployed equivalent attention to the talker's eyes and mouth and, as a group, they attended more to the talker's mouth compared to native speakers who attended more to the talker's eyes.

The present findings provide new insights into adults' reliance on different selective attention strategies when processing fluent audiovisual speech and, together with the results from prior studies, suggest that adults' distribution of attention to a talker's eyes and mouth depends on specific speech processing demands. This is illustrated by several sets of findings, including the current ones. For example, Lewkowicz & Hansen-Tift's (2012) presented 45 s segments of native and non-native audiovisual speech utterances to native English speakers and instructed them to simply watch the videos. Findings showed that the participants attended more to the talker's eyes than mouth in both language conditions. Barenholtz et al. (2016) first presented pairs of 3 s utterances to native English speakers and then presented an audio-only version of one of the videos and instructed the participants to indicate whether the audio-only utterance corresponded to the first or second video. In contrast to the adults in the Lewkowicz & Hansen-Tift's (2012) study, the participants in this study attended more to the talker's mouth in response to native and non-native speech and, in addition, they attended more to the mouth in the non-native speech condition. In the aggregate, these findings

demonstrate that the distribution of selective attention to audiovisual speech depends on whether adults are assigned a specific speech processing task or not and what the specific task requires them to do.

On the one hand, the results from Experiment 1 partly replicate the Barenholtz et al. (2016) findings by showing that adults attend more to a talker's mouth when they are engaged in fluent audiovisual speech processing. On the other, they differ from those of Barenholtz et al. (2016) in showing that adults deploy more overall attention to a talker's eyes than mouth and that they shift their attention away from the eyes to the mouth in response to non-native speech. The most likely reason for this difference is that Barenholtz et al. (2016) presented very short segments of audiovisual speech, whereas here we presented longer (60 s) segments of audiovisual speech. The very short speech segments presented by Barenholtz et al. (2016) required participants to quickly focus on the critical information to identify the speech segment, whereas the longer speech segments in the current study provided participants with more time to explore the talker's face and probably contributed to their greater exploration of the talker's eyes. This interpretation is consistent with Võ et al.'s (2012) results showing that participants shifted their attention to the mouth whenever the talker spoke but directed their attention to other parts of the talker's face in the absence of speech.

The current results are 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 noise (Buchan et al., 2007; Lansing & McConkie, 2003; Vatikiotis-Bateson et al., 1998), by participants' older age (Thompson & Malloy, 2004), or when speech processing becomes highly relevant (Buchan et al., 2007; Lusk & Mitchel, 2016). Together, these findings suggest that the greater attention accorded to a talker's mouth provides access to the redundant and, thus, highly salient audiovisual

speech cues which 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 also 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, while the other was that the highly proficient group may still need to rely more on audiovisual redundancy and, thus, attend more to the mouth because even highly proficient 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 similar patterns of selective attention in that they attended more to the mouth than did the native-language group. In addition, as in Experiment 1, the non-native 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 also in line with the fact that increases in processing difficulty correspond with increases in selective 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 speakers with different levels of non-native language expertise, increasing expertise does not correspond with decreasing levels of selective attention to a talker's mouth. On the one hand, such results are consistent with previous evidence showing that adults' selective attention patterns 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 and that more sensitive measures might be required.

Nonetheless, the fact that the highly proficient group clearly differed from the native group is consistent with findings from second-language learning studies showing that the production and perception of L2 phonology is quite an arduous task for L2 learners. These studies have shown 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 (Borghini & Hazan, 2018) when processing non-native speech because they rely on strategies that tend to be less efficient than those of native speakers. 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, the combination of such findings with those of Experiment 2 suggests that even highly proficient participants find second language speech perception challenging, and hence they cannot engage in the earlier noted "relatively automatic speech processing" as do native speakers. Instead, they still need to rely more on the maximally salient audiovisual speech cues located in the talker's mouth – when they are available– to augment their L2 comprehension.

The current results corroborate findings from other studies by demonstrating that greater speech-processing difficulty elicits greater reliance on the highly salient audiovisual perceptual cues available in a talker's mouth. In addition, our findings show for the first time that this general principle also applies to people with differing levels of non-native language proficiency but with an important caveat: the degree of selective attention to a talker's mouth is not affected by the level of non-native language expertise. Overall, findings to date suggest that (1) perceivers resort to a greater reliance on highly salient audiovisual speech cues located in a talker's mouth to enhance their speech comprehension and that (2) they rely on such cues even if they are expert L2 speakers.

Finally, our findings have practical implications; they suggest that second-language learning can be maximized by audiovisual training with audiovisual, rather than auditory-only, non-native speech materials (Bernstein, Auer, Eberhardt, & Jiang, 2013; Heikkilä et al., 2018). Future studies that incorporate other more fine-grained measures of L2 perception and processing will contribute to gain a better understanding of the current results.

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

Birmingham, E., & Kingstone, A. (2009). Human social attention: A new look at past, present, and future investigations. *Annals of the New York Academy of Sciences*, *1156*, 118–140. https://doi.org/10.1111/j.1749-6632.2009.04468.x

Birules, 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., 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

Johnstone, R. A. (1996). Multiple displays in animal communication: “backup signals” and “muliple messages.” *Proceedings of the Royal Society of London Series B-Biological Sciences*, *351*(Real 1990), 329–338.

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

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 behavior 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

Figure Captions:

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 (a) the English Test Scores, and (b) the Post-viewing comprehension test of non-native participants. Dots represent individual means, the line represents a fitted linear model, and the shaded area represents standard errors of the mean.

Footnotes:

1 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.