# Speech perception: a comparative study of audiovisual matching in four and ten months-old infants

**Léna Kervran**

**Msc Brain and Cognition, UPF**

Director: Núria Sebastián-Gallés

Tutor: Mathilde Fort

Group: SAP – Speech Acquisition and Perception Group, UPF.

Date: 24/08/2015

# Contents

# Abstract

The present study looks at audiovisual association in four and ten months-old infants, in the context of complex audiovisual connected speech events. In particular, we wanted to look at whether and how bilingual and monolingual infants at four and ten months were able to match a talking face with the corresponding sentence they heard. We measured the amount of infants' looking time to the eyes, the mouth, and the face of two side-by-side speaking faces. One of these faces visually articulated the auditory sentence that was being heard whereas the other face did not. Higher looking time to the articulating face that matched with the auditory sentence meant they identified the congruent face, following the paradigm developed by Kuhl & Meltzoff (1982). Our results indicate that infants cannot match the visual and auditory sentence either at four or ten months. However we saw that at four months, bilingual infants look significantly more than monolinguals at the mouth of the speaker. This is coherent with previous studies and it is partly what we had predicted. At ten months, our results suggest that infants could possibly be in the process of learning matching.


*Keywords*: audiovisual speech; talking faces, infancy; early bilingualism, audiovisual matching; eye-tracking, attention.

# Introduction

It is widely accepted that speech perception is a multimodal event (e.g., Rosenblum, 2008). Most of the time, it occurs in a bimodal fashion, in the sense that a perceiver crucially has access to at least two sensorial types of information while a speaker is talking: he/she has access to both visual information on the speaker's face, and to auditory information present as an acoustic signal. Therefore, these two types of information can be processed together in order to create a coherent perception of what is being said.

It is known that for adults, visual information presents an advantage for processing streams of connected speech, especially under adverse conditions of speech perception. For instance, seeing the face of a person talking improves speech processing when the acoustic signal is noisy (e.g., Sumby & Pollack, 1954) or when it is produced in a second language (e.g., Navarra & Soto, 2007). Furthermore, having access to the facial articulatory gestures of the speaker provides the complementary information that has been masked or distorted in the acoustic signal (e.g., Schwartz, Berthommier & Savariaux, 2004; Ross, Saint-Amour, Lavitt, Javitt & Foxe, 2007). In these contexts, adults focus their visual attention on the mouth region of speakers, giving them access to redundant audiovisual speech cues[1]. However, adults prefer focusing on the upper part of the face when the speech signal is easy to process, in the cases of native language or of a clear acoustic signal (e.g., Everdell, Marsh, Yurick, Munhall & Paré, 2007) allowing them to access socio-emotional information coming from the eyes region (e.g., Buchan, Paré & Munhall, 2007).

In their first months of life, infants already show rudimentary skills to associate visual speech with its corresponding acoustic signal[2]; but this capacity greatly varies over the

1 The speech cues present in the mouth region are termed *redundant* when these speech cues enable to 'lipread' the acoustic signal.

[2] Patterson & Werker, J.F (1999); Patterson & Werker (2002); Patterson & Werker (2003); Kuhl & Meltzoff (1982); Burnham & Dodd (1996); Burnham & Dodd (2004); Yeung & Werker (2013); Kushnerenko, Teinonen, Volein, Csibra (2008); Pons, Lewkowicz, Soto-Faraco, Sebastián-Gallés (2009).

course of infancy and differs for monolinguals and bilinguals (e.g., Pons, Lewkowicz, Soto-Faraco & Sebastián Gallés, 2009; Lewkowicz & Hansen-Tift, 2012; see Lewkowicz, Minar, Tift & Brandon, 2015; and Lewkowicz & Ghazanfar, 2006, for reviews). When watching audiovisual talking faces, research indicates that monolingual infants seem to exhibit two shifts of visual attention towards the eyes/mouth of the speaker (Lewkowicz & Hansen-Tift 2012). During the first shift, four and eight months-old infants direct their attention from the eyes to the mouth of the speaker: while four months-old infants look proportionally more to the eyes than the mouth, eight months-old infants look proportionally more the mouth than the eyes. A second shift by twelve months (Lewkowicz & Hansen-Tift, 2012) shows an increase of attention back to the eyes region compared to what happens at eight months. As for bilingual infants, they pay attention earlier – at four months of age – to the mouth region of the speaker, and this preference is kept through infancy up to twelve months (Pons, Bosch & Lewkowicz, 2015). The different patterns of attention between monolinguals and bilinguals are attributed to the fact that bilinguals need to learn two languages at a time, and therefore, they need to keep them apart; they do this by paying more selective attention to the mouth region, which gives them access to redundant speech cues. The aforementioned patterns of visual attention in infancy suggest that, like adults, infants can selectively deploy their attention to the mouth of the speaker to improve their ability to process speech signals.

Thus, one goal of this study was to investigate how much human infants, who are in the process of learning their native language(s), rely on audiovisual speech cues to process talking faces. In order to look at visual attention we used a procedure very similar to that of Kuhl & Meltzoff (1982) and Lewkowicz et al. (2015). We used the Intermodal Preferential Looking procedure, where the perceiver can hear only one auditory signal, but he/she can see two side-by-side faces articulating two distinct visual sounds/sentences. Only one of these visually articulated sounds/sentences matches the auditory signal. Higher proportion of looking time to the matching face shows that perceivers can associate a speech sound with its corresponding articulatory gesture, or at least with its visual correlate (Kuhl & Meltzoff, 1982). Thanks to this procedure, Kuhl & Meltzoff (1982) showed that vowels presented in isolation (e.g., /i/,

/a/) can be matched audiovisually as early as four-and-a-half months-old[3]. Interestingly, Lewkowicz et al. (2015) showed with the same procedure that with running speech – in his study, these were sentences produced in infant-directed speech – infants only managed to perform the audiovisual association by twelve months. In our study we used the same procedure; but the differences are that instead of using infant-directed speech like Lewkowicz et al., we used adult-directed speech. Moreover, while both Kuhl & Meltzoff and Lewkowicz et al. used a familiarization phase by showing the silent faces pronouncing either the vowels or the sentences prior to testing, our study does not include any kind of familiarization.

In a nuthshell, the present study investigated whether 1) four and ten months-old infants can match an articulating face with its corresponding acoustic signal 2) if/how monolinguals and bilinguals are different in their performance and selective attention patterns. We predicted that a) if the matching is possible, the strategy in terms of selective attention would be to look at the mouth in higher proportions, because having access to visual speech cues could enable infants to encode its sensori-motor components, to improve their speech perception processing abilities, and enable infants to solve the complex audiovisual situations they encounter. We also predicted that b) given that bilinguals pay more attention to the mouth area early in their development, they should be better than their monolingual peers at matching visual speech with its auditory component due to a).

3 A complementary study showed that this is possible as soon as two months of age (Patterson & Werker, 2003).

# Method

## *Participants*

Twelve healthy and full-term four-months-old infants growing up in a Spanish (N=6, of which 6 girls) and Catalan (N=6, of which 3 girls and 3 boys) monolingual environment, and Spanish-Catalan bilingual environment (N=8, of which 4 girls and 4 boys) were tested in this study. The data from ten more infants were excluded from the final analyses due to the total looking time to the screen being less than 10% (N=8), and for fussiness (N=2).

Ten healthy and full-term ten-months-old infants growing up in a Spanish (N=5, of which 2 girls and 3 boys) and Catalan (N=2, of which 1 girl and 1 boy) monolingual environment, and Spanish-Catalan bilingual environment (N=5, of which 2 girls and 3 boy) were tested in this study. The data from one more infant were excluded from the final analyses due to the total looking time to the screen being less than 10% (N=1).

Prior to the study, parents were asked to fill in the informed consent for their child's participation. A detailed questionnaire (Bosch & Sebastián-Gallés, 2001) was filled in by the investigators in order to determine the infants' linguistic background. Infants with less than 20% of direct exposure to a second language were considered monolingual.

## *Stimuli*

The stimuli consisted of 16 sentences in Spanish, and 16 in Catalan. These were produced in adult-directed-speech in either Spanish or Catalan by a native Spanish/Catalan bilingual female speaker. The sentences were taken from *The Little Prince* by Saint-Exupéry and were modified[4] for the purpose of this study. For instance,

[4] The sentences were modified in the sense that their length was equalised, and a break – coma – was introduced so that both auditorily and visually, a break would be perceived at the same moment on the two side-by-side faces. Sentences were paired depending on these parameters : length and break.

one of the sentences in Spanish was: "Es el mejor momento de mi aventura en el desierto, aunque ya me he bebido todo el agua" ("It is the best moment of my adventure in the desert, even though I have already drunk all the water") (see Appendix). The sentences lasted between 5010 ms and 6080 ms; the duration of Spanish and Catalan sentences did not differ statistically (mean Spanish = 5314 ms, mean Catalan = 5369 ms, $t < 1$).

The stimuli consisted of audiovisual movies that were constructed with Adobe Premiere Pro CS3 and consisted of two side-by-side video clips of the same female speaker looking directly at the camera and uttering a prepared script[5]. Across all the movies, which were counterbalanced for side across trials, one of the talking faces matched with the soundtrack while the other did not.

Crucially, both the matching and non-matching talking faces were synchronized relative to the onset and the offset of the auditory soundtrack. These talking faces on the screen had approximately the same size as that of a human person.

## *Procedure*

The participants were tested in a white, dimly lit, sound-attenuated laboratory room. They were seated on their caregiver's lap and they were positioned approximately 60 cm away from a 1080x1920 screen. The stimuli were launched using MATLAB and Tobii Analytics Software Development Kit (Tobii Analytics SDK).

5 Appendix.

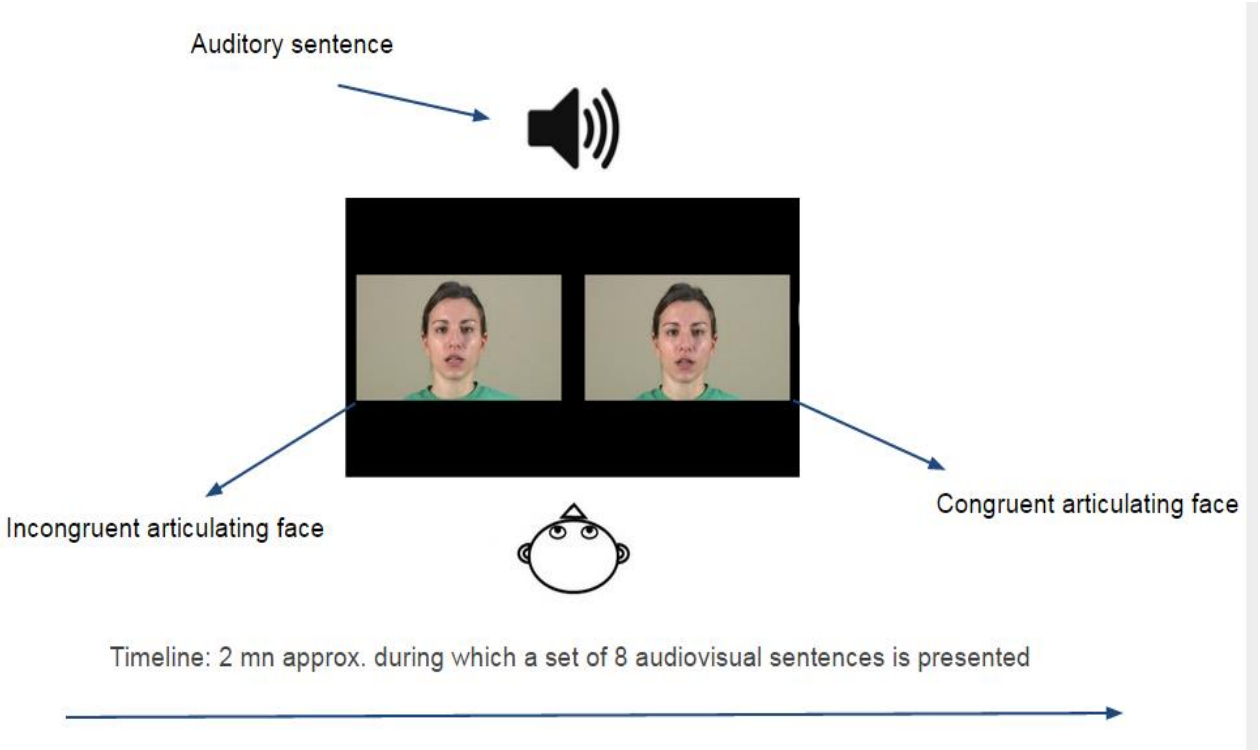


*Figure 1*. Experimental setting.

Prior to the experiment, each infant's eye movements were calibrated using a five-point routine in order to ensure positional validity of gaze measurements. We asked the parents not to look at the screen in order to ensure there would not be any interference with the infant experiment: we asked them to continuously fixate the infant's hair. Once the phase of calibration of the eye-tracker was completed, the experiment could start. Each trial began with a central, visual and auditory attention-getter, displayed at the center of screen, in order to make sure infants focused.

*Figure 2*. Attention-getter.

The attention-getter lasted a minimum of 1000 ms but remained on the screen until the infant looked at it. Then the experimenter pressed a key to launch the stimuli.

We used an Intermodal Preferential Looking procedure (IPL). Infants watched two side-by-side talking faces pronouncing audiovisual sentences (N=8); the female speaker looked directly at the camera and uttered the stimuli. Infants heard an auditory signal that matched only one of the videos and that came from the center of the screen, at equal distance from both sides. Visual sentences were always presented by pairs. For example, visual sentence A in Spanish was always paired with visual sentence B in Spanish[6]. For half of the babies the auditory stream matched the video for visual sentence A, and for the other half it matched the video for visual sentence B. This was counterbalanced across lists so that each member of the video pairs was seen in the matching or non-matching condition but only once by each baby. We also made sure that the matching video appeared on the right side of the screen for half of the trials, and on the left side of the screen for the other half; this counterbalancing was done across 16 lists. The order of presentation of the stimuli was pseudo-randomized so that across participants, each sentence was displayed at a different position (either in the first third, the second, or the last third of the lists). The order of sentences was counterbalanced across 8 experimental lists so that each participant perceived each audiovisual sentence only once. The visual component of the video clip was displayed at the same resolution as the screen, at a frequency of 25 images/second, whereas the auditory component was displayed at a frequency of 44,100 Hz. Infants' eye movements were recorded by a Tobii TX300 (Tobii Technology AB, Danderyd, Sweden) stand-alone eye-tracker at a sampling rate of 300 Hz.

## *Data Analysis*

To determine which part of the talker's face infants were looking at, we divided each video into three areas of interest (AOIs) for each of the two talking faces on the screen: one around the eyes, one around the mouth and one for the rest of the face. These three AOIs are represented by rectangles in *Figure 3* below.

[6] This is because, as example before, sentences were paired depending on their respective length and depending on where the break happened in the sentence.

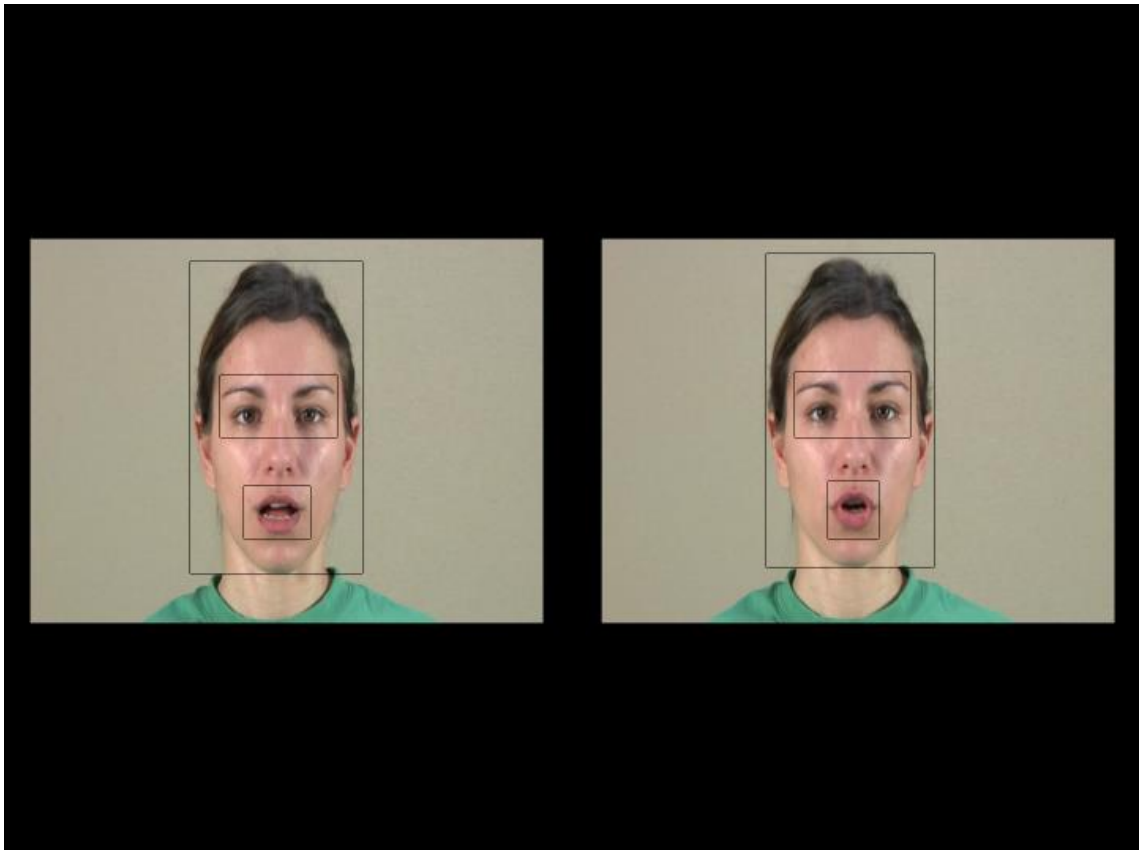

*Figure 3*. The three AOIs that were defined for analysing the eye-tracking data.

The speaker's position was rather constant across recordings; this is why we could define only one AOI for each region of interest for each sentence. Defining the AOIs was done by determining the coordinates, in pixels, of the rectangles above; we used the minimum and the maximum of all the values across all the videos as coordinates. In the videos, we caught the image in which the mouth had the largest aperture in order to define the AOImouth. Using MATLAB, we then transformed the raw data collected by the eye-tracker (coordinates in pixels) by computing whether infants looked at the defined AOIs.

# Results

## *Preprocessing and graphs*

First, we computed across all trials the proportion of total looking time (PTLT) for each AOI. To do so we divided the time spent looking at each AOI (e.g., AOI Eyes) by the sum of total looking time to all three AOIs (i.e., AOI Eyes, AOI Mouth and AOI Rest of the Face), as in Lewkowicz & Hansen-Tift (2012):

**PTLT AOIn=PTLT AOIn/(PTLT AOI1+PTLT AOI2+PTLT AOI3)**

The component ***n*** of the formula was replaced by each of our areas of interest: AOIeyes, AOImouth, AOIface.

We computed this measure separately for the matching and non-matching sides but also separating monolingual from bilingual infants. Statistical tests were done with R and Statistica[7].

## *Four months-old infants*

We computed a 2 (Bilinguals vs. Monolinguals) x 2 (Matching vs. Non-Matching) mixed ANOVA using R and Statistica on the mean PTLT. The results (cf. illustrated in *Figure 4*) show no significant main effect nor interaction between the two factors (all $Fs < 1$). Moreover, the performances were at chance level given that they did not significantly differ from 50% (all $t < 1$).

[7] Statistica 8, StatSoft Inc.

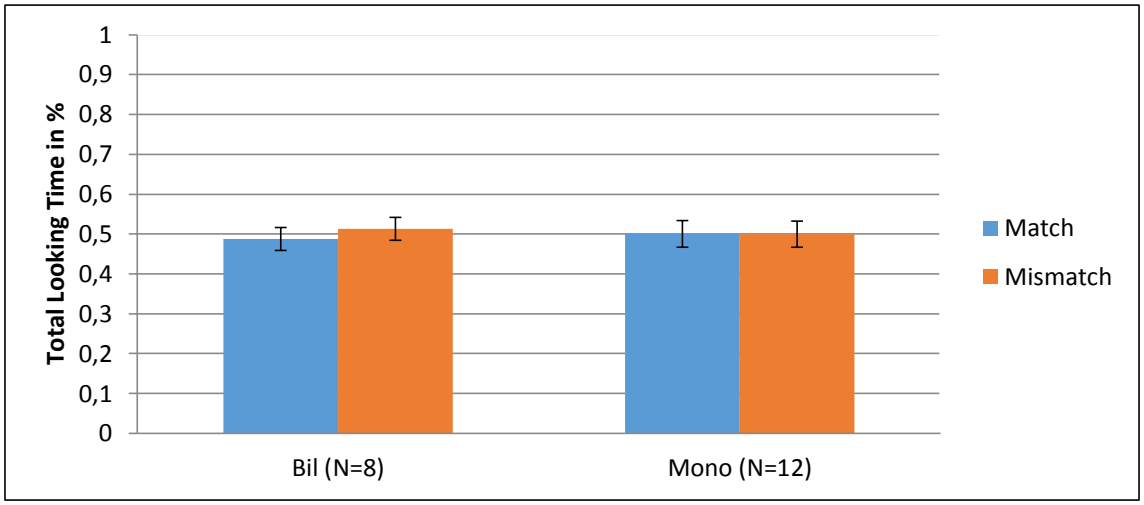


*Figure 4*. PTLT summed for all AOIs for matching and non-matching side respectively and for bilinguals (Bil) and monolinguals (Mono) respectively, for 4 months-old infants. The error bars represent standard deviation.

Then we computed a 2(Bilinguals vs. Monolinguals) x 2(Matching vs. Non-Matching) x3(AOIEyes, AOIMouth, AOIRest of the face) mixed ANOVA on the mean PTLT computed for each AOI. A main effect of bilingualism was observed ($F(1,18)=8.863$, $p=.008$), indicating that bilingual infants looked overall more than monolingual infants. There was no main effect of the Matching factor ($F<1$).The ANOVA returned a significant effect of AOI ($F(1,18)=9.159$, $p=.0006$). The interaction between AOI and bilingualism approached significance ($F(2,36)=2.674$, $p=.08$). Planned comparisons indicated that bilingual infants looked longer to the AOImouth than monolinguals ($F(1,18)=5.388$, $p=.03$). The other planned comparisons between monolingual and bilingual infants for the AOI Eyes and AOI Face did not reach significance (respectively: $F(1,18)=0.310$, $p=.584$; $F(1,18)=2.182$, $p=.157$).

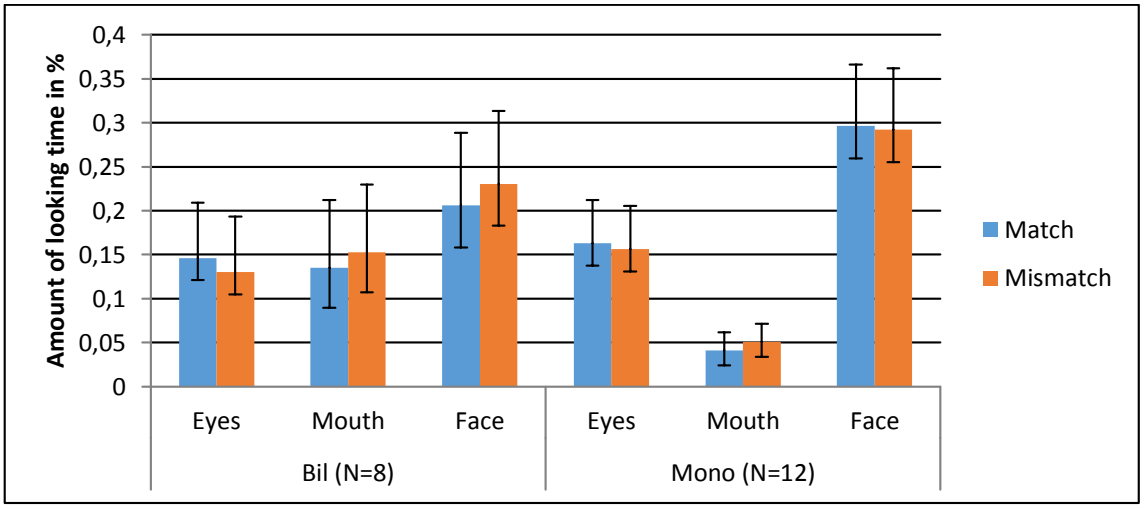


*Figure 5*. Mean PTLT for all AOIs respectively, distinguished for matching VS non-matching side, and for bilinguals (Bil) and monolinguals (Mono), for four month infants. The error bars represent standard deviation.

## *Ten months-old infants*

We computed a 2(Bilinguals vs. Monolinguals) x 2(Matching vs. Non-Matching) mixed ANOVA using R and Statistica on the mean PTLT. The results (cf. illustrated in Figure 6) show no significant main effect or interaction between the two factors (all $F<1$). Moreover, the sum of all performances is close to chance level given that it did not significantly differ from 50% (all $ts < 1$).

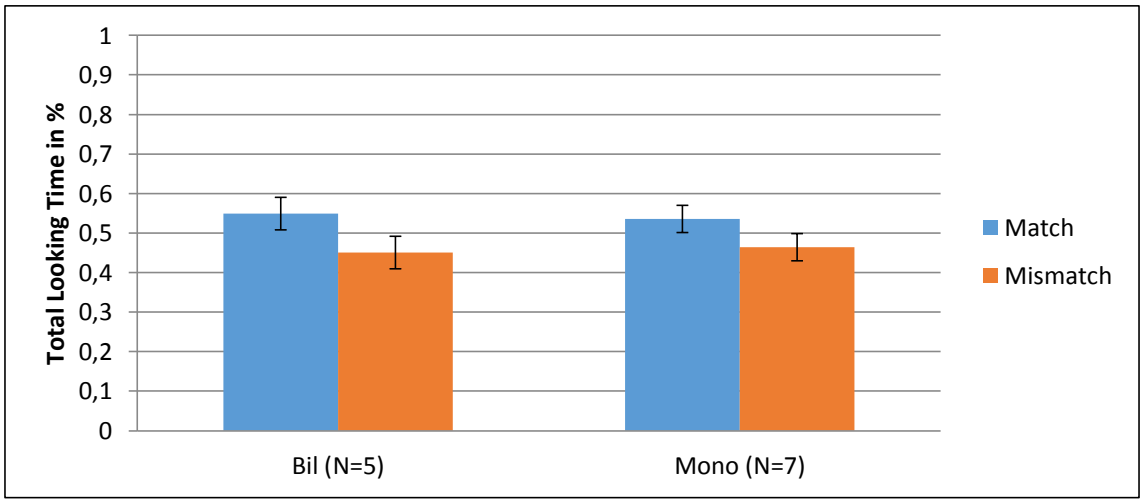


*Figure 6*. PTLT summed for all AOIs for matching and non-matching side and for bilinguals (Bil) and monolinguals (Mono), for 10 months-old infants. The error bars

represent standard deviation.

Then we computed a 2 (Bilinguals vs. Monolinguals) x 2 (Matching vs. Non-Matching) x3 (AOIEyes, AOIMouth, AOIRest of the face) mixed ANOVA on the mean PTLT computed for each AOI. The ANOVA returned that nothing reached significance (all $F<1$); the factor matching is however still close to significance ($F(1,18)$=1.948, $p$=.193).

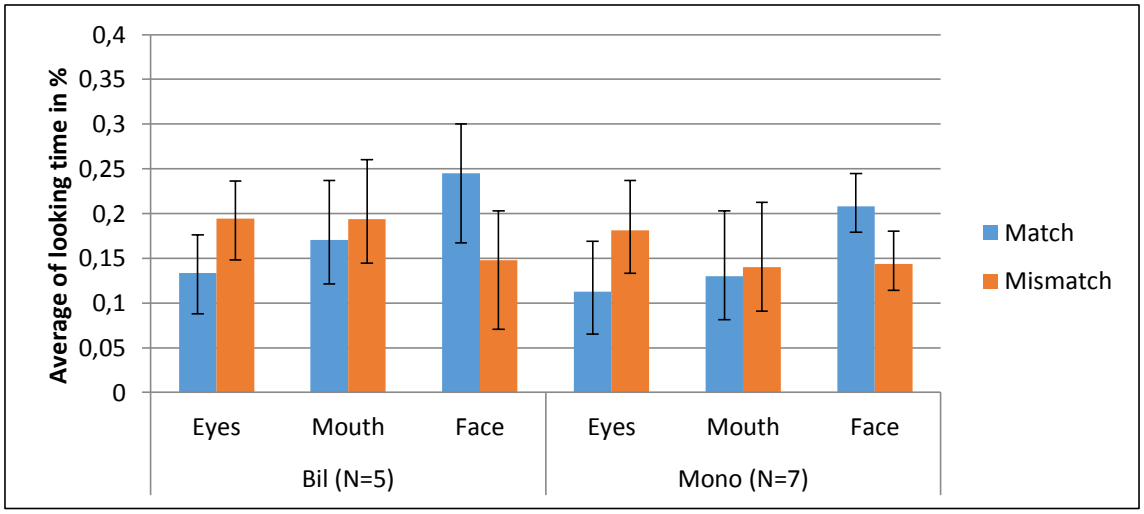


*Figure 7.* Mean PTLT for all AOIs respectively, distinguished for matching VS non-matching side, and for for bilinguals (Bil) and monolinguals (Mono), for ten months-old infants. The error bars represent standard deviation.

# Discussion

The goal of this study was to investigate how much human infants, who are in the process of learning their native language(s), rely on audiovisual speech cues to process talking faces when they are presented with a complex audiovisual situation. We examined the eye gaze of four and ten months-old monolingual and bilingual infants while they were watching side-by-side videos of a female native speaker uttering different sentences in adult-directed speech; only one matched the auditory input coming from the center of the screen.

We found that four months-old infants look at the matching and non-matching side at chance level, suggesting that, when perceiving connected speech, they are not able to identify which face is articulating the auditory sentence that they hear. This is also what Lewkowicz et al. (2015) found when they tested four months-old infants with a similar task. Testing a bigger sample could provide further evidence that while infants are able to do this task with isolated vowels by four months of age (Kuhl & Meltzoff, 1982) and as early as by two months of age (Patterson & Werker, 2003), four months-old infants are not able to match a visual sentence with its auditory counterpart (Lewkowicz et al., 2015).

Our data also indicate that bilingual infants at four months looked longer than monolingual infants at the mouth area, suggesting that they use attentional strategies that are different from those used by their monolingual peers when perceiving talking faces. Said another way, our results suggest in line with Pons et al. (2015), that as early as four months of age, being a bilingual seems to constrain in specific ways infants' attention to audiovisual talking faces. As explained by Pons et al. (2015), the fact that bilingual infants look more to the mouth area could be a strategy for keeping apart the two languages they are learning, by accessing articulatory-motor information of the acoustic sounds they hear. It therefore seems that bilingual infants, due to being faced

with two languages early on, acquire specific attentional strategies for processing articulating faces visually.

By ten months of age, we found that both bilinguals and monolinguals look at the matching and the non-matching side at chance level. This is also what Lewkowicz et al. (2015) found with ten months-old infants. Our results also indicate that bilinguals no longer look more to the mouth compared to monolinguals by that age. In terms of selective attention, this suggests that by ten months, the mouth area is used as much by bilinguals and monolinguals as an informative locus.

By ten months, we also found a trend suggesting that infants look more at the matching face. If this result is confirmed with more participants (i.e., becomes statistically significant), this could imply that by ten months, the ability to perform matching *emerges* in both bilingual and monolingual infants; indeed, matching is acquired by twelve months of age with audiovisual sentences (Lewkowicz et al., 2015)[8]. If our results are confirmed by testing a larger sample, they will need to be discussed in the light of findings on both perceptual expertise and perceptual narrowing (e.g., Pons et al., 2009, Lewkowicz & Hansen-Tift., 2012), and how these phenomena that occur between six and eleven months in infancy (Pons et al., 2009) might play a role in obtaining these results. As the perceptual system tunes to the native language(s), it is possible that audiovisual patterns of attention change, and matching becomes easier.

Two last implications are the following ones. The preliminary results of our study support that of Lewkowicz et al. (2015). If they are confirmed by further testing, we could provide evidence that the same results can be found whether using infant-direct speech (Lewkowicz et al., 2015), or adult-directed speech: the only difference with using infant-directed speech should be one of degree; as proved by Fernald (1985) and Lewkowicz & Hansen-Tift (2012) infant-directed speech is just more salient and slower and increases the overall looking time, but in principle, the same results should be observed. We could also say that no phase of familiarization – exposure to two silent faces pronouncing the sentences prior to testing (e.g. Kuhl & Meltzoff, 1982; Patterson

8 As aforementioned, matching is possible earlier on for isolated vowels, as soon as by two months of age (Patterson & Werker., 2003).

& Werker, 1999; Patterson & Werker, 2003; Kubicek, Boisferon, Dupperix, Pascalls, Loevenbruck, Gervain, Schwarzer, 2014; Lewkowicz et al., 2015) – is necessary in audiovisual studies using the Intermodal Preferential Looking procedure, at least when sentences are being seen and heard[9].

9 Isolated vowels (Kuhl & Meltzoff, 1982; Patterson & Werker, 1999; Patterson & Werker, 2003) might be processed differently as compared to sentences, and identification might fall apart without familiarization.

# Conclusion

In this preliminary study, we show that neither four nor ten months-old infants can identify the articulating face that matches one auditory sentence that is being played, which supports Lewkowicz et al. (2015)'s conclusions. However we unravel that there is a tendency for identifying the matching face by ten months, for both bilinguals and monolinguals, which could itself correspond to the emergence of matching that was found by Lewkowicz et al. (2015) at twelve months of age. Also, our four months data confirms what was found in previous studies, i.e. that bilingual infants pay more selective attention to the mouth of a speaker, and that this happens earlier than for monolinguals.

# Perspectives

If testing a bigger sample confirms the aforementioned results, one further condition that will be added is a *diverted, oblique gaze condition*. The procedure will remain the same, but instead of having the speaker's eyes open, she will have her eyes half closed and she will not look at the perceiver. By adding this condition, it will be possible to check what happens in terms of audiovisual attention when access to pieces of information from the eyes is not available, or available but in an unusual fashion. This may totally collapse patterns of selective attention. Alternatively, not being able to access the information from the eyes might engender a higher focus on the mouth of the speaker; following our hypothesis that matching side-by-side faces is easier when focusing on the mouth, this may better performances on the task. This diverted gaze condition could bring precious pieces of information because it could be proved that not looking directly at the infant while talking could disrupt learning patterns of audiovisual attention, and potentially impact further learning.

# References

Buchan, J.N, et al (2007). *Spatial Statistics of Gaze Fixations during Dynamic Face Processing*, Soc Neurosci, vol. 2, pp. 1-13.

Burnham D. and Dodd, B. (1996*). Auditory-visual speech perception as a direct process: The McGurk effect in infants and across languages*, in Speechreading by humans and machines. vol. 150, D. G. Stork and M. E. Hennecke, Eds., ed: Springer-Verlag, pp. 103-114.

Burnham, D. and Dodd, B. (2004). *Auditory-visual speech integration by prelinguistic infants: perception of an emergent consonant in the McGurk effect*. Developmental Psychobiology, vol. 45, pp. 204-220.

Bosch, L., Sebastian-Galles, N. (2011). *Evidence of early language discrimination abilities in infants from bilingual environments*. Infancy, 2(1), 29-49.

De Saint Exupéry, A. (1943). *Le Petit Prince*. Gallimard.

Everdell, I.T, et al. (2007). *Gaze Behaviour in Audiovisual Speech Perception: Asymmetrical Distribution of Face-Directed Fixations*, Perception, vol. 36, pp. 1535-45.

Fernald, A. (1985). *Four-month-old infants prefer to listen to motherese.* Infant Behavior and Development, 8, 181- 195.

Kubicek, C. et al. (2014). *Cross-Modal Matching of Audio-Visual German and French Fluent Speech in Infancy*. PLoS One, vol. 9, p. e89275, 2014.

Kuhl, P.K. and Meltzoff, A.N. (1982). *The bimodal perception of speech in infancy*. Science, vol. 218, pp. 1138-1141.

Kushnerenko, E. et al. (2008). *Electrophysiological evidence of illusory audiovisual speech percept in human infants*. Proc Natl Acad Sci U S A, vol. 105, pp. 11442-5, Aug 12.

Lewkowicz, D.J., Ghazanfar, A.A. (2006). *The decline of cross-species intersensory perception in human infants*. Proc Natl Acad Sci U S A, vol. 103, pp. 6771-4, Apr 25.

Lewkowicz, D.J., Hansen-Tift, A. (2012). *Infants deploy selective attention to the mouth of a talking face when learning speech*. PNAS, vol. 109, pp. 1431-6, Jan 31.

Lewkowicz, D.J., et al. (2015), *Perception of the multisensory coherence of fluent audiovisual speech in infancy: its emergence and the role of experience*, Journal of Experimental Child Psychology, vol. 130, pp. 147-62, Feb

Navarra, J., Soto-Faraco, S. (2007). *Hearing lips in a second language: visual articulatory information enables the perception of second language sounds*, Psychological Research, vol. 71, pp. 4-12.

Patterson, M. and Weker, J.F (1999). *Matching Phonetic Information in Lips and Voice is Robust in 4.5-month-old infants*. Infant Behavior and Development, vol. 22, pp. 237-247.

Patterson, M.L. and Werker, J.F (2002). *Infants' Ability to Match Dynamic Phonetic and Gender Information in the Face and Voice.* Journal of Experimental Child Psychology, vol. 81, pp. 93-115.

Patterson, M.L and Werker, J.F (2003). *Two-month-old infants match phonetic information in lips and voice*. Developmental Science, vol. 6, pp. 191-196.

Pons, F. et al. (2009). *Narrowing of intersensory speech perception in infancy*. Proc Natl Acad Sci U S A, vol. 106, pp. 10598-602, Jun 30.

Pons, F. et al. (2015), *Bilingualism modulates infants' selective attention to the mouth of a talking face*, Psychological Science, vol. 26, pp. 490-8, April.

Rosenblum, L.D. (2008). *Speech perception as a multimodal phenomenon*. Curr Dir Psychol Sci. 2008 Dec; 17(6): 405–409. doi: 10.1111/j.1467-8721.2008.00615.x

Ross, L.A., et al. (2007). *Do you see what I am saying? Exploring Visual Enhancement of Speech Comprehension in Noisy Environments*, Cerebral Cortex, vol. 17, pp. 1147-53.

Schwartz, J.L., et al. (2004). *Seeing to hear better: evidence for early audio-visual interactions in speech identification*, Cognition, vol. 93, pp. B69-78.

Sumby, W.H, Pollack, I. (1954). *Visual Contribution to Speech Intelligibility in Noise*, Journal of the Acoustical Society of America, vol. 26, pp. 212-215, 1954.

Yeung, H.H and Werker, J.H (2013). *Lip movements affect infants' audiovisual speech perception*. Psychol Sci, vol. 24, pp. 603-12, May.

# Appendix: Sentences

The sentences used in the experiment are adapted from the Little Prince (Spanish and Catalan). Visual sentences were paired depending on their duration and depending on where breaks happen in the sentences. Below is indicated the audiovisual duration of each audiovisual sentence. The audiovisual duration varies between 5020ms and 6040ms in the Sentences in Spanish. The audiovisual duration varies between 5010ms and 6080ms in the Sentences in Catalan.

***Sentences in Spanish***

| **Sentences in Spanish** | **Duration (ms)** | **Sentences in English** |
|---|---|---|
| Le hice comprender al niño que los pinos no son arbustos sino árboles muy grandes. | 6010 | I conveyed to the boy that pine trees are not shrubs but very high trees. |
| No he sido muy honesta al hablar de los carteros, y corro el riesgo de que no me traigan el correo. | 5210 | I have not been very honest when I talked about the postmen, and I take the risk of not having my mail delivered. |
| Es el mejor momento de mi aventura en el desierto, aunque ya me he bebido todo el agua. | 5090 | It is the best moment of my adventure in the desert, even though I have already drunk all the water. |
| Las únicas montañas que él conoce, son tres volcanes que cada mes entran en erupción. | 5210 | The only mountains that he knows are three volcanoes that erupt every month. |
| Hay que cuidar y regar las plantas regularmente, para que crezcan, y hagan flores bonitas. | 6010 | It is necessary to water and take care of the plants regularly, so that they grow and have beautiful flowers. |
| Los gatos no duermen mucho durante la noche, y prefieren ir fuera a cazar ratones. | 5140 | Cats do not sleep a lot at night, and they prefer to go out and hunt mice. |
| Le dije al niño que no había estudiado geografía, y que no sabía dibujar bien con estos lápices. | 6040 | I told the boy that I hadn't studied geography, and that I didn´t know how to draw well with these pencils. |
| Cuando uno ve una hierba mala en su jardín, es necesario quitarla inmediatamente. | 5160 | When one sees a weed in their garden, one has to pull it up immediately. |
| Si todos los habitantes del pueblo se pusieran | 5200 | If all the inhabitants of the village stood up, they |

| | | |
|---|---|---|
| de pie, cabrían fácilmente en la plaza mayor. | | could all easily fit in the plaza mayor. |
| Si los mercados estuvieran siempre abiertos, todos los días de la semana se parecerían. | 5200 | If the markets were always open, all the days of the week would look alike. |
| No tenía la apariencia de un niño perdido, mientras estaba lejos de cualquier lugar habitado. | 5140 | The boy didn´t look like he was lost, while he was far away from any inhabited place. |
| Para mi familia los jueves son un buen día, porque podemos ir de paseo hasta el pueblo. | 5180 | Thursdays are a nice day for my family, because we can have a walk to the village. |
| Si un pequeño árbol no se arranca a tiempo, no hay manera de deshacerse de él más tarde. | 5170 | If a small tree is not pulled out soon enough, it is impossible to get rid of it later. |
| Las personas mayores nunca entienden nada, y es muy aburrido tener que darles explicaciones. | 5210 | Elder people never understand anything, and it is very boring to have to explain everything to them. |
| Cogí un tren muy rápido que hacía mucho ruido cuando pasaba por en medio de la ciudad. | 5040 | I took a quick train that made a lot of noise when it went through the centre of the city. |
| Como nunca he dibujado un cordero, haré la única cosa que soy capaz de hacer. | 5020 | As I have never drawn a lamb, I will do the only thing that I am able of doing. |

***Sentences in Catalan***

| **Sentences in Catalan** | **Duration (ms)** | **Sentences in English** |
|---|---|---|
| Vaig fer entendre al nen que els pins no són arbustos si nó arbres molt grans. | 5190 | I conveyed to the boy that pine trees are not shrubs but very high trees. |
| Na he estat gaire honesta al parlar dels carters i corro el risc que no em portin el correu. | 6040 | I have not been very honest at talking about the postmen, and I take the risk of not having my mail delivered. |
| És el millor momento de la meva aventura pel desert, tot i haver begut tot l'aigua. | 5010 | It is the best moment of my adventure in the desert, even if I have already drunk all the water. |
| Les úniques muntanyes que ell coneix són tres volcans que sempre entren en erupció | 5220 | The only mountains that he knows are three volcanoes that erupt every month. |
| S'ha de cuidar i regar regularment les plantes | 5210 | It is necessary to water and take care of the |

| | | |
|---|---|---|
| perquè creixin i facin flors boniques. | | plants regularly, so that they grow and have beautiful flowers. |
| Els gats no dormen gaire per la nit I prefereixen anar fora a caçar ratolins. | 5110 | Cats do not sleep a lot at night, and they prefer to go out and hunt mice. |
| Vaig dir al nen que no havia estudiat geografia i que no sabia dibuixar bé amb aquests llapis. | 6000 | I told the boy that I hadn't studied geography, and that I didn´t know how to draw well with these pencils. |
| Quan un veu una mala herba al seu jardí és necessari treure-la ràpidament. | 6030 | When one sees a weed in their garden, one has to pull it up immediately. |
| Si tots els habitants del poble es poséssin dempeus, cabrien fàcilment a la plaça major. | 5210 | If all the inhabitants of the village stood up, they could all easily fit in the plaza mayor. |
| Si els mercats estiguéssin sempre oberts, tots els dies de la semana s'assemblarien. | 5170 | If the markets were always open, all the days of the week would look alike. |
| No tenía l'aparença d'un nen perdut mentre fos lluny de qualsevol lloc habitat. | 5180 | The boy didn´t look like he was lost, while he was far away from any inhabited place. |
| Per la meva família els dijous són un bon dia perquè podem anar de passeig fins al poble. | 5180 | Thursdays are a nice day for my family, because we can have a walk to the village. |
| Si un arbre petit no s'arranca a temps, no hi ha manera de desfer-se d'ell més tard. | 5180 | If a small tree is not pulled out soon enough, it is impossible to get rid of it later. |
| Las persones grans mai entenen res i és molt avorrit haver de donar-los explicacions. | 6080 | Elder people never understand anything, and it is very boring to have to explain everything to them. |
| Vaig agafar un tren ràpid que feia molt soroll quan passava pel mig de la ciutat. | 5090 | I took a quick train that made a lot of noise when it went through the centre of the city. |
| Com mai he dibuixat un xai, faré l'única cosa que sóc capaç de fer. | 5010 | As I have never drawn a lamb, I will do the only thing that I am able of doing. |