Journal of Experimental Psychology:
Human Perception and Performance
2016, Vol. 42, No. 9, 1275–1281

0096-1523/16/$12.00 http://dx.doi.org/10.1037/xhp0000208

Newborns' Sensitivity to the Visual Aspects of Infant-Directed Speech: Evidence From Point-Line Displays of Talking Faces

Bahia Guellaï
Université Paris Ouest Nanterre La Défense

Arlette Streri
Université Paris Descartes

Adrien Chopin
University of California, and Ecole Normale Supérieure & CNRS, Paris, France

Delphine Rider
Université Paris Descartes

Christine Kitamura
University of Western Sydney

The first time a newborn is held, he is attracted by the human's face. A talking face is even more captivating, as it is the first time he or she hears *and* sees another human talking. Older infants are relatively good at detecting the relationship between images and sounds when someone is addressing to them, but it is unclear whether this ability is dependent on experience or not. Using an intermodal matching procedure, we presented newborns with 2 silent point-line displays representing the same face uttering different sentences while they were hearing a vocal-only utterance that matched 1 of the 2 stimuli. Nearly all of the newborns looked longer at the matching point-line face than at the mismatching 1, with prior exposure to the stimuli (Experiment 1) or without (Experiment 2). These results are interpreted in terms of newborns' ability to extract common visual and auditory information of continuous speech events despite a short experience with talking faces. The implications are discussed in the light of the language processing and acquisition literature.

Keywords: newborns, audiovisual speech perception, multimodality, visual prosody

Speech perception is a multimodal experience: it does not only involve audition but also vision, to the extent that seeing someone talking influences adults' (McGurk & MacDonald, 1976) and infants' (Rosenblum, Schmuckler, & Johnson, 1997) interpretation of speech. Infants have a relatively robust capacity to detect congruencies between seen and heard speech events: they match vowel information in lips and voice (Kuhl & Meltzoff, 1982; Patterson & Werker, 2003) and learn phonetic contrasts better in auditory-visual than auditory-only conditions (Teinonen, Aslin, Alku, & Csibra, 2008). Speech, however, is more than phonetic segments: it is continuous and includes important prosodic markers, such as rhythm, intonation, and stress. Adults use these markers to parse the speech stream, and to get access to the semantic and pragmatic linguistic levels. Recent studies (Lewkowicz, Minar, Tift, & Brandon, 2015) provided evidences that infants are able to match a sentence to the corresponding talking face, both in native and nonnative languages by the end of their first year. Eight-month-old infants reliably detect congruency between matching auditory and visual point-line displays of a talking face, mainly based on prosodic information (Kitamura, Guellaï, & Kim, 2014). In other words, infants are able to perceive the correspondences between auditory and visual speech information before understanding the meaning of words. Daily exposition to talking faces may have affected the development of this ability. This is why we tested in the present study the possible role of experience. We presented newborns with audiovisual speech events to test their ability to match a sentence to the corresponding talking face.

Talking faces are one of the first objects that infants encounter when they enter the world. Newborns are familiar with auditory speech because the fetus hears (mainly prosodic) speech informa-

This article was published Online First April 28, 2016.

Bahia Guellaï, Laboratoire Ethologie, Cognition, Développement, Université Paris Ouest Nanterre La Défense; Arlette Streri, Laboratoire Psychologie de la Perception, Université Paris Descartes; Adrien Chopin, Department of Psychology, University of California, and Laboratoire des systèmes perceptifs, Ecole Normale Supérieure & CNRS, Paris, France; Delphine Rider, Laboratoire Psychologie de la Perception, Université Paris Descartes; Christine Kitamura, School of Social Sciences and Psychology, University of Western Sydney, and MARCS Institute, University of Western Sydney.

This study obtained the authorization of the Graduate School of Paris Descartes as well as the written consent of the nursing staff of the Bichat Maternity Hospital and the parents. It was funded by the French Ministry of Research and Institut Universitaire de France. We thank the parents and infants who participated in the study as well as Pr. Luton, Laetitia de Lorgeril, and the nursing staff at the Bichat Maternity Hospital for their collaboration. We thank also Chris Davis and Pauline Vabre for technical support.

Correspondence concerning this article should be addressed to Bahia Guellaï, Laboratoire Ethologie, Cognition, Développement, Université Paris Ouest Nanterre La Défense, 200, avenue de la République, 92000 Nanterre, France. E-mail: bahia.guellai@gmail.com

tion through the low-pass filter of their mother's womb, starting from around 24 weeks of gestation (Querleu, Renard, Versyp, Paris-Delrue, & Crèpin, 1988). This experience underpins their ability to recognize their native language structure (Mehler et al., 1988) and their increased sensitivity to native language rhythm (Nazzi, Jusczyk, & Johnson, 2000). Exposure to visual speech, however, begins at birth when the newborn can see *and* hear a person talking for the first time. Remarkably, they can match the lip image and sound of a vowel when presented in isolation (Aldridge, Braga, Walton, & Bower, 1999) and when a monkey vocalizes (Lewkowicz, Leo, & Simion, 2010). However, the singular audiovisual events that have been used so far for matching experiments do not contain the complexities of continuous speech.

Indeed, a whole sentence is exceedingly rich in information. The conveyed message forms a system in which multiple correlated inputs contribute to the interpretation of a speaker's message in diverse overlapping ways. In that sense, visual speech information is not confined to articulatory gestures from the mouth and lips. For example, features such as head motion also play a role in the intelligibility of expressive speech (Cvejic, Kim, & Davis, 2010; Davis & Kim, 2006). The movements produced during continuous speech include both rigid motion (i.e., from the whole head) such as up-down and side-to-side movements; and nonrigid motion (i.e., from the internal facial features) that change the spatial configuration of the facial surface over time (Roark, Barrett, Spence, Abdi, & O'Toole, 2003). Despite a weak visual system (Allen, Tyler, & Norcia, 1996), newborns are highly sensitive to motion (Otsuka & Yamaguchi, 2003). At birth, the presence of both rigid and nonrigid motion in a talking face enhances the recognition of interactive faces (Guellaï, Coulon, & Streri, 2011) whereas rigid head motion alone is an important source of information for newborns' recognition of photographs of faces (Bulf & Turati, 2010). Still, it is not known whether newborns can link facial motion cues to the corresponding auditory speech when seeing someone talking.

The present study addressed this question using faces talking in an infant-directed speech style. Indeed, when addressing to infants, adults usually use a different speech register than the one they use with adults. This particular speech register has been studied broadly in the auditory domain in the past decades: it is steeped with prosodic, phonetic, and affective exaggeration (Kitamura & Burnham, 2003). This particular speech register is preferred by newborns over normal speech from another adult (Cooper & Aslin, 1990; Fernald, 1985). Interestingly, specific characteristics of infant-directed speech have also been observed in the visual domain. The very few studies present in the literature on this topic showed that when addressing to an infant, adults exaggerate their facial movements (Chong, Werker, Russell, & Carroll, 2003; Green, Nip, Wilson, Mefferd, & Yunusova, 2010; Smith & Strader, 2014). The particular properties of the infant-directed speech seem to be universal across human cultures and some authors claim that the acoustical and visual properties of the infant-directed speech attract infants' attention to the speaker and help them to parse the speech stream (Kitamura & Burnham, 2003). A first step to test this possibility would be to see if infants are able to recognize similarities present in the exaggerated auditory and visual characteristics of the infant-directed speech.

To our knowledge, two studies explored the question (Kitamura et al., 2014; Lewkowicz et al., 2015) through an intermodal matching paradigm. In Lewkowicz and colleagues' study (2015), infants 4 to 14 months old were presented with the same face uttering one sentence in a native and in a nonnative language while only one of the two speech streams was heard. The authors observed that only 12- to 14-month-old infants are able to match what they see to what they hear. In Kitamura and colleagues' study (2014), the authors presented 8-month-old infants with two visual displays of the same talking face and an auditive utterance matching only one of the two displays. The stimuli used in this study consisted of point-line displays derived from the head and face motion of three women recorded while reading short sentences to their 8-month-old babies (those babies were not included in the final sample). The use of such stimuli instead of normal faces allowed for enhancing attention to the rigid and nonrigid movements of the face. Besides, it helps controlling potentially distracting information such as form, texture, and more importantly, distraction from the eyes' region (Guellaï & Streri, 2011). Moreover, point-line rather than point-light displays were used to provide enough coherent information given the relatively sparse motion captors used in the displays. It is also important to note that the placement of the mouth markers on the outer edges of the lips deemphasizes the contrastive details of consonant articulation and tends to emphasize the rhythmic cycling of mouth movements. Infants reliably detected auditory and visual congruencies in the displays even when only the suprasegmental information (i.e., the speech sound was low-filtered) of the sentences are provided, and also when only rigid motion of the head was presented (Kitamura et al., 2014). Therefore, before the end of their first year, infants are already able to match a sentence to the corresponding facial movements of a talking face, and they do so even with prosodic cues only. These results are different from those of Lewkowicz and colleagues (2015), which is possibly related to the presence of the eyes in this latter study. Indeed, the presence of direct gaze can interfere with audiovisual speech processing in younger infants (Guellaï & Streri, 2011).

In these previous studies, infants were tested during the second half of their first year: they already had experienced talking faces and especially faces addressing to them using the infant-directed speech. One could ask to what extent is this ability linked to experience with talking faces? One way to answer this question is to test infants with very little audiovisual experiences such as during the neonatal period.

Newborns bring to the world biases that underpin perceptual experience and guide learning. They are attracted to faces (Goren, Sarty, & Wu, 1975), and even more to a talking face (Coulon, Guellaï, & Streri, 2011; Guellaï & Streri, 2011). They also find auditory speech more interesting than complex nonspeech analogues (Vouloumanos & Werker, 2007). These findings are indicative of a system that guides infant's attention to stimuli that advantage learning and development. Newborns are also capable of detecting redundant information across modalities, for example, recognizing the shape of a previously felt object (Streri & Gentaz, 2003). This implies that infants are capable of intersensory organization very early in life (Bahrick & Lickliter, 2003; Gibson, 1984). Furthermore, recent evidence indicates that intersensory perception is initially very broadly specified and narrows over the following months (Lewkowicz & Ghazanfar, 2006; Pons, Lewkowicz, Soto-Faraco, & Sebastián-Gallés, 2009). Perceptual processing at very young ages is said to be based on sensitivity to low-level auditory-visual relations, and at this level, specified by attention to synchronous onsets and offsets or to the temporal synchrony between auditory and visual events (Lewkowicz, 2000). Studies using vocal events that support the role of temporal synchrony typically showed that infants look longer at a face/voice set when face and voice are in synchrony than when they are out of synchrony (Blossom & Morgan, 2006; Dodd, 1979; Lewkowicz et al., 2010). However, the out of synchrony condition of these experiments featured auditory and visual stimuli ending at different times while they are presented side-by-side. This detection of in- versus out-of-phase audiovisual events is more straightforward, we suggest, than detecting a match between a continuous sentence and its visual counterpart using the intermodal preference procedure. In that procedure, infants are forced to make a choice between two visual stimuli that start and end at the same time while listening to an auditory input.

In the present study, we tested whether newborns can already use the rigid and nonrigid features of a talking face to match what they see to what they hear.

Experiment 1

Participants

Participants were 24 full-term newborns (13 males), all in good health (Apgar scores > 9). They were recruited and tested in a maternity hospital in France. The mean age was 49 hr (*SE* = 4.21; range 10.5–83.5 hr). Additional infants were excluded for irritability (6), falling asleep (2), technical problems (2), or because they did not look at one of the two stimuli during the familiarization (1). Newborns were from French monolingual families (9), from other language monolingual families (2), or from bilingual French/other language families (13).

Recording of Audiovisual Stimuli

Three women recorded sentences in Australian English addressed to their infants using a Northern Digital Optotrak 3020 system. In our study, the presentation of faces talking a nonnative language is not problematic, given that newborns process native and nonnative talking faces in the same way (see Guellaï, Mersad, & Streri, 2015). A Northern Digital Optotrak 3020 system and a Sennheiser MKH416P48U-3 floor microphone were used to record the stimuli. Each female face (*n* = 3) was a mother who sat facing her 8-month-old infant during the recording session. She was instructed to speak a list of sentences as she would if she were reading to her infant at home. Optotrak markers were attached to the female's face-internal features and to a head rig to measure rigid head movement. There were 21 nonrigid face internal markers on the eyebrows (4), cheeks (6), mouth (8), and jaw (3); and 3 markers on the head rig. Each frame of speech movement was extracted from the raw marker positions and represented in terms of its displacement from the first frame. Data reduction was achieved by using principle components (PC) analysis; the PC's coefficients were used to generate point-line visual speech motion. An animated schema of the face was generated by displaying the PC's coefficients generated in Matlab using a marker visualization tool. Rendering was achieved in Togl and OpenGL, and each frame was written to an image file and compiled to a video sequence at 25 frames per second. This gave a line-joined version of visual speech. The faces depicted rigid (i.e., the movement of the whole head) and nonrigid (i.e., the movement of face-internal features) motion but no fine articulatory details, since the motion detection markers were placed on the outer edge of the lips. Such a sparse representation was chosen to emphasize rhythmic speech motion and the perception of simple onsets and offsets. Nevertheless, the motion of the display did correlate with auditory properties as measured over a rendition. We measured how changes in speech amplitude (root mean square [RMS] energy) and mean F0 correlated with the rigid motion of the display and with the motion of the eyebrows, mouth (lip height and width), and jaw. For each of the sentences, at least one correlation coefficient was 0.45 or greater indicating that there was always a relationship between some of the auditory and visual features. According to the results of the Kitamura and colleagues' study (2014), this particularity of the stimuli did not impact infants' ability to match what they heard to what they saw. Therefore, the same audiovisual speech events (i.e., point-line displays of talking faces with both rigid and nonrigid motion) were presented to newborns in the present study.

Sentence Stimuli

The same stimuli as those used in Kitamura and colleagues study (2014) were presented to newborns. The speech stimuli consisted of three pairs of sentences, one pair from each of the three female talkers (see Table 1). The sentences were adapted from the IEEE list of phonetically balanced sentences and were "Yum, clams are round, small, soft and tasty" (9 syllables); "Look, the big red apple fell to the ground" (10 syllables); "It's tea time, let's feed the white mouse some flower seeds" (12 syllables); and the utterance "Did you know, Woolly is a sheep" (8 syllables). Adaptations to facilitate a more infant-directed style to the sentences were for instance, adding the word "yum" and "look" to the beginning of the utterances. From the list of sentences for each of the mother, pairs of sentences were selected that matched in duration but differed in syllable number. For some of the original sentence pairs, there were small differences in duration (<500 ms). They were corrected by slightly accelerating the longer utterance and slightly slowing the shorter utterance using Adobe Premier. The sentences and syllable differences for each mother are shown in Table 1.

Table 1
Sentence Pairs (A and B) for the Three Mothers With the Syllable Number of Each Sentence in Parentheses

Mother	Sentence A (syllable n°)	Sentence B (syllable n°)	Syllable diff	Duration (secs)
1	Sheep (8)	Apple (10)	2	4
2	Clams (9)	Mouse (12)	3	6
3	Sheep (8)	Mouse (12)	4	5

Note. Each pair differed in syllable number but was matched for duration. The sentences were: "Did you know, Woolly is a sheep" (sheep); "Yum, clams are round, small, soft, and tasty" (clams); "Look, the big red apple fell to the ground" (apple); and "It's tea time, let's feed the white mouse some flower seeds" (mouse).

Apparatus

Before testing, parents and medical staff gave their informed consent. Newborns were observed in a quiet room inside the maternity ward of the hospital. They were positioned in a semiupright position (30°) in an adapted baby seat. The seat was placed on a table facing a 19-in. color monitor 35 cm from the infants' eyes. The screen presented the two visual displays that subtended a visual angle of 16° ×14°, and the two images were 12 cm apart from each other. Two loudspeakers were placed on each side of the monitor facing the baby and the auditory stimuli were played at ~65 dBA SPL. A digital video camera placed midline above the screen and facing the infant transmitted its output to a video monitor, which was used by Experimenter 1, blind to the conditions, to judge the infant's eye movements (left, right, or not looking). Experimenter 2 stood behind the newborn during testing to monitor any signs of discomfort.

Testing Procedure

The same procedure as the one used with older infants in the Kitamura and colleagues' study (2014) was used here with some minor differences such as less test trials and less repetitions of the stimuli per trial, due to newborns' smaller attentional capacities.

Each infant was tested with one of the three talkers to avoid any bias due to the potential of dynamic signatures differing between speakers (see Soken & Pick, 1992). An example of a pair of the visual displays is shown in Figure 1. The infant's gaze was centered at the start of all trials by the mean of a looming red ball presented on the middle screen. Because the visual stimuli were unfamiliar point-line displays, we initially played the infants two audiovisual repetitions of each sentence on the left side of the screen, and two repetitions of each sentence on the right side of the screen. The order of presentation of the displays was counterbalanced across subjects. Then, in a test phase, the auditory-only target sentence was played through the loud-speakers with the two silent point-line displays presented simultaneously on the left and right side of the monitor. There were three test trials that lasted for six repetitions of the sentence. The side of presentation of the visual displays and the target sentence were counterbalanced across infants. Sides of presentation of the in-phase visual stimulus remained constant across trials. In total, there were eight possibilities for each of the three faces (i.e., four for the familiarization and four for the test). We counterbalanced the target sentence so that half of the infants heard one sentence in the pair and the other half heard the other sentence in the pair. The target sentence was also counterbalanced so that it was displayed on the left or right side equally. The MatLab sofware was used to program the sequencing of the experiment and to record newborns' looking times in real time. Two experimenters, blind to the conditions, coded newborns' looking behaviors (left, right, or not looking). Experimenter 1 coded in real time, and Experimenter 2 coded offline. Intercoder reliability between the two independent coders for the first experiment was calculated on 67% of the infants (Pearson's $r = .90$, $p < .05$, for duration of looking; Pearson's $r = .92$, $p < .05$, for the first look; and Pearson's $r = .92$, $p < .05$, for the latency of the first look).

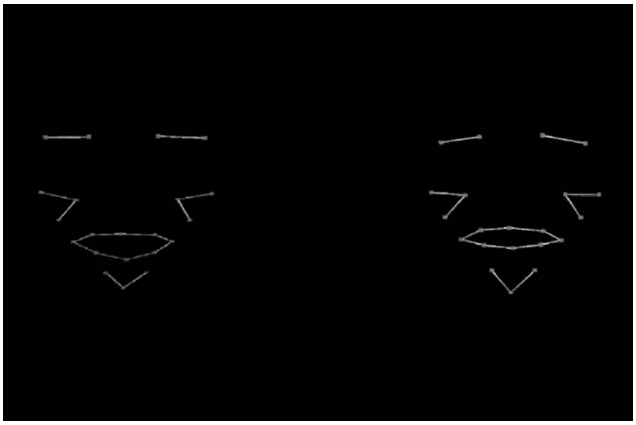

Figure 1. A captured image of the side-by-side videos of talking faces represented by point-line displays and presented to the newborns in Experiment 1 and Experiment 2.

Results

The extent to which newborns fixated the matching visual display was used to index the degree to which they detected congruency between the auditory and visual speech signals. Because of each sentence pair (1, 2 and 3) having different durations (see Table 1), the dependent variable was the fixation duration as a percentage of the total trial duration for each test sentence pair. Assumptions of normality were met, with both Shapiro-Wilk and Kolmogorov–Smirnov tests being nonsignificant for all the three experiments presented here (p values > 0.1). Preliminary analyses of variance (ANOVAs) testing the effects of the three counterbalancing variables (i.e., (i) order of presentation during familiarization; (ii) left or right presentation of visual target sentence; and (iii) target auditory sentence 1 or 2 in pair) were nonsignificant. We report partial eta^2 for effect size.

Of the 24 newborns tested, 22 spent a greater percentage of time fixating the matching than the mismatching display. The difference between proportions of time watching the matching ($M = 59.60\%$; $SE = 1.88$) and mismatching display ($M = 27.82\%$; $SE = 2.63$) was significant (Student t test $t_{(23)} = 8.11$, $p < .001$; see Figure 2). A repeated-measure ANOVA testing syllable difference (2, 3, or 4) between participants, and matching/mismatching within participants revealed a main effect of matching/mismatching: newborns looked more at the matching than at the mismatching display, $F_{(1,21)} = 64.55$, $p < .0001$; $\eta_p^2 = 0.76$. No other effect or interaction was significant. Moreover, we observed that percentages of looking time to the matching display increased across the three test trials (first trial = 43.44%, second trial = 57.01%, third trial = 78.35%). The difference of looking time to the matching display was significant between the first and third trial ($t_{(23)} = -4.98$, $p < .0001$). We analyzed also the percentages of first look to the displays and observed that they were more important for the matching ($M = 75\%$, $SE = 5.40$) than for the mismatching displays ($t_{(23)} = 4.63$, $p < .001$). Finally, we analyzed the latencies of the first look to the matching display compared with the mismatching one and observed that they were shorter for the matching ($M = 3.63$ s, $SE = 0.90$) than for the mismatching ($M = 10.24$ s, $SE = 1.32$) displays across the three test trials ($t_{(23)} = -4.12$, $p < .001$; see Figure 2).

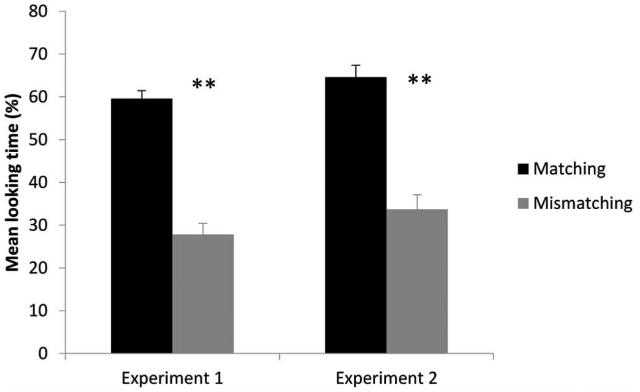

Figure 2. Mean percentage looking time to matching and mismatching point-line displays by newborn groups with familiarization and without familiarization. Error bars represent standard error. ** $p < .001$.

The results of Experiment 1 revealed that newborns are sensitive to the congruency between the auditory characteristics of infant-directed speech and its corresponding face and head movements: (1) They spent more time looking at the matching than at the mismatching display; (2) they direct their first look more often to the matching than to the mismatching display; and (3) latencies of this first look are shorter for the matching than for the mismatching stimulus. This result patterns are constant despite the fact that different talking women were presented. While results of Experiment 1 provide evidence that even newborns, with only little exposure to talking faces, are able to match a sentence to its visual counterpart, it is possible that exposition to the familiarization audiovisual stimuli prior to the test session have provided enough experience for the newborns to detect the auditory-visual relations in the test trials. Experiment 2 was designed to test this possibility.

Experiment 2

Participants

Participants were 12 full-term newborns (3 males), all in good health (Apgar scores > 9). The mean age was 35 hr ($SE = 4.08$; range 14–53 hr). Additional infants were excluded for irritability (3), or falling asleep (2). Newborns were from French monolingual families (7), or from bilingual French/other language families (5).

Stimuli and Apparatus

The stimuli and apparatus were the same as those used in Experiment 1.

Testing Procedure

The procedure was the same as in Experiment 1 with the difference that there was no familiarization with the audiovisual stimuli prior to the test phase. Newborns were presented with the two visual point-line displays belonging to the same woman uttering different sentences and only one of two corresponding utterances. There was four possibilities for each of the three faces. Again, the two experimenters coded newborns' looking behaviors. Intercoder reliability between the two independent coders for this second experiment was calculated again on 67% of the sample (Pearson's $r = .88$, $p < .05$, for duration of looking; Pearson's $r = .90$, $p < .05$, for the first look; and Pearson's $r = .90$, $p < .05$, for the latency of the first look).

Results

Of the 12 newborns, 11 looked significantly longer at the matching point-line display than at the mismatching one and the difference between time spent watching the matching ($M = 64.59\%$; $SE = 2.77$) and the mismatching one ($M = 33.72\%$; $SE = 3.42$) was significant, $t(11) = 6.13$, $p < .0001$. The ANOVA revealed a main effect of matching/mismatching, $F(1, 9) = 11.45$, $p < .01$; $\eta_p^2 = 0.56$, confirming that newborns detect congruency in auditory and visual displays using point-line displays of visual speech motion even with no familiarization period (see Figure 2). No other effect or interaction was significant. Again, percentages of looking time to the matching display increased across the three test trials (first trial = 51.81%, second trial = 64.51%, third trial = 77.46%). The difference of looking time to the matching display was significant between the first and third trial, $t_{(11)} = -4.11$, $p < .01$. This result is congruent with results of the previous group of newborns tested and could be interpreted in terms of infants being better at tracking the congruent display across trials.

Results of this second experiment confirmed that newborns are able to match a continuous sentence to the corresponding rigid- and nonrigid motion from point-line displays of talking faces, even with no prior exposure to the stimuli.

Discussion

The present study explored the sensitivity of infants, only hours after birth, to the relations between continuous auditory speech and its visual counterpart. The findings reveal that newborns recognize the characteristics of an auditory sentence presented with both rigid and nonrigid motion of the face and head: in Experiment 1 and Experiment 2, nearly all of the newborns (a total of 33 out of 36) preferentially fixated the matching talking face. Even without familiarization, newborns spontaneously oriented their gaze more toward the display that matched the auditory sentence than toward the nonmatching one. Evidently, the ability to recognize the relations between auditory and visual speech motion does not depend on experience. We also show that the involved cognitive process is rapid: the newborn's behavior differs between the congruent and incongruent displays as early as the first look, which is directed more often and with a shorter latency toward the congruent display than toward the noncongruent one.

Given the coarse grain nature of both the rigid and nonrigid motion in the auditory visual displays, the results suggest that newborns rely rather on the rhythmic qualities of the voice and face-head kinematics to perform the task. This is consistent with the idea that prosodic cues, such as rhythm and intonation are not specific to the auditory modality. For instance, adults perceive prosody from eyebrow movements (Krahmer & Swerts, 2004), hand gestures (Guellaï, Langus, & Nespor, 2014), and head movements of adult speakers (Cvejic et al., 2010). In addition, newborns have prenatal experience with auditory prosody (Querleu et al., 1988), and they rely on this experience to recognize their native language (Mehler et al., 1988). Newborns are likely to use the auditory information to detect the correlations between the voice and face/head kinematics.

Our results suggest that infants come to the world perceptually tuned to perceive the relations between complex auditory and visual events. Nonetheless, it is possible that even a few hours of exposure to talking faces primes this capacity. In addition, it is possible that more general capacities based on low-level correspondences drive this early audiovisual speech matching ability. Indeed, it is possible that manipulating the level of perceptual information available, such as the low-level spatiotemporal correspondences, would yield a different conclusion. Newborns may experience other intersensory events that could influence the development of audiovisual speech processing. For instance, sensorimotor stimulation begins in utero (see Reissland & Francis, 2010) and could be an important precursor to later intermodal speech perception. A recent study showed that 4.5-month-old infants' oral behaviors, such as chewing and sucking, which are also observed in fetuses, can influence infants' audiovisual speech perception abilities (Yeung & Werker, 2013). Another source of sensorimotor experience that might entrain newborns' robust preference for rhythmically aligned auditory-visual speech is that newborns have experienced prenatal inputs such as the alignment of mothers' body movements with her speech flow. Indeed, speech production involves the whole body, as not only the face and head moves but also, for example, the hands and the arms (Guellaï et al., 2014). This would suggest that newborn intermodal speech perception is nurtured by experiences of other types of in utero sensorimotor matching.

Another aspect of our findings is that the ability to detect relations between auditory and visual speech information appears to be quite broad. Here, French born infants with a variety of non-English language backgrounds were able to match what they see to what they hear, even when they were presented with auditory-visual displays presented in English. This is in accordance with other evidence showing auditory-visual and indeed, visual-only speech perception is initially broad and narrows over the first year (Weikum et al., 2007). Newborns exhibit cross-species intersensory matching for monkey calls (Lewkowicz et al., 2010) but lose the ability before the age of 8 months (Lewkowicz & Ghazanfar, 2006). Similarly, intersensory matching of nonnative phonetic contrasts occurs at 6 months, but dissipates by 11 months of age (Pons et al., 2009).

The implications of this study are twofold. First, the results highlight the possibility that infant-directed speech, with its exaggerated dynamics in the auditory and visual domain, is already perceived crossmodaly at birth. Second, the existence of such a bias to match auditory and visual speech information in the first hours after birth might play a foundational role in the development of later language acquisition. Future studies will examine the specificity of the ability to detect spatiotemporal correspondences in the case of audiovisual speech processing. This will help understanding the developmental trajectory of auditory visual speech perception and the interactions between auditory and motor coordination in speech acquisition.

References

Aldridge, M. A., Braga, E. S., Walton, G. E., & Bower, T. G. R. (1999). The intermodal representation of speech in newborns. *Developmental Science, 2,* 42–46. http://dx.doi.org/10.1111/1467-7687.00052

Bahrick, L. E., & Lickliter, R. (2003). Intersensory redundancy guides early perceptual and cognitive development. *Advances in Child Development and Behavior, 30,* 153–187. http://dx.doi.org/10.1016/S0065-2407(02)80041-6

Blossom, M., & Morgan, J. (2006). *Does the face say what the mouth says? A study of infants' sensitivity to visual prosody*. Paper presented at the 30th Annual Boston University Conference on Language Development.

Bulf, H., & Turati, C. (2010). The role of rigid motion in newborns' face recognition. *Visual Cognition, 18,* 504–512. http://dx.doi.org/10.1080/13506280903272037

Chong, S. C. F., Werker, J. F., Russell, J. A., & Carroll, J. M. (2003). Three facial expressions mothers direct to their infants. *Infant and Child Development, 12,* 211–232. http://dx.doi.org/10.1002/icd.286

Cooper, R. P., & Aslin, R. N. (1990). Preference for infant-directed speech in the first month after birth. *Child Development, 61,* 1584–1595. http://dx.doi.org/10.2307/1130766

Coulon, M., Guellaï, B., & Streri, A. (2011). Recognition of unfamiliar talking faces at birth. *International Journal of Behavioral Development, 35,* 282–287. http://dx.doi.org/10.1177/0165025410396765

Cvejic, E., Kim, J., & Davis, C. (2010). Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion. *Speech Communication, 52,* 555–564. http://dx.doi.org/10.1016/j.specom.2010.02.006

Davis, C., & Kim, J. (2006). Audio-visual speech perception off the top of the head. *Cognition, 100,* B21–B31. http://dx.doi.org/10.1016/j.cognition.2005.09.002

Dodd, B. (1979). Lip reading in infants: Attention to speech presented in- and out-of-synchrony. *Cognitive Psychology, 11,* 478–484. http://dx.doi.org/10.1016/0010-0285(79)90021-5

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior & Development, 8,* 181–195. http://dx.doi.org/10.1016/S0163-6383(85)80005-9

Gibson, E. J. (1984). Shedding the light of the ecological approach on differentiation and enrichment. In M. E. Lamb, A. L. Brown, & B. Rogoff (Eds.), *Advances in development* (Vol. 3). Hillsdale, NJ: Erlbaum.

Goren, C. C., Sarty, M., & Wu, P. Y. (1975). Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics, 56,* 544–549.

Green, J. R., Nip, I. S., Wilson, E. M., Mefferd, A. S., & Yunusova, Y. (2010). Lip movement exaggerations during infant-directed speech. *Journal of Speech, Language, and Hearing Research, 53,* 1529–1542. http://dx.doi.org/10.1044/1092-4388(2010/09-0005)

Guellaï, B., Coulon, M., & Streri, A. (2011). The role of motion and speech in face recognition at birth. *Visual Cognition, 19,* 1212–1233. http://dx.doi.org/10.1080/13506285.2011.620578

Guellaï, B., Langus, A., & Nespor, M. (2014). Prosody in the hands of the speaker. *Frontiers in Psychology, 5,* 700.

Guellaï, B., Mersad, K., & Streri, A. (2015). Suprasegmental information affects processing of talking faces at birth. *Infant Behavior and Development, 38,* 11–19.

Guellaï, B., & Streri, A. (2011). Cues for early social skills: Direct gaze modulates newborns' recognition of talking faces. *PLoS ONE, 6*(4), e18610. http://dx.doi.org/10.1371/journal.pone.0018610

Kitamura, C., & Burnham, D. (2003). Pitch and communicative intent in mothers' speech: Adjustments for age and sex in the first year. *Infancy, 4,* 85–110. http://dx.doi.org/10.1207/S15327078IN0401_5

Kitamura, C., Guellaï, B., & Kim, J. (2014). Motherese by eye and ear: Infants perceive visual prosody in point-line displays of talking heads. *PLoS ONE, 9,* e111467. http://dx.doi.org/10.1371/journal.pone.0111467

Krahmer, E. J., & Swerts, M. (2004). More about brows: A cross-linguistic analysis-by-synthesis study. In C. Pelachaud & Zs. Ruttkay (Eds.), *From brows to trust: Evaluating embodied conversational agents* (pp. 191–216). Amsterdam: Kluwer Academic Publishers. http://dx.doi.org/10.1007/1-4020-2730-3_7

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science, 218,* 1138–1141. http://dx.doi.org/10.1126/science.7146899

Lewkowicz, D. J. (2000). The development of intersensory temporal perception: An epigenetic systems/limitations view. *Psychological Bulletin, 126,* 281–308. http://dx.doi.org/10.1037/0033-2909.126.2.281

Lewkowicz, D. J., & Ghazanfar, A. A. (2006). The decline of cross-species intersensory perception in human infants. *Proceedings of the National Academy of Sciences of the United States of America, 103,* 6771–6774. http://dx.doi.org/10.1073/pnas.0602027103

Lewkowicz, D. J., Leo, I., & Simion, F. (2010). Intersensory perception at birth: Newborns match nonhuman primate faces and voices. *Infancy, 15,* 46–60. http://dx.doi.org/10.1111/j.1532-7078.2009.00005.x

Lewkowicz, D. J., Minar, N. J., Tift, A. H., & Brandon, M. (2015). Perception of the multisensory coherence of fluent audiovisual speech in infancy: Its emergence and the role of experience. *Journal of Experimental Child Psychology, 130,* 147–162. http://dx.doi.org/10.1016/j.jecp.2014.10.006

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748. http://dx.doi.org/10.1038/264746a0

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition, 29,* 143–178. http://dx.doi.org/10.1016/0010-0277(88)90035-2

Nazzi, T., Jusczyk, P., & Johnson, E. K. (2000). Language discrimination by English-learning 5-month-olds. *Journal of Memory and Language, 43,* 1–19. http://dx.doi.org/10.1006/jmla.2000.2698

Otsuka, Y., & Yamaguchi, M. K. (2003). Infants' perception of illusory contours in static and moving figures. *Journal of Experimental Child Psychology, 86,* 244–251. http://dx.doi.org/10.1016/S0022-0965(03)00126-7

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonemic information in lips and voice. *Developmental Science, 6,* 191–196. http://dx.doi.org/10.1111/1467-7687.00271

Pons, F., Lewkowicz, D. J., Soto-Faraco, S., & Sebastián-Gallés, N. (2009). Narrowing of intersensory speech perception in infancy. *Proceedings of the National Academy of Sciences of the United States of America, 106,* 10598–10602. http://dx.doi.org/10.1073/pnas.0904134106

Querleu, D., Renard, X., Versyp, F., Paris-Delrue, L., & Crèpin, G. (1988). Fetal hearing. *European Journal of Obstetrics, Gynecology, and Reproductive Biology, 28,* 191–212. http://dx.doi.org/10.1016/0028-2243(88)90030-5

Reissland, N., & Francis, B. (2010). The quality of fetal arm movements as indicators of fetal stress. *Early Human Development, 86,* 813–816. http://dx.doi.org/10.1016/j.earlhumdev.2010.09.005

Roark, D. A., Barrett, S. E., Spence, M. J., Abdi, H., & O'Toole, A. J. (2003). Psychological and neural perspectives on the role of motion in face recognition. *Behavioral and Cognitive Neuroscience Reviews, 2,* 15–46. http://dx.doi.org/10.1177/1534582303002001002

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics, 59,* 347–357. http://dx.doi.org/10.3758/BF03211902

Smith, N. A., & Strader, H. L. (2014). Infant-directed visual prosody: Mothers' head movements and speech acoustics. *Interaction Studies, 15,* 38–54. http://dx.doi.org/10.1075/is.15.1.02smi

Soken, N. H., & Pick, A. D. (1992). Intermodal perception of happy and angry expressive behaviors by seven-month-old infants. *Child Development, 63,* 787–795. http://dx.doi.org/10.2307/1131233

Streri, A., & Gentaz, E. (2003). Cross-modal recognition of shape from hand to eyes in human newborns. *Somatosensory and Motor Research, 20,* 13–18.

Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition, 108,* 850–855. http://dx.doi.org/10.1016/j.cognition.2008.05.009

Vouloumanos, A., & Werker, J. F. (2007). Listening to language at birth: Evidence for a bias for speech in neonates. *Developmental Science, 10,* 159–164. http://dx.doi.org/10.1111/j.1467-7687.2007.00549.x

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., & Werker, J. F. (2007). Visual language discrimination in infancy. *Science, 316*(5828), 1159–1159.

Yeung, H. H., & Werker, J. F. (2013). Lip movements affect infants' audiovisual speech perception. *Psychological Science, 24,* 603–612. http://dx.doi.org/10.1177/0956797612458802

Received March 19, 2015
Revision received March 16, 2016
Accepted March 16, 2016 ■