Hearing Faces: How the Infant Brain Matches the Face It Sees with the Speech It Hears

Davina Bristow[1,2], Ghislaine Dehaene-Lambertz[1,3,4], Jeremie Mattout[1,4], Catherine Soares[1,4], Teodora Gliga[1,4], Sylvain Baillet[4,5], and Jean-François Mangin[4,6]

Abstract

■ Speech is not a purely auditory signal. From around 2 months of age, infants are able to correctly match the vowel they hear with the appropriate articulating face. However, there is no behavioral evidence of integrated audiovisual perception until 4 months of age, at the earliest, when an illusory percept can be created by the fusion of the auditory stimulus and of the facial cues (McGurk effect). To understand how infants initially match the articulatory movements they see with the sounds they hear, we recorded high-density ERPs in response to auditory vowels that followed a congruent or incongruent silently articulating face in 10-week-old infants. In a first experiment, we determined that auditory–visual integration occurs during the early stages of perception as in adults. The mismatch response was similar in timing and in topography whether the preceding vowels were presented visually or aurally. In the second experiment, we studied audiovisual integration in the linguistic (vowel perception) and nonlinguistic (gender perception) domain. We observed a mismatch response for both types of change at similar latencies. Their topographies were significantly different demonstrating that cross-modal integration of these features is computed in parallel by two different networks. Indeed, brain source modeling revealed that phoneme and gender computations were lateralized toward the left and toward the right hemisphere, respectively, suggesting that each hemisphere possesses an early processing bias. We also observed repetition suppression in temporal regions and repetition enhancement in frontal regions. These results underscore how complex and structured is the human cortical organization which sustains communication from the first weeks of life on. ■

INTRODUCTION

Phonemes are the elementary building blocks of speech, the combination of which allows the rich verbal exchanges between members of the human species. In adults, they are processed by a specialized and highly efficient left-hemispheric network, which subserves the singular properties of phoneme perception, such as categorical perception, normalization across nonpertinent acoustical variations, and so forth. There is a long-lasting debate about whether a similar network that might favor language acquisition is present in infants. Striking similarities are observed between infant and adult performances. Like adults, infants perceive phonemes categorically from birth (Eimas, Siqueland, Jusczyk, & Vigorito, 1971) and are able to normalize speech input (Kuhl, 1983). ERP studies have revealed that these capacities are based on a fast computation of a phonetic representation. Indeed, in infants, the response to a change of syllable occurring after several repetitions of the same syllable (mismatch response, MMR) demonstrates two of the main properties used to define phonetic representations in adults. First, the MMR is not affected by irrelevant variation in speaker (Dehaene-Lambertz & Pena, 2001). Second, it is categorical, being dependent on the linguistic value of the acoustical change. The MMR is larger after a change of syllable that crosses a phonetic boundary than after an acoustically similar change that occurs within the phonetic category (Dehaene-Lambertz & Baillet, 1998). Dipole modeling of the sources of this MMR points toward a temporal origin in infants (Dehaene-Lambertz & Baillet, 1998; Dehaene-Lambertz & Dehaene, 1994) as in adults (Näätänen et al., 1997).

To increase our understanding of speech perception early in life, we can turn to another property of phoneme perception observed in adults—the rapid and automatic integration of facial cues. Seeing the articulatory movements of the speaker improves the detection (Grant & Seitz, 2000) and understanding of speech sounds (Erber, 1975), especially in noisy situations. Moreover, facial information can modify the perception of speech sounds giving rise to a novel percept, a

[1]INSERM, Neurospin, Gif/Yvette, France, [2]University College London, UK, [3]AP-HP, Le Kremlin-Bicêtre, France, [4]IFR49, Neurospin, Gif/Yvette, France, [5]CNRS, LENA, Paris, France, [6]CEA, UNAF, Neurospin, Gif/Yvette, France

phenomenon known as the McGurk effect (McGurk & MacDonald, 1976). For example, an auditory presented syllable /ba/ is irrepressibly perceived as /da/ when associated with a face articulating /ga/. This phenomenon occurs even when subjects know what the stimuli actually are, demonstrating automatic cross-modal integration (Rosenblum & Saldana, 1996; Summerfield & McGrath, 1984). ERP studies have revealed that this audiovisual integration occurs during the early stages of perception. The mismatch negativity induced by the illusory phoneme change created by a McGurk effect is not delayed relative to a mismatch negativity recorded after a real auditory change (around 150–250 msec) and the topography of the MMR is similar in both cases (Saint-Amour, De Sanctis, Molholm, Ritter, & Foxe, 2007; Colin et al., 2002). Early cross-modal integration is confirmed by the difference recorded between ERPs to audiovisual speech and the sum of the ERPs to unimodal auditory and visual stimuli on the N1/P2 auditory complex (van Wassenhove, Grant, & Poeppel, 2005). This integration is realized within the superior temporal region (Reale et al., 2007; Calvert, Campbell, & Brammer, 2000), predominantly along the STS, along which fMRI has revealed intermixed patches of voxels responding to unimodal auditory, visual stimuli and to audiovisual stimuli (Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004). Other regions, such as the Broca's area and the precentral cortex, may also contribute depending on the task demand, for instance, when visual and auditory information is conflicting or when syllables are presented in a noisy environment (Ojanen et al., 2005; Callan et al., 2003). This reveals a complex and dynamic interplay between the different regions involved in speech perception but also production (Nishitani & Hari, 2002).

Infants' first encounter with speech during the last trimester of fetal life is purely auditory; however, after birth, speech rapidly becomes a multimodal stimulus as visual and motor capacities develop. By around 2 months, infants become able to match silently articulating faces with the appropriate auditory vowel (Patterson & Werker, 2003; Kuhl & Meltzoff, 1982). By 3 to 5 months of age, they produce vocalizations approaching the target vowel in response to an audiovisual model (Kuhl & Meltzoff, 1982). Around 5 months of age, seen and heard speech can be integrated inducing adult-like McGurk percepts (Burnham & Dodd, 2004; Rosenblum, Schmuckler, & Johnson, 1997). However, another study at 4 months found an inconsistent pattern of integration. Both the sex of the infant and the positioning of the illusory percept (as habituation or test stimuli) affected the outcome (Desjardins & Werker, 2004). The authors therefore suggested that experience may be necessary to consolidate the integration of audiovisual speech and that cross-modal fusion may not be as robust or as consistent in infants as it is in adults.

Thus, prior to 4 months of age, at the earliest, there is no evidence of integrated audiovisual perception. The earlier ability to match auditory and visual speech information (Patterson & Werker, 2003; Kuhl & Meltzoff, 1982) does not necessarily indicate that infants are already integrating auditory and visual speech information into a single percept rather than merely being sensitive to the association between two frequently co-occurring stimuli. Studies in animals indicate that multisensory integration is rudimentary during early postnatal life and develops only gradually thereafter (Wallace, Carriere, Perrault, Vaughan, & Stein, 2006). Furthermore, because of the heterogeneous maturation of the different brain regions and the connections between them (Dubois et al., 2008; Yakovlev & Lecours, 1967), the complex adult linguistic network may initially be limited. Potential cross-modal areas (i.e., in the STS) may not have received sufficient input or be sufficiently mature to take advantage of their cross-modal properties, or the adults' network may be scattered, requiring efficient connections to unify and synchronize its subcomponents. However, it is possible that these connections are efficient from very early on, favoring a rapid coupling between seen, heard, and produced speech in infants.

To study the neural bases underlying infants' early capacities to match auditory and visual speech cues, we used ERPs in a mismatch paradigm, in which a change (deviant stimulus) is introduced after several repetitions of a standard stimulus. In previous auditory experiments, it has been shown that the MMRs to speech stimuli possess phonetic properties (i.e., normalization and categorical perception). Because this component can be modulated by visual cues in adults (i.e., McGurk illusion; Colin et al., 2002), we hypothesized that the ERP in response to an auditory vowel might be affected by the prior presentation of visual cues. Therefore, in the first experiment, we compared the auditory evoked response to a change of vowel when the preceding vowels were presented either aurally (unimodal auditory context) or visually (cross-modal context). The latency of the MMR in both contexts would indicate when cross-modal integration occurs relative to unimodal perception. If the phonetic representation is cross-modal as in adults, the MMRs would be comparable in both contexts, whereas if infants are only associating visual and auditory stimuli, the MMR to deviant vowels would be delayed in the cross-modal context and its topography would be different from that of the unimodal auditory MMR. In the second experiment, we replicated the cross-modal MMR observed in Experiment 1 and contrasted two types of information conveyed by the face and voice: ''who'' (is speaking) and ''what'' (is said), by allowing the gender of the speaker, as well as the vowel spoken, to vary between the visual context and the auditory test. Both attributes are important for our social species, yet behavioral experiments suggest that audiovisual integration of linguistic information occurs much earlier than the integration of other cues, such as the gender of the face or the voice. This observation has been interpreted as suggesting that

specific learning mechanisms are involved for linguistic stimuli (Patterson & Werker, 2002). Furthermore, we took advantage of our high-density recordings to compute the sources of the MMRs and clarify the difference in processing of auditory and visual linguistic and nonlinguistic cues in infants.

EXPERIMENT 1

The goals of the first experiment were to examine at which stage cross-modal integration occurs relative to unimodal perception and to determine whether the phonetic MMR reflects a purely auditory computation or whether repetition and change could be computed across-modalities. Context information on the vowel category was either given through the visual modality [i.e., cross-modal trials: a face silently articulating a vowel (/a/ or /i/) twice in succession was presented before the auditory vowel (/a/ or /i/)], or through the auditory modality [i.e., unimodal trials: the auditory test stimulus (/a/ or /i/) was preceded by two different auditory exemplars of /a/ or /i/, in association with a face, whose mouth was hidden]. We compared infants' ERPs in response to the auditory test stimulus when a change occurred between the test vowel and the previous stimuli, or not, in both contexts (Figure 1). This paradigm has several advantages. First, we were able to study the response to the exact same stimuli (auditory vowels) while targeting a unimodal or a cross-modal representation depending on the context in which they were presented. Second, we presented only natural stimuli in order to target regular speech processing and avoid surprise effects induced by incongruent auditory–visual pairing. Finally, the comparison of the latency of the MMR in unimodal and cross-modal contexts was not biased by any visual cues, the face being replaced by a bull's eye during the presentation of the test vowel.

Methods

Subjects

Twenty-one full-term infants (12 boys and 9 girls) were tested between 9 and 12 weeks after birth (mean age = 10.2 weeks, *SD* = 0.7 weeks). Fifteen additional infants were tested but rejected for fussiness, excessive movement, or bad recording. The study was approved by the regional ethical committee for biomedical research, and parents gave their written informed consent.

Visual Stimuli

Two male and two female actors were filmed articulating /a/ and /i/ against a white background. Four frames were

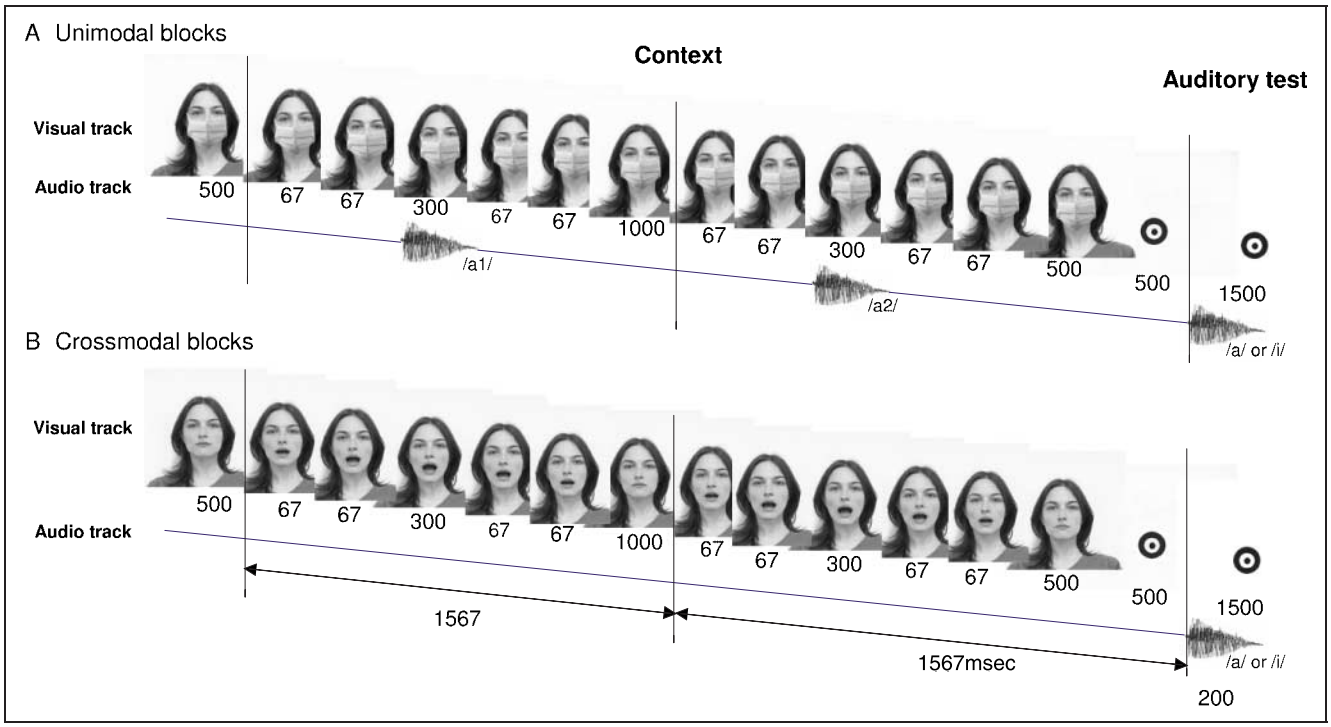

Figure 1. Trial structure: Each trial consisted of the visual presentation of faces (context) followed by the presentation of an auditory vowel (/a/ or /i/) (test) while the face was replaced by a bull's eye. In unimodal blocks (A), the face's mouth was masked and two different exemplars of the same auditory vowel were successively presented (SOA = 1567 msec) before the test vowel. In cross-modal blocks (B), the silent succession of frames depicted two articulatory movements (a shape or i shape). Eight different faces (4 male and 4 female) and four vowel exemplars (2 male and 2 female) were used for the context. The auditory test vowel /a/ or /i/ was either congruent or incongruent with the context.

extracted from each clip: (i) mouth closed, (ii) beginning of movement, (iii) mouth semi-extended, (iv) mouth fully extended. These images were used to recreate natural articulatory movements, which were used as the context stimulus in the cross-modal trials. A still image of each actor was also extracted to be used in unimodal trials. The mouth region was hidden with a surgical mask.

Auditory Stimuli

As context stimuli in unimodal auditory trials, four different /a/ and /i/, lasting 190 to 200 msec and produced by two males and two females, were used. To keep some acoustical variations, no effort was made to match these vowels. Two other vowels /a/ and /i/ produced by a male and a female speaker were kept as test vowels and were matched for duration (190 msec) and subjective intensity.

Procedure

With the EEG net placed on their head, infants were seated on their parent's lap facing a projector screen and two loudspeakers hidden behind the screen on each side. The screen was located 80 ± 10 cm away, spanning a visual angle of 25° × 25°.

Each trial had an AAX structure, with two repetitions of a context stimulus followed by an auditory vowel (test stimulus) congruent or incongruent with the previous context. Two types of trials, unimodal and cross-modal trials, were presented, which depended on the modality of the context stimuli. The test phase was similar across unimodal and cross-modal trials (Figure 1).

In cross-modal trials, the context (visual /a/ and visual /i/) consisted of the silent succession of visual frames during 2636 msec depicting two articulatory movements. The extrema of the articulatory movements for /i/ and /a/ were presented for 300 msec with two intermediate frames of 67 msec preceding and following this to create a natural articulatory movement of the mouth lasting 568 msec. The articulation was preceded and followed by 500 msec of the mouth being fully closed. In unimodal trials (auditory /a/ and auditory /i/), a still face was presented for the same duration, with the mouth hidden by a surgical mask to avoid a lack of mouth movement conflicting with the sounds heard. Two different auditory exemplars (from different voices but within the same gender) of the same vowel were presented at 634 msec and 2201 msec after the onset of the trial (corresponding to the onset of the full-mouth slide in cross-modal trials).

For both types of trials, the face was then replaced by a brightly colored ''bull's eye'' in the same location as the mouth and nose in the visual stimuli, and an auditory vowel (the test stimulus) was presented 500 msec after the appearance of the bulls eye. This vowel was either congruent or incongruent with the context stimuli. The gender of the test vowel was always kept congruent with the gender of the context stimuli within the same trial. The ''bull's eye'' remained for 1500 msec and was then replaced by a new face randomly chosen among the four possibilities for the next trial. When the infant looked away from the screen, the experiment was stopped and the infant's gaze was attracted back to the screen before the experiment resumed. If it was not possible, the experiment was terminated. If the infant looked away midway through a trial, the experiment recommenced at the start of the interrupted trial.

The context (visual /a/, visual /i/, auditory /a/ or auditory /i/) was kept constant within blocks of 16 trials, and then was changed for one of the other possibilities, with /a/ and /i/ alternating as the context vowel. Within each block, faces were randomly selected for each trial among the four possible actors. The order of blocks was counterbalanced across subjects and eight blocks were presented, giving a total of 128 trials. Each infant was thus exposed to each condition. Each test vowel (male and female, /a/ and /i/) was presented equally in each condition (congruent and incongruent), in each context (uni and cross-modal), and in each subject.

ERP Recordings

Scalp voltages were recorded from a Geodesic sensor net (EGI, 129 channels) referenced to the vertex. They were amplified, digitized at 250 Hz, and filtered between 0.5 and 20 Hz. The EEG was segmented into epochs starting 500 msec before and ending 1500 msec after each auditory test stimulus. Channels contaminated by eye or motion artifacts were automatically rejected and trials with more than 50% bad channels were excluded. Our choice was to present short blocks of 16 trials to ensure that each infant was exposed to each condition, with each vowel appearing in standard and test position while keeping the experiment short enough for this age. However, we observed that the response to the test auditory vowels during the first trials in a cross-modal block following a unimodal block was affected by the numerous repetition of the standard auditory vowel in the previous block with a progressive decay (see Winkler, Cowan, Csépe, Czigler, & Näätänen, 1996 for an example in adults of how the MMN is also affected by the change of the standard sound). The reverse effect (i.e., interference of the standard visual vowel of a cross-modal block on the first test vowels of the following unimodal block) was also present but weaker. To avoid this type of carried-over effect, we excluded the first eight trials following a change of type of blocks. Therefore, after artifact rejection, only 34 trials were retained, on average, per infant (8.9, 8.2, 9.2, and 9.8 trials per infant for cross-modal congruent vowel, cross-modal incongruent vowel, unimodal congruent vowel, and unimodal incongruent vowel conditions, respectively[1]). The artifact-free trials were averaged for each infant and for

each condition. Averages were then baseline-corrected (−200 to 0 msec) and an average reference transformation was applied to obtain reference-independent potentials. The data were then down-sampled to 65 electrodes in order to compare Experiments 1 and 2, for which it was not possible to use the 129-channels net.

ERP Analyses

As the same test vowels were used across blocks in congruent and incongruent conditions, any significant difference between the waveforms indicates that, in incongruent trials, infants have detected a change between the auditory test vowel and the preceding context stimuli. Classically, in adults, the introduction of a new sound after a series of repeated sounds induces an early MMR characterized by an inversion of polarity between a negativity over right frontal region and a positivity over left occipital–temporal regions (Näätänen, 1990). In infants, a similar response is observed but the polarity is usually, but not always, positive over the frontal lobe and negative over the posterior regions. Its latency is also delayed relative to adults, and occurs between 100 and 400 msec (see Dehaene-Lambertz & Gliga, 2004 for a review and a discussion about auditory MMR in infants). We inspected the time course of two-dimensional reconstructions of the difference between all congruent and incongruent trials across both contexts (main effect of congruency). A dipolar topography corresponding to an infant's MMR was present from 152 to 500 msec. We report here only analyses performed on the onset of this response which corresponds to the ascending slope of the first auditory peak (152–300 msec) because we were interested in early differences between unimodal and cross-modal contexts. We selected clusters of electrodes at the positive and negative maxima of the dipole configuration (Figure 2). These clusters, which were consistent with the topography of the MMR both in adults (Näätänen, 1990) and in infants (Dehaene-Lambertz & Dehaene, 1994), comprised 12 electrodes over right frontal areas (encompassing C4, Fz, F4, and F8) and 10 electrodes over left posterior regions (comprising O1, P3, T5, and the mastoid).

ANOVAs were performed on the voltage averaged across electrodes in these two clusters over the selected time window (152–316 msec), with electrodes (frontal and posterior), congruency (congruent and incongruent), and context (unimodal and cross-modal) as within-subjects factors. Because of the selection of the electrodes at the dipoles maxima, a main effect of electrodes is not informative, thus only interactions between electrodes and the other factors are examined.

To investigate whether a difference between two voltage topographies is related to a change in source configuration or in source strength, McCarthy and Wood (1985) suggest performing analyses on normalized data. The MMR in each subject was thus scaled by the vector length defined as the square root of the sum of squared voltages over all electrode locations before being entered in an ANOVA with context (unimodal vs. cross-modal) as within-subject variable.

Finally, we examined the latencies of the peak of the MMR. The peak latency was determined as being the first time point showing the maximum voltage difference between the positive and negative poles of the studied MMR.

Results

In response to the auditory test vowels, we recorded a classical infant auditory evoked response potential (Dehaene-Lambertz & Dehaene, 1994) with two peaks, at 298 msec (SE = 8 msec) and at 440 msec (SE = 17 msec) after stimulus onset. The auditory ERP was weaker when the test vowel was congruent with the preceding context stimulus in both unimodal and cross-modal contexts, giving rise to a long-lasting (152–500 msec) difference (MMR) above the frontal and posterior electrodes (Figure 2).

Because we were interested by the onset of the MMR in both contexts, we report only analyses of the first peak of the auditory response (152–316 msec), but results were similar if the entire mismatch period (152 to 500 msec) was examined. The MMR was significant during this early time window [Electrodes × Congruency: $F(1, 20) = 11.45$, $p = .003$] and did not interact with the context [Electrodes × Congruency: × Context: $F(1, 20) < 1$]. Post hoc analyses restricted to each context confirmed that an MMR was present in the unimodal [Electrodes × Congruency: $F(1, 20) = 5.00$, $p = .037$] and cross-modal context [Electrodes × Congruency: $F(1, 20) = 6.67$, $p = .018$]. At each electrode site, there was a mismatch effect [frontal site: $F(1, 20) = 11.17$, $p = .003$; posterior site: $F(1, 20) = 6.67$, $p = .018$] with no significant interaction between congruency and context [frontal site: $F(1, 20) = 1.31$, $p = .266$; posterior site: $F(1, 20) < 1$].

The response evoked by incongruent trials in cross-modal context began no later than the response evoked by incongruent trials in unimodal context. The peak latency of the MMR was similar in the unimodal (293 msec, SE = 32 msec) and cross-modal conditions [314 msec, SE = 27 msec, $F(1, 20) < 1$; Figure 2C]. In both conditions, its topography, a right frontal positivity synchronous with a left temporo-occipital negativity, corresponded to the classical auditory MMR previously recorded at this age (Dehaene-Lambertz & Dehaene, 1994). Its topography was slightly rotated clockwise in the unimodal context relative to the cross-modal context (Figure 2). The difference between the two MMR isolated a cluster of 10 frontal electrodes comprising Fz and F3. However, analysis of the normalized MMRs (see Methods) averaged over this cluster and over the studied time window (152–316 msec) revealed only a marginal effect of context [$F(1, 20) = 4.04$, $p = .058$], suggesting that the source configuration might

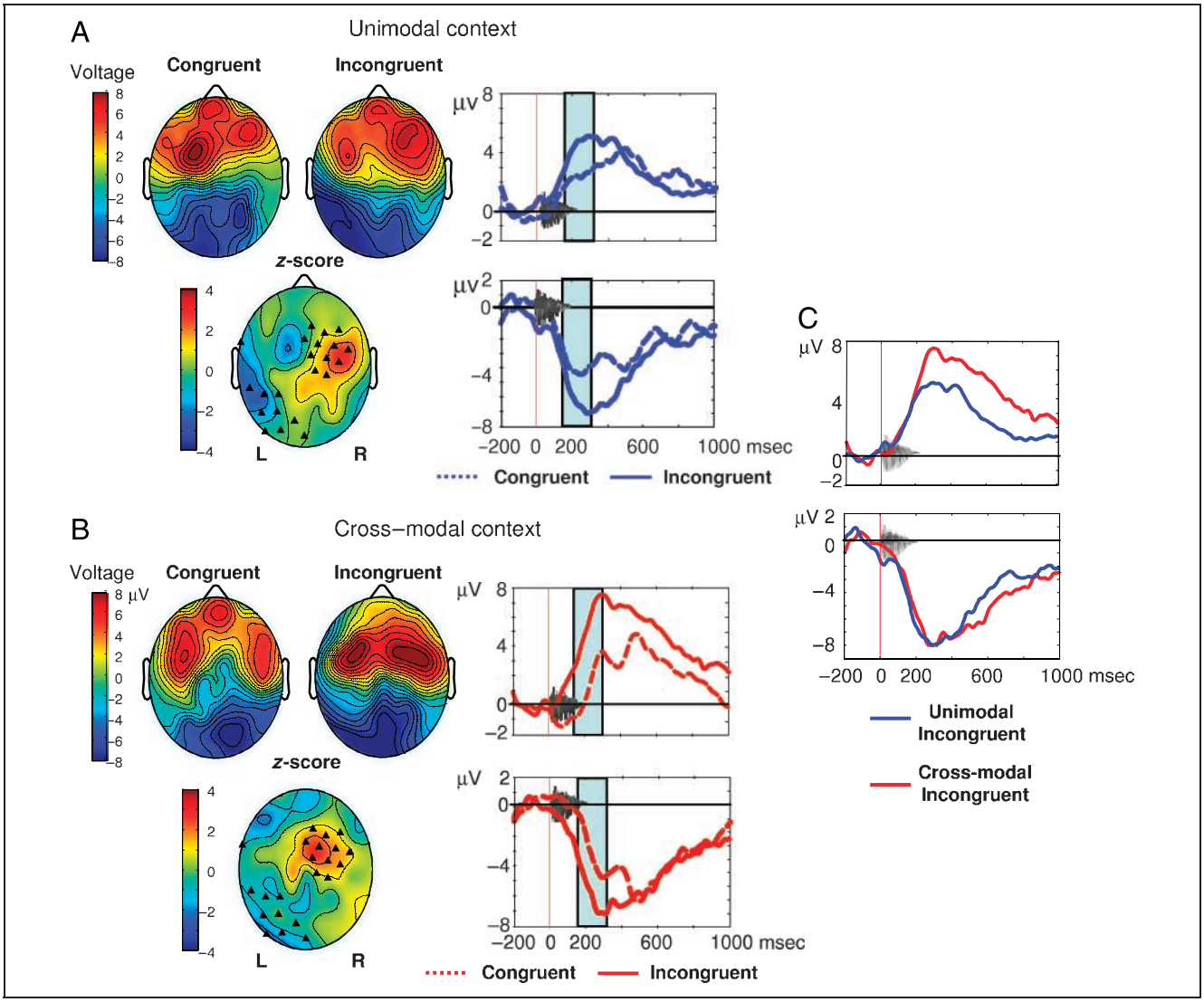

Figure 2. Mismatch responses in Group 1 in unimodal (A) and cross-modal blocks (B). In each panel, the upper row displays the topographies of the evoked potentials in response to the auditory test stimulus averaged in congruent and incongruent conditions from 152 to 316 msec poststimulus onset (this time window is presented as a rectangle on the waveforms on the right). The lower row displays a 2-D map of the *z*-score of the comparison between these conditions. On this map, triangles mark the location of the electrodes selected for statistical analysis at the positive (frontal site) and negative maxima (posterior site) of the dipole configuration of the mismatch response. The right column displays the time course of the voltage averaged over the same groups of electrodes in both contexts. In C, the waveforms of incongruent trials for both contexts are displayed on the same graphs to illustrate the absence of delay of the mismatch response in cross-modal context relative to unimodal context.

be slightly different for each context. However, because of the small number of trials contributing to the averages in each infant, the topographies may be noisy, making it difficult to draw any conclusions on the basis of the weak difference seen here.

To summarize, an MMR was observed, similar in timing and topography, in both cross-modal and unimodal contexts.

Discussion

The repeated prior presentation of visually articulated vowels habituated the infants' evoked response to auditory vowels in a phoneme-specific manner. We hypothesized that if infants have a neural network encoding a cross-modal or amodal representation of phonemes, the response of such a network to auditory phonemes would be affected by the prior presentation of that same phoneme visually. The significant difference between congruent and incongruent trials thus demonstrates that 9- to 12-week-old infants have a cross-modal neural representation of phonemes that responds to specific phonetic information irrespective of the modality it is perceived in. Furthermore, the MMR was similar in timing and with a close topography, in both cross-modal and unimodal contexts, suggesting

that auditory–visual integration occurs early on in speech processing and not at the level of semantic associations.

Our results thus suggest close similarities between infant and adult representations of speech. However, Jaaskelainen et al. (2004), using a similar paradigm to ours, observed different responses in adults. They presented adults with either pairs of auditory vowels or pairs of a visual (i.e., a movie clip of a person articulating the vowel) and an auditory vowel, or single auditory vowels. The N1 response to the second member of the pair was suppressed when it followed either another auditory vowel or its visual equivalent, although the suppression effect was weaker in the visual–auditory pair. However, suppression was not significantly stronger when the paired vowels belonged to the same vowel category than when they differed. Thus, seeing articulatory movements habituated the response to the auditory vowel in adults, as observed here, but this suppression was not phoneme specific unlike what we observed in infants. The absence of a specific effect in Jaaskelainen et al.'s experiment contradicts other evidence of early and specific cross-modal integration in adults. For example, Saint-Amour et al. (2007) and Colin et al. (2002) observed an MMN when the syllable change was an illusory percept created by the visual cues with no real auditory change. Jaaskelainen et al. suggested that the relative proximity of the mouth movements of the studied vowels (/ae/ vs /ø/) could explain the absence of a phoneme-specific effect in their study, or it maybe that a single repetition of the vowel was not enough to cause habituation irrespective of modality as they did not observe phoneme-specific habituation with auditory pairs either. Note also that in this experiment as in ours, the context faces should be coded as visual speech in order to elicit an MMR. Because mouth movements do not always have a linguistic value, a phonetic representation of a moving face might not be automatic when it is presented alone. The participant needs to recognize them as communicating faces. This is a very different situation from experiments studying the McGurk effect, or experiments with congruent and incongruent pairing of visual and auditory stimuli, in which the simultaneous presentation of auditory and visual stimuli is automatically coded as a communication act. In infants, either because faces have generally been seen in the context of communication or because they are biased to spontaneously code a moving face as a communicating face, or because articulating faces elicit imitation movements (Kuhl & Meltzoff, 1982; Meltzoff & Borton, 1979), the visual stimulus is apparently interpreted within the phonetic domain, permitting it to habituate the following auditory vowel, as demonstrated by the observed MMR. To confirm that infants are, indeed, able to compute phonetic cross-modal representations, we performed a second experiment using only cross-modal trials.

EXPERIMENT 2

The low number of trials retained in Experiment 1 made replication of our findings in a second group of infants desirable. Our second goal was to compare audiovisual integration of linguistic and nonlinguistic cues conveyed by the face and the voice (i.e., the gender of the speaker and the vowel produced). On one hand, infants should ignore individual variations to acquire language, but on the other hand, they should be attentive to individual differences in order to recognize a particular person. How are these antagonist requirements resolved by the infant's brain and how are cues from the voice and face integrated to fulfill these commitments? Behavioral studies suggest that audiovisual integration of gender occurs later than audiovisual integration of linguistic information and is difficult to observe before 8 months of age (Patterson & Werker, 2002; Walker-Andrews, Bahrick, Raglioni, & Diaz, 1991). This has been interpreted as suggesting that specific learning mechanisms are involved for linguistic stimuli (Patterson & Werker, 2002). For example, a motor representation can be used as a common frame to code visual and auditory representations of speech as postulated by the motor theory of speech (Liberman, 1996). According to this hypothesis, the imitation capacities displayed from the first days of life (visual–motor matching) and the auditory feedback provided by their own articulatory movements (motor–auditory matching) provide important information that enables the stabilization of a common motor representation of phonemes as a necessary step toward achieving visual–auditory vowel matching (Skipper, van Wassenhove, Nusbaum, & Small, 2007). Gender cross-modal representations cannot benefit from a similar mechanism and can only rely on associations between some voice features (e.g., low vs. high pitch) and face characteristics (e.g., thin vs. square face). Thus, if gender cross-modal representations are present, it suggests that there is no need to postulate a special mechanism for cross-modal speech integration.

Methods

Subjects

Sixteen full-term infants (4 boys and 12 girls) were tested between 9 and 12 weeks after birth (mean age = 10.5 weeks, *SD* = 0.92 weeks). Ten additional infants were rejected for fussiness, excessive movement, or bad electrode recording.

Procedure

The same stimuli and procedure as in Experiment 1 were used but only cross-modal trials were presented (Figure 3). The face gender (four different faces were used for each gender) and the articulatory movement

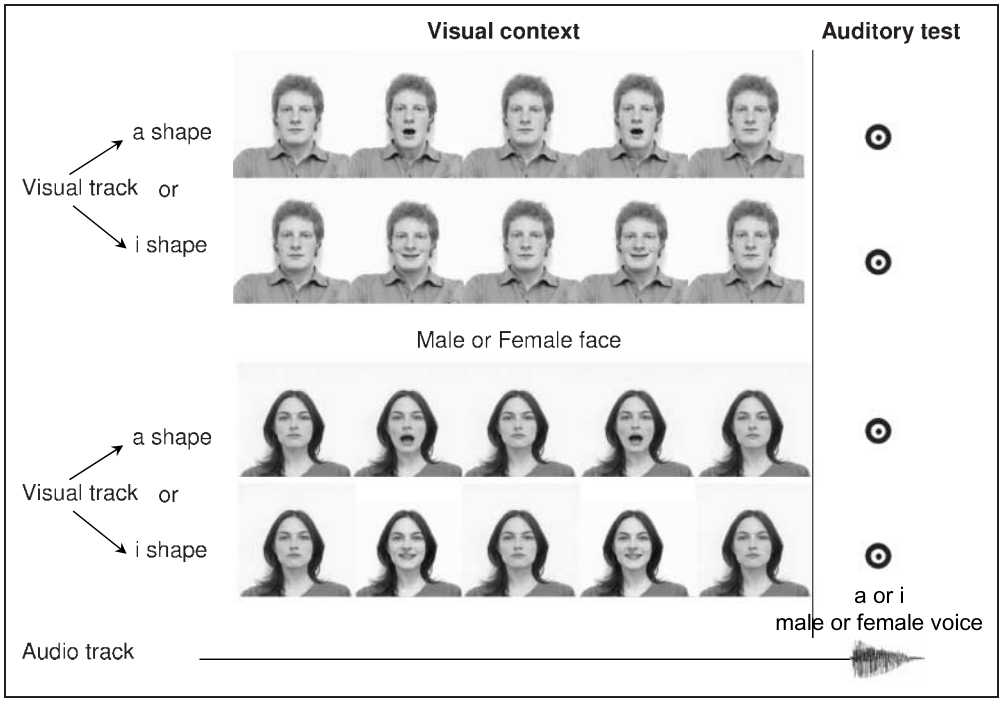

Figure 3. Trial structure in Experiment 2. Only cross-modal trials are presented in this experiment. The test vowel can be congruent or not with the visual context along the vowel or the gender dimension.

were kept constant within blocks of 32 trials. The order of blocks was counterbalanced across subjects and four blocks were used (128 trials). The test auditory vowel was congruent or incongruent with the visual stimuli along two dimensions, vowel and gender, creating four types of trials: (1) vowel congruent, gender congruent; (2) vowel incongruent, gender congruent; (3) vowel congruent, gender incongruent; and (4) vowel incongruent, gender incongruent. Each test vowel (male and female, /a/ and /i/) was presented equally in each condition and in each subject.

ERP Recordings

ERP recordings and data preprocessing were similar to Experiment 1, except that the net comprised 65 channels. After artifact rejection, on average, 104 trials were retained per infant (i.e., 26.8, 25.9, 26.1, and 25.7 for the vowel-congruent/gender-congruent, vowel-congruent/gender-incongruent, vowel-incongruent/gender-congruent, and vowel-incongruent/gender-incongruent conditions, respectively). The individual averages were averaged together to obtain four conditions (vowel congruent and vowel incongruent; gender congruent and gender incongruent) comprising around 50 trials in each subject and each condition.

ERP Analyses

The same analyses as in Experiment 1 were conducted. First, an ANOVA was computed on the voltage averaged across the same time window (152–316 msec) and the same clusters of electrodes as in Experiment 1, with electrodes (frontal and posterior), congruency (congruent and incongruent), and type of change (vowel and gender) as within-subjects factors.

We also tested the replication of the vowel cross-modal MMR with an ANOVA between the two groups of infants on the normalized data. The variance was significantly higher in Experiment 1 than in Experiment 2, violating the assumption of homogeneity required by ANOVAs (Var1/Var2 = 16). However, data normalization corrected this problem allowing the comparison [Var1/Var2 = 1.78 for a critical value of $F_{\text{Max}}(2, 20) = 2.46$ for $\alpha = .05$].

The inspection of the 2-D maps of the difference between gender-congruent and -incongruent trials revealed an MMR at the same latency as the vowel MMR but with a different topography. The topography was quite complex with two clusters of opposite polarities developing above the right anterior region, rapidly followed by a negative and a positive response over the right and left parieto-temporal channels. We thus selected clusters of electrodes at these four poles and conducted a second ANOVA with the same factors and on the same time window as above. The four selected sites consisted of three electrodes over the right anterior frontal region (F8 and under); three over the right anterior temporal region (anterior to T4); six over the right posterior region (comprising P4, T6, and mastoid); and six over the left posterior region (P3, T5, and under; see Figure 4).

Next, vowel and gender MMRs were normalized as explained above to further examine their topographical differences. Two ANOVAs, with type of change (vowel vs.

gender) and electrodes as within-subject variables, were computed. The electrodes corresponded to the clusters defined above and comprised two sites for the vowel MMR and four sites for the gender MMR.

Finally, the latencies of the peak of the MMRs defined as in Experiment 1 were analyzed.

Source Modeling

The topographical differences between the vowel and the gender MMR needed to be further explored. We took advantage of our dense spatial recordings and of the anatomical images obtained from our previous magnetic resonance experiments with infants (Dehaene-Lambertz, Hertz-Pannier, et al., 2006; Dubois, Hertz-Pannier, Dehaene-Lambertz, Cointepas, & Le Bihan, 2006) to compute a detailed model of the infant head and cortical folds to reconstruct a plausible distribution of the cortical origins of the surface voltage. Cortical current density mapping was obtained using a realistically distributed model consisting of 10,000 elementary current dipoles that sampled the cortical surface obtained from fine grained structural magnetic resonance imaging of a normal 3-month-old infant. Automatic brain surface extraction from infant MR image series is still an open challenge for the following reasons: (1) poor and heterogeneous myelinization of the brain at this age make automatic segmentation algorithms fail to accurately discriminate between white and gray matter and (2) the partial volume effect due to the relatively larger size of voxels with respect to the smaller size of the infant brain makes it difficult to track automatically the cortical gyrification patterns. Therefore, head and brain surfaces were semi-automatically extracted from a single infant MR image series using the BrainVisa software package (http://brainvisa.info/) to obtain a generic, although realistic, model of head tissues for EEG modeling. The locations and orientations of the elementary dipoles were constrained to the cortical mantle and the geometry of the EEG sensor net was warped to the head mesh. EEG forward modeling was computed using the overlapping-sphere analytical model with three shells (scalp, skull, and CSF). This technique has proven to reach comparable accuracy as numerical Boundary Element Model's, although with greater computational efficiency (Ermer, Mosher, Baillet, & Leah, 2001). All sphere and conductivity parameters were adjusted to the typical

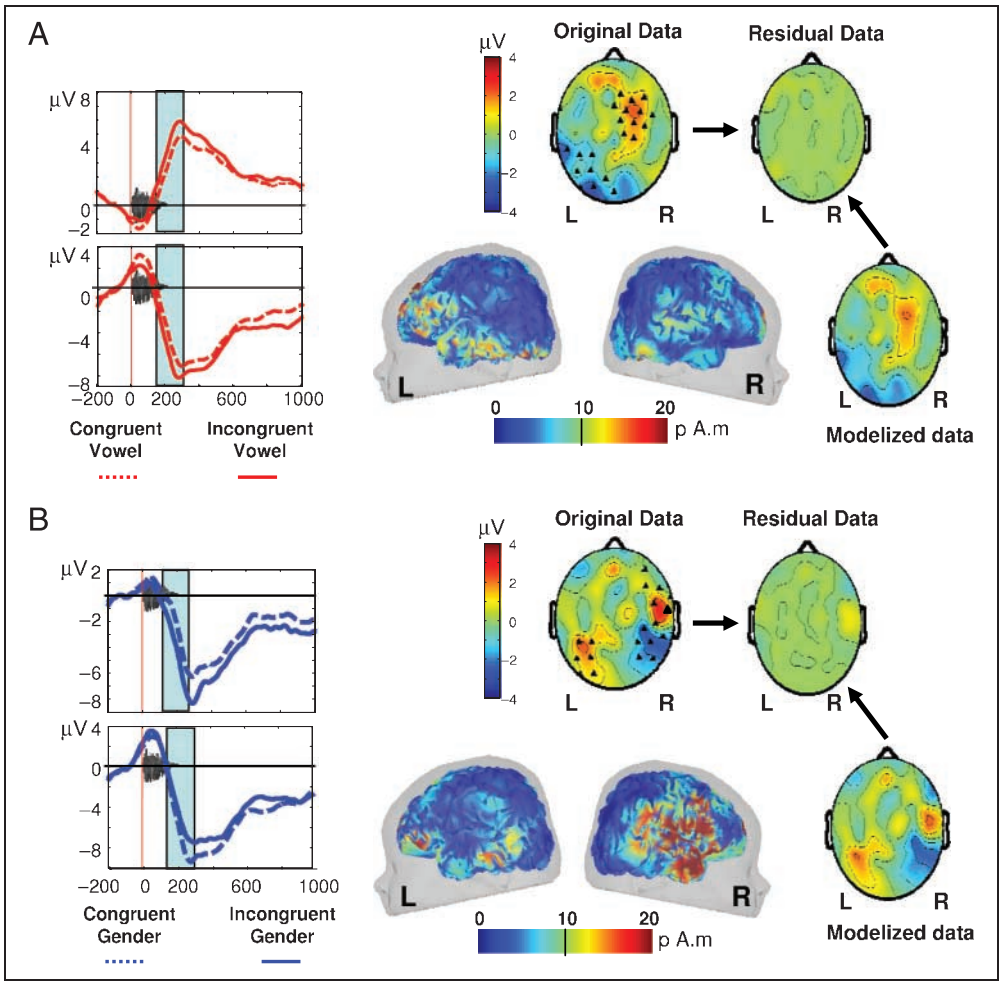

Figure 4. Vowel (A) and gender (B) mismatch responses in Experiment 2. The graphs on the left show the time course of the voltage averaged over groups of electrodes used in statistical analyses (frontal and temporal sites for vowel contrast and left posterior, right posterior sites for gender contrast). For each panel, 2-D maps of the difference between congruent and incongruent conditions, averaged across the 152–316 msec time window, are presented. Triangles represent the electrodes selected for statistical analysis. The sources of the vowel and gender mismatch responses averaged across the same time window, are presented on a 3-month-old 3-D brain in the lower part of each panel, with in the rightmost column, a 2-D map of their projection back on the surface of the head. The residual data map corresponds to the difference between the original data and the modelized data.

infants' head tissue properties which contain more water than the adult head tissues (Gibson, Bayford, & Holder, 2000). Cortical current maps were computed from the grand averages of the different conditions using a linear inverse estimator (weighted minimum-norm current estimate—WMNE). All these procedures were conducted with BrainStorm (http://neuroimage.usc.edu/brainstorm).

Results

ERP Analyses

The auditory evoked response potential developed with two peaks, at 300 msec (*SE* = 9 msec) and at 417 msec (*SE* = 17 msec) after stimulus onset.

Vowel change. We replicated the results observed in Experiment 1. The amplitude of the response was weaker for congruent vowel trials and an MMR was observed when the auditory test vowel was incongruent with the preceding silent articulations [Vowel Congruency × Electrodes: $F(1, 15) = 10.27$, $p = .006$; Figure 4]. There was no significant difference in the vowel cross-modal MMR between Experiments 1 and 2 [Experiment × Vowel Congruency × Electrodes on normalized data: $F(1, 35) < 1$]. By contrast, these two clusters of electrodes were not sensitive to gender information [Gender congruency × Electrodes: $F(1, 15) < 1$] and a significant three-way interaction between electrodes, congruency, and change type was observed [$F(1, 15) = 17.81$, $p < .001$]. The interaction Congruency × Change type was particularly significant over the posterior cluster [$F(1, 15) = 23.27$, $p < .001$].

Gender change. Whereas the effect of gender was not significant over the clusters of electrodes sensitive to a vowel change, a gender MMR was, nevertheless, present during the same time window but with a different topography (Figure 4). To evaluate whether this difference was significant, a similar ANOVA was conducted over the four sites displaying the response maxima. There was a significant Change type × Congruency × Electrode interaction [$F(3, 45) = 7.15$, $p = .0005$] due to an effect of congruency (i.e., a MMR) for gender [Congruency × Electrode: $F(3, 45) = 8.54$, $p = .0001$], but not for vowel [Congruency × Electrode: $F(3, 45) < 1$].

Topographical differences between vowel and gender MMR. To verify that the difference between these MMRs cannot be attributed to a difference of amplitude, but was related to a genuine difference of sources, data were normalized as explained above and entered in an ANOVA with electrodes (frontal and posterior) used in the vowel mismatch analysis, and type of change (vowel and gender) as within-subjects factors. This analysis confirmed the significant interaction between electrodes and change type [$F(1, 15) = 16.00$, $p = .001$]. When the electrodes were the four groups used to study the gender MMR, the interaction Electrodes × Change type was also significant [$F(3, 45) = 6.98$, $p < .001$].

Latencies analysis. Finally, we performed an analysis of the latencies of the peaks of the MMRs, which showed no difference in peak latencies [Vowel: mean = 287 msec, *SE* = 30 msec and Gender: mean = 286 msec, SE = 33 msec; $F(1, 15) < 1$].

To summarize, we recorded two significant MMRs for the vowel and the gender change. Their latency was similar but their topography was significantly different, suggesting that these features are computed in parallel by two different networks.

Brain Sources

Cortical sources modeling confirm that different brain regions were involved in both responses. At the maximum of the MMRs (286 msec postvowel onset), the sources were differently lateralized (Figure 4). For the vowel MMR, they were localized to the left inferior frontal cortex, the left anterior superior temporal region, and the left inferior temporal gyrus, whereas the source of the gender MMR was a larger more diffuse region along the right Sylvain fissure. The proposed model explained 85% and 82% of the surface voltage, respectively, for the vowel and the gender MMR during the 152–316 msec time window. As can be seen in Figure 4, the particular surface topography of both MMRs is correctly captured by the brain sources modeling.

Temporal Profile of the Brain Sources

By exploiting the high temporal resolution of EEG and the spatial information of source modeling, we were also able to examine the temporal profile of activity for congruent and incongruent conditions in regions of interest highlighted by studies of cross-modal integration in adults, around the Sylvian scissure (Jones & Callan, 2003; Nishitani & Hari, 2002; Calvert et al., 2000). In the left hemisphere, we observed three different response profiles (Figures 5 and 6). In the left superior temporal gyrus, activity was similar for vowel congruent and incongruent conditions. In the adjacent ventral regions (e.g., STS), activity decreased with cross-modal vowel repetition. Conversely, in the left inferior frontal cortex, cross-modal repetition enhanced activity relative to incongruent trials. Meanwhile, in the right hemisphere, a similar pattern was observed with a decrease of activity for congruent relative to incongruent gender in superior temporal regions and the reverse pattern in the frontal region (Figure 5).

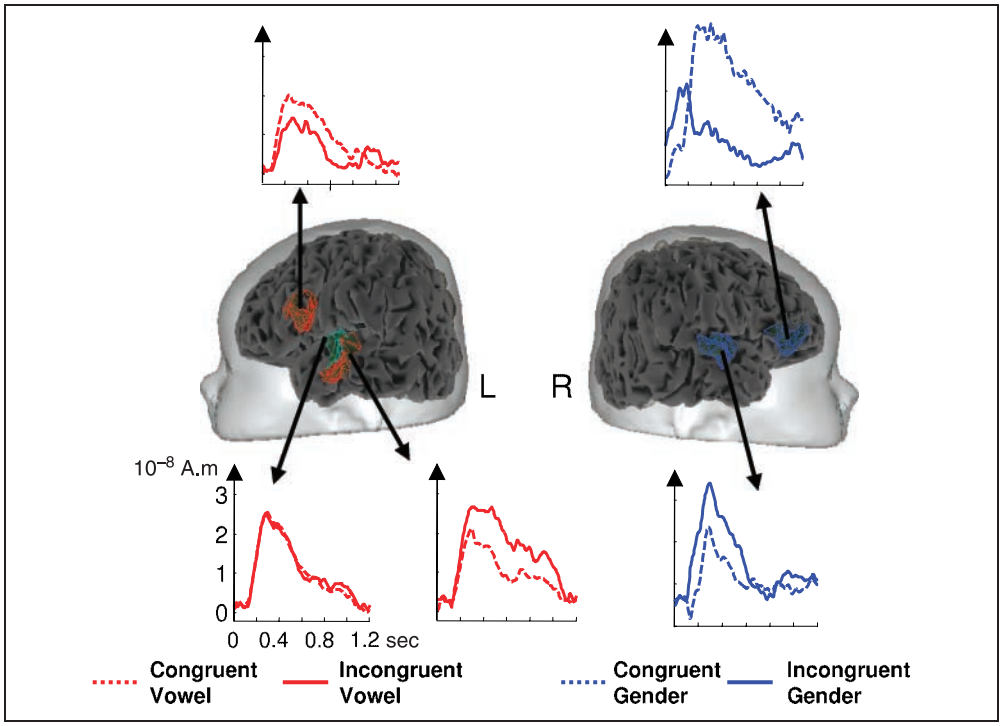

Figure 5. Cortical source reconstructions. Time courses of cortical sources from brain areas representative of the three types of observed activity are presented for congruent and incongruent conditions for vowel and gender contrasts. Each waveform represents the mean of activity of the region of interest color marked on the infant's 3-D brain. No difference is observed between vowel congruent and incongruent trials in the left superior temporal gyrus. In the left superior temporal sulcus and more ventral areas, activity is reduced in vowel congruent trials relative to incongruent trials, whereas the opposite pattern of activity is observed in left inferior frontal regions. Activity decreased for gender-congruent trials relative to incongruent trials in the right temporal cortex, but not in the frontal region.

Discussion

This second experiment confirmed the results obtained in the first. In two independent groups of infants, we observed an MMR when the auditory test vowel was incongruent with the previously seen mouth movements. This demonstrates early integration of visual cues into phonetic representations. The response to a change of

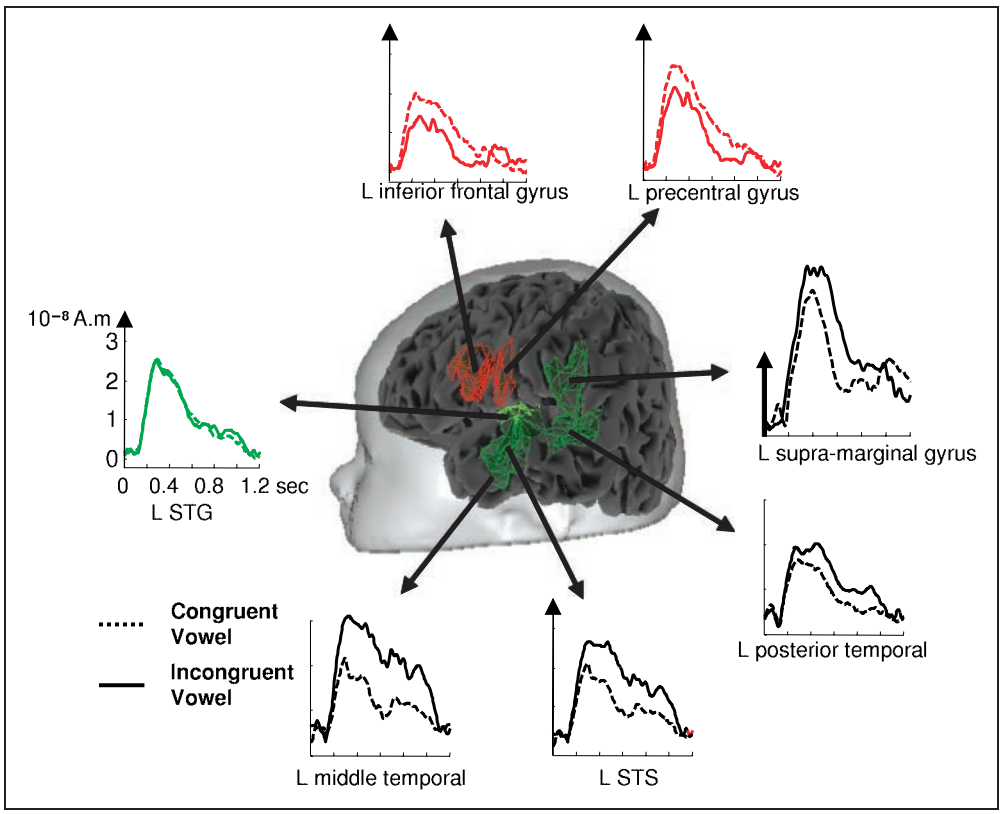

Figure 6. Cortical source reconstructions of vowel responses. Time courses of cortical sources from different brain areas that are involved in speech perception and cross-modal integration in adults (Nishitani & Hari, 2002). Each waveform represents the mean activity of the region of interest, color marked on the infant's 3-D brain (left hemisphere). Waveforms of regions with a similar response profile are in the same color: green = the response is similar for congruent and incongruent trials (superior temporal gyrus [STG]); black = activity is reduced for congruent trials relative to incongruent trials (middle temporal gyrus, superior temporal sulcus [STS], posterior temporal region, supramarginal gyrus); red = activity is increased for congruent trials relative to incongruent trials (inferior frontal gyrus, precentral gyrus).

phoneme presented similar functional properties to that described in adults (categorical perception, normalization and integration of facial cues), confirming a continuity between infant and adult speech computations (Dehaene-Lambertz & Gliga, 2004). Furthermore, we recorded two MMRs with two different topographies when either the gender of the speaker or the vowel spoken was incongruent with the previous visual information. The gender effect cannot be a spurious effect related to noise in the data because vowel and gender were orthogonal variables, (i.e., the same trials contributed to gender and vowel analyses), and thus, the vowel comparison would be similarly affected by residual noise in the data. Thus, contrary to what had been suggested by behavioral experiments, auditory–visual integration of face and voice is not initially limited to the linguistic domain and encompasses other categories, such as gender. In both cases, this integration is realized at an early processing stage, the MMR being as fast for cross-modal vowel and gender mismatches as for a unimodal auditory vowel mismatch (around 300 msec). This latency is compatible with auditory MMRs recorded during the first year of life (Winkler et al., 2003; Cheour et al., 1998).

ERP Results Conflict with Behavioral Results

Our finding of cross-modal integration of gender information was surprising as it appears to conflict with the results of behavioral studies. Although very early on infants are sensitive to individual facial cues (Blass & Camp, 2003) and can discriminate voices (Floccia, Nazzi, & Bertoncini, 2000), gender categorization is difficult to observe with behavioral measures before 3 to 4 months of age in faces (Quinn, Yahr, Kuhn, Slater, & Pascalils, 2002; Leinbach & Fagot, 1993), 5 to 6 months in voices (Miller, 1983), and 6 to 8 months for cross-modal transfer of gender between faces and voices (Patterson & Werker, 2002; Walker-Andrews et al., 1991). Nevertheless, some behavioral data suggest that infants may be sensitive to gender equivalences in face and voice earlier on. For example, a 4-month-old's ability to match vowel information in faces and voices is disrupted when the gender of the face and voice are put in full conflict with vowel matching (Patterson & Werker, 2002). Three-month-old infants habituated with a particular combination of a face and voice looked longer to a new pairing, especially if the gender of the voice was changed (Brookes et al., 2001). Because the gap between a female voice and a male voice is greater than between two voices of the same gender, it is possible that infants were surprised by the size of the change rather than by the change of gender category. However, Experiment 2 shows at least that infants are sensitive to specific association between facial features and voice quality, and thus, should possess the tools to construct gender classes.

ERPs show greater sensitivity relative to behavioral measures due to two factors. First, behavior is often a composite measure reflecting the combined effects of several processing stages. Meanwhile, ERPs may be able to track the response of different brain systems even when this does not lead to an overt behavioral response. Many examples of ERP–behavior dissociations exist in the adult literature. For instance, high-density ERPs demonstrate a series of processing stages of subliminal visual stimuli that subjects deny seeing (Sergent, Baillet, & Dehaene, 2005). In a training task where subjects learn to attend to subtle phonetic differences, ERP evidence for stimulus discrimination (MMN) may appear as much as 24 hours before a change in overt behavior occurs (Tremblay, Kraus, & McGee, 1998). In infants, conflicts between several cues often lead to surprising behavioral failures. For example, infants can no longer discriminate numerosity when it conflicts with other object features (Feigenson, 2005). Their performance can drop when categorizing objects when they have to process both the identity and the location of the object (Mareschal & Johnson, 2003). ERPs can be sensitive to these different levels of computations (Izard, Dehaene-Lambertz, & Dehaene, 2008). Second, behavioral results are based on a few measures obtained after several minutes of habituation to one type of stimulus, whereas infants were exposed here to more than 100 trials. Such fast recurrent presentation can be more engaging. It also provides more evidence about standard and deviant stimuli helping to calibrate a scale on which the two voices and eight faces used here could have been classified in two different clusters.

Two Different Networks for Gender and Vowel Representations

The differences in the topography of the vowel and gender MMRs confirm that there is no univocal MMR but that this component reflects the network targeted by the repeated then changed feature. As in adults (Giard et al., 1995), the same stimulus is thus coded in parallel along its different features by different networks. To further characterize the brain regions involved, we used source modeling. Localization of scalp evoked responses is an ill-posed inverse problem as the relationship between surface voltages and cortical activity is a one-to-many mapping. Nevertheless, methods for ERP source reconstruction have greatly improved in the recent years, mostly based on the ability to derive realistic head models from high-density MRI (Baillet, Mosher, & Leahy, 2001). These distributed approaches yield a unique and most probable solution in a Bayesian sense (Mattout, Phillips, Penny, Rugg, & Friston, 2006). The interest of source modeling relies in its unique way of combining spatial and temporal information and such models have proved efficient in uncovering brain mechanisms and their temporal sequence in adults (Sergent et al., 2005; Nishitani & Hari, 2002) as well as in infants (Izard et al., 2008; Richards, 2005; Dehaene-Lambertz & Dehaene, 1994). In

this work, we based our source estimate on a dedicated realistic head model for infants and on grand averages to obtain more stable data. Although the spatial accuracy of ERP is coarser than fMRI, these models reveal a distinct lateralization for gender and vowel representations. Current intensity was stronger over the left hemisphere for the vowel MMR and on the right for the gender MMR (Figure 4). Although vowel processing is classically considered to be less strongly lateralized than consonant processing, numerous experiments in adults have reported a predominantly left-hemisphere response to vowel stimuli (Shestakova et al., 2002; Tervaniemi et al., 2000; Näätänen et al., 1997). By contrast, voice characteristics are predominantly processed by the right hemisphere (von Kriegstein, Kleinschmidt, Sterzer, & Giraud, 2005; Belin, Fecteau, & Bedard, 2004) where there is direct functional connectivity between auditory voice processing regions and visual face processing regions (von Kriegstein et al., 2005). Furthermore, depending on whether adults' attention is directed toward the linguistic content or the voice characteristics of the same speech stimuli, the left–right asymmetry can be reversed (Belin et al., 2004; Zatorre, Evans, Meyer, & Gjedde, 1992).

What is remarkable here is that our subjects are only 10 weeks old. For the first time, the lateralization of different aspects of speech processing observed in adults is clearly demonstrated during the first weeks of life. The left lateralization proposed by the brain source modeling is compatible with previous results obtained in infants with ERPs using syllables (Dehaene-Lambertz & Dehaene, 1994) and fMRI and NIRS using sentences (Pena et al., 2003; Dehaene-Lambertz, Dehaene, & Hertz-Pannier, 2002), suggesting a left-hemisphere bias toward processing speech stimuli from the first weeks of life. However, this left-hemisphere advantage was not specific to speech stimuli (Dehaene-Lambertz et al., 2002; Dehaene-Lambertz, 2000) and could have been related to structural asymmetries (e.g., left larger planum temporale, right shorter and steeper Sylvian scissure). Here, the fact that two different features of the same stimulus are channeled in parallel to a different hemisphere demonstrates a genuine functional difference between hemispheres beyond structural differences.

The left lateralization of linguistic processing has been attributed to better processing of fast-temporal transitions by this hemisphere (Zatorre & Belin, 2001; Schwartz & Tallal, 1980), whereas the right hemisphere would be better at processing spectral content. However, the left-hand lateralization of phonetic processing in adults goes beyond acoustical characteristics (Dehaene-Lambertz et al., 2005; Jacquemot, Pallier, Le Bihan, Dehaene, & Dupoux, 2003), suggesting that constraints other than low-level perceptual processes are driving lateralization in adults. Both vowel identification and voice categorization rely largely on the analysis of spectral content. The different lateralization of the sources in infants leads us to question whether the functional lateralization observed in infants goes beyond the physical features of the stimuli, as it does in adults, and is driven by the brain general organization itself. In particular, interactions with other functional networks that are themselves biased toward a particular hemisphere might channel processing toward that same hemisphere. For example, a right-hemisphere advantage has been described for face processing at this age (Tzourio et al., 1992). The involvement of emotional and face processing in voice identification might thus be one factor favoring the involvement of the right hemisphere in the processing of gender categorization.

Repetition Suppression in Temporal Regions and Repetition Enhancement in Frontal Regions

By exploiting the high temporal resolution of EEG and the spatial information of source modeling, we were also able to examine the temporal profile of activity for congruent and incongruent conditions in regions of interest highlighted by studies of cross-modal integration in adults (Jones & Callan, 2003; Nishitani & Hari, 2002; Calvert et al., 2000). Although the reconstructed activations are probably accurate only to within 1 or 2 cm, we observed a clear distinction between supra- and infra-sylvian responses. This distinction was more apparent on the left side for phonetic computations and on the right side for gender computations (Figure 5). Three types of response profile were observed. First, in the left superior temporal gyrus, activity was similar for vowel congruent and incongruent conditions, suggesting that either this region does not compute cross-modal representation of vowels or is not sensitive to vowel repetition, coding features rather than complex auditory objects as suggested by Wessinger et al. (2001) in adults. Second, in the adjacent left ventral regions, activity was weaker for cross-modal vowel repetition than for a vowel change. The same phenomenon was observed on the right side for cross-modal gender repetition. This pattern of a weaker response for congruent than to incongruent stimuli in the temporal lobe fits with our previous brain sources modeling of auditory responses in infants using dipolar models (Dehaene-Lambertz & Baillet, 1998; Dehaene-Lambertz & Dehaene, 1994). In these studies, repetition of the auditory syllable induced a significant decrease in amplitude of both the ERPs and theirs generators as soon as the first repetition. As a decrease in firing has been recorded at the single-cell level with repetitive exposure to the same stimulus (Dehaene-Lambertz, Dehaene, et al., 2006; Desimone, 1996), this response decrease at the macroscopic level has been interpreted as signaling that the same neural representation was repeatedly accessed. Here, the weaker response for congruent trials establishes that the same neural representation was accessed first by the visual stimulus, then by the auditory stimulus, and thus, that these regions contain cross-modal neural populations

that represent phonetic and gender information irrespective of modality. This is in agreement with data from human adults (Beauchamp et al., 2004) and monkeys (Poremba et al., 2003) showing that regions responding to auditory and visual stimuli overlap along the entire length of the STS.

Third, cross-modal repetition enhanced activity relative to incongruent trials in inferior frontal regions. The localization may seem improbable, as the frontal regions are thought to be immature at this age. However, the frontal lobe should not be considered a homogeneous lobe. It sustains multiple functions that have different developmental time course. For example, the motor part of the frontal cortex develops very fast during the first year of life, whereas other portions have a protracted development until puberty. Using fMRI, it was shown that the dorsolateral prefrontal cortex is activated when awake and not sleeping 3-month-old infants were listening to their native language (Dehaene-Lambertz et al., 2002). At the same age, activity in Broca's area is also observed especially when sentences are repeated (Dehaene-Lambertz, Hertz-Pannier, et al., 2006). Grossmann, Johnson, Farroni, and Csibra (2007) recorded a peak of gamma activity over the right prefrontal cortex when faces with direct gaze were presented to infants. These studies demonstrate that although immature, the frontal lobe is certainly not a silent region at this age. Being affected by the prior representation of visual stimuli, the frontal region is thus a cross-modal region. This is also the case in monkeys, where only small sections of the inferior frontal region appear to be modality-specific (Poremba et al., 2003). In human adults, activity in frontal areas as measured with fMRI is not always reported in natural cross-modal speech perception. However, these regions are recruited on the left side when visual and auditory information conflicts and when syllables are presented in a noisy environment (Ojanen et al., 2005; Callan et al., 2003), or on the right side when adults have to recognize a particular voice after having learned specific voice–face pairing (von Kriegstein & Giraud, 2006). At early stages of acquisition, frontal regions might be more commonly involved in infants than in adults.

Note also that both for vowel in the left and for gender in the right hemisphere, the frontal regions increased their activity for repetition contrary to what was observed for temporal regions. Repetition enhancement has been observed in adults and monkeys when the repeated stimulus needs to be maintained in working memory (Desimone, 1996). In an fMRI experiment, in which short sentences were presented every 14 sec to 3-month-old infants, an increase of activity was observed in the left inferior frontal region when sentences were repeated (Dehaene-Lambertz, Hertz-Pannier, et al., 2006). Behavioral experiments have shown that infants of this age are able to spontaneously maintain verbal content for several minutes. Two-month-old infants are able to detect a small phonetic changes between ''the rat chased white mice'' presented during a habituation phase and ''the cat . . .'' presented in the test phase, even though those phases were separated by 2 min of silent delay (Mandel, Jusczyk, & Nelson, 1994). They are also able to detect a change of word order between two sentences using the same experimental paradigm (Mandel, Kemler Nelson, & Jusczyk, 1996). Frontal regions may thus be already involved at this age in short-term memory, and here might be involved in storing the characteristics of the standard stimuli throughout the block.

The observation of a gender MMR that can be explained only by learned associations between some voice features (e.g., low vs. high pitch) and face characteristics (e.g., thin vs. square face) demonstrates that infants possess an efficient capacity to match visual and auditory cues at the perceptive level. Thus, it is possible that phonetic cross-modal representations might be constructed as gender representations from visual and auditory representations without postulating a motor component. The frontal lobes may be crucially involved in such learning, by coding the associative significance of the stimulus independent of its physical properties, as it has been described for some frontal areas in monkeys (Watanabe, 1992). For phonetic knowledge, the frontal activity revealed by source modeling may be an important node in binding comprehension and production of speech. This region which is close to the motor cortex, and in which audiovisual mirror neurons are observed in humans (Buccino et al., 2001) as in monkeys (Kohler et al., 2002), may be critically involved in driving motor learning in infants by matching infants' motor output with stored audiovisual templates. The precise nature of the role of the frontal regions in infant learning needs to be explored further.

Conclusion

Our results do not reveal a ''blooming buzzing confusion'' (James, 1892) in the infant brain but a complex interplay between a large set of brain areas with differentiated functions within hemispheres, and lateralization of the processing of different features of the same stimulus to opposite hemispheres. Although these regions and their interconnections mature at different rates, we show that they are already sufficiently functional to sustain cross-modal representations by 10 weeks of age. Progress in brain imaging techniques usable in infants should help us to understand how this organization of the human brain assists learning during the first weeks of life.

Acknowledgments

This study was supported by the McDonnell foundation, APETREIMC, the European NEST-PATHFINDER program 'Neurocom' (Contract no. 12738), l'Agence Nationale pour la Recherche NIBB, and the Wellcome Trust. We thank C. Billard, P. Landrieu, M. Tardieu, and S. Dehaene for their support.

Reprint requests should be sent to Ghislaine Dehaene-Lambertz, Unité INSERM 562, CEA/SAC/DSV/DRM/NeuroSpin, Bat145, point courrier 156, 91191 Gif/Yvette, France, or via e-mail: ghislaine.dehaene@cea.fr.

Note

1. Although the average number of trials might seem small for each condition, the event-related voltage is large in infants, due to the conductive properties of their tissues. ERPs were clearly visible in each subject.

REFERENCES

Baillet, S., Mosher, J., & Leahy, R. (2001). Electromagnetic brain mapping. *IEEE Signal Processing Magazine, 18,* 14–30.

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., & Martin, A. (2004). Unraveling multisensory integration: Patchy organization within human STS multisensory cortex. *Nature Neuroscience, 7,* 1190–1192.

Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences, 8,* 129–135.

Blass, E. M., & Camp, C. A. (2003). Biological bases of face preference in 6-weeks-old infants. *Developmental Science, 6,* 524–536.

Brookes, H., Slater, A., Quinn, P. C., Lewkowicz, D. J., Hayes, R., & Brown, E. (2001). Three-month-old infants learn arbitrary auditory–visual pairings between voices and faces. *Infant and Child Development, 10,* 75–82.

Buccino, G., Binkofski, F., Fink, G. R., Fadiga, L., Fogassi, L., Gallese, V., et al. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: An fMRI study. *European Journal of Neuroscience, 13,* 400–404.

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

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *NeuroReport, 14,* 2213–2218.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10,* 649–657.

Cheour, M., Ceponiene, R., Lehtokoski, A., Luuk, A., Allik, J., Alho, K., et al. (1998). Development of language-specific phoneme representations in the infant brain. *Nature Neuroscience, 1,* 351–353.

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk–MacDonald effect: A phonetic representation within short-term memory. *Clinical Neurophysiology, 113,* 495–506.

Dehaene-Lambertz, G. (2000). Cerebral specialization for speech and non-speech stimuli in infants. *Journal of Cognitive Neuroscience, 12,* 449–460.

Dehaene-Lambertz, G., & Baillet, S. (1998). A phonological representation in the infant brain. *NeuroReport, 9,* 1885–1888.

Dehaene-Lambertz, G., & Dehaene, S. (1994). Speed and cerebral correlates of syllable discrimination in infants. *Nature, 370,* 292–295.

Dehaene-Lambertz, G., Dehaene, S., Anton, J. L., Campagne, A., Ciuciu, P., Dehaene, G. P., et al. (2006). Functional segregation of cortical language areas by sentence repetition. *Human Brain Mapping, 27,* 360–371.

Dehaene-Lambertz, G., Dehaene, S., & Hertz-Pannier, L. (2002). Functional neuroimaging of speech perception in infants. *Science, 298,* 2013–2015.

Dehaene-Lambertz, G., & Gliga, T. (2004). Common neural basis for phoneme processing in infants and adults. *Journal of Cognitive Neuroscience, 16,* 1375–1387.

Dehaene-Lambertz, G., Hertz-Pannier, L., Dubois, J., Meriaux, S., Roche, A., Sigman, M., et al. (2006). Functional organization of perisylvian activation during presentation of sentences in preverbal infants. *Proceedings of the National Academy of Sciences, U.S.A., 103,* 14240–14245.

Dehaene-Lambertz, G., Pallier, C., Serniklaes, W., Sprenger-Charolles, L., Jobert, A., & Dehaene, S. (2005). Neural correlates of switching from auditory to speech perception. *Neuroimage, 24,* 21–33.

Dehaene-Lambertz, G., & Pena, M. (2001). Electrophysiological evidence for automatic phonetic processing in neonates. *NeuroReport, 12,* 3155–3158.

Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention. *Proceedings of the National Academy of Sciences, U.S.A., 93,* 13494–13499.

Desjardins, R. N., & Werker, J. F. (2004). Is the integration of heard and seen speech mandatory for infants? *Developmental Psychobiology, 45,* 187–203.

Dubois, J., Dehaene-Lambertz, G., Perrin, M., Mangin, J. F., Cointepas, Y., Duchesnay, E., et al. (2008). Asynchrony of the early maturation of white matter bundles in healthy infants: Quantitative landmarks revealed noninvasively by diffusion tensor imaging. *Human Brain Mapping, 29,* 14–27.

Dubois, J., Hertz-Pannier, L., Dehaene-Lambertz, G., Cointepas, Y., & Le Bihan, D. (2006). Assessment of the early organization and maturation of infants' cerebral white matter fiber bundles: A feasibility study using quantitative diffusion tensor imaging and tractography. *Neuroimage, 30,* 1121–1132.

Eimas, P. D., Siqueland, E. R., Jusczyk, P. W., & Vigorito, J. (1971). Speech perception in infants. *Science, 171,* 303–306.

Erber, N. P. (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing Disorder, 40,* 481–492.

Ermer, J. J., Mosher, J. C., Baillet, S., & Leah, R. M. (2001). Rapidly recomputable EEG forward models for realistic head shapes. *Physics in Medicine and Biology, 46,* 1265–1281.

Feigenson, L. (2005). A double-dissociation in infants' representations of object arrays. *Cognition, 95,* B37–B48.

Floccia, C., Nazzi, T., & Bertoncini, J. (2000). Unfamiliar voice discrimination for short stimuli in newborns. *Developmental Science, 3,* 333–343.

Giard, M. H., Lavikainen, J., Reinikainen, R., Perrin, F., Bertrand, O., Pernier, J., et al. (1995). Separate representations of stimulus frequency, intensity, and duration in auditory sensory memory: An event-related potential and dipole-model analysis. *Journal of Cognitive Neuroscience, 7,* 133–143.

Gibson, A., Bayford, R. H., & Holder, D. S. (2000). Two-dimensional finite element modelling of the neonatal head. *Physiological Measurement, 21,* 45–52.

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America, 108,* 1197–1208.

Grossmann, T., Johnson, M. H., Farroni, T., & Csibra, G. (2007). Social perception in the infant brain: Gamma oscillatory activity in response to eye gaze. *Scan, 2,* 284–291.
Izard, V., Dehaene-Lambertz, G., & Dehaene, S. (2008). Distinct cerebral pathways for object identity and number in human infants. *PLoS Biology, 6,* e11.
Jaaskelainen, I. P., Ojanen, V., Ahveninen, J., Auranen, T., Levanen, S., Mottonen, R., et al. (2004). Adaptation of neuromagnetic N1 responses to phonetic stimuli by visual speech in humans. *NeuroReport, 15,* 2741–2744.
Jacquemot, C., Pallier, C., Le Bihan, D., Dehaene, S., & Dupoux, E. (2003). Phonological grammar shapes the auditory cortex: A functional magnetic resonance imaging study. *Journal of Neuroscience, 23,* 9541–9546.
James, W. (1892). Psychology. Briefer Course. In: *Writings (1878–1899)* (pp. 18–34). New York: Holt (published again in Library of America, 1212 pp., 1992).
Jones, J. A., & Callan, D. E. (2003). Brain activity during audiovisual speech perception: An fMRI study of the McGurk effect. *NeuroReport, 14,* 1129–1133.
Kohler, E., Keysers, C., Umilta, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science, 297,* 846–848.
Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior and Development, 6,* 263–285.
Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science, 218,* 1138–1141.
Leinbach, M. D., & Fagot, B. I. (1993). Categorical habituation to male and female faces: Gender schematic processing in infancy. *Infant Behavior and Development, 16,* 317–332.
Liberman, A. M. (1996). *Speech: A special code*. Cambridge: Bradford Books/MIT Press.
Mandel, D. R., Jusczyk, P. W., & Nelson, D. G. (1994). Does sentential prosody help infants organize and remember speech information? *Cognition, 53,* 155–180.
Mandel, D. R., Kemler Nelson, D. G., & Jusczyk, P. W. (1996). Infants remember the order of words in a spoken sentence. *Cognitive Development, 11,* 181–196.
Mareschal, D., & Johnson, M. H. (2003). The ''what'' and ''where'' of object representations in infancy. *Cognition, 88,* 259–276.
Mattout, J., Phillips, C., Penny, W. D., Rugg, M. D., & Friston, K. J. (2006). MEG source localization under multiple constraints: An extended Bayesian framework. *Neuroimage, 30,* 753–767.
McCarthy, G., & Wood, C. C. (1985). Scalp distributions of event-related potentials: An ambiguity associated with analysis of variance models. *Electroencephalography and Clinical Neurophysiology, 62,* 203–208.
McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748.
Meltzoff, A. N., & Borton, R. W. (1979). Intermodal matching by human neonates. *Nature, 282,* 403–404.
Miller, C. L. (1983). Developmental changes in male/female voice classification by infants. *Infant Behavior and Development, 6,* 313–330.
Näätänen, R. (1990). The role of attention in auditory information processing as revealed by event-related potentials and other brain measures of cognitive function. *Behavioral and Brain Sciences, 13,* 201–288.
Näätänen, R., Lehtokovski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., et al. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature, 385,* 432–434.
Nishitani, N., & Hari, R. (2002). Viewing lip forms: Cortical dynamics. *Neuron, 36,* 1211–1220.
Ojanen, V., Mottonen, R., Pekkola, J., Jaaskelainen, I. P., Joensuu, R., Autti, T., et al. (2005). Processing of audiovisual speech in Broca's area. *Neuroimage, 25,* 333–338.
Patterson, M. L., & Werker, J. F. (2002). Infants' ability to match dynamic phonetic and gender information in the face and voice. *Journal of Experimental Child Psychology, 81,* 93–115.
Patterson, M. L., & Werker, J. F. (2003). Two-month old infants match phonetic information in lips and voice. *Developmental Science, 6,* 193–198.
Pena, M., Maki, A., Kovacic, D., Dehaene-Lambertz, G., Koizumi, H., Bouquet, F., et al. (2003). Sounds and silence: An optical topography study of language recognition at birth. *Proceedings of the National Academy of Sciences, U.S.A., 100,* 11702–11705.
Poremba, A., Saunders, R. C., Crane, A. M., Cook, M., Sokoloff, L., & Mishkin, M. (2003). Functional mapping of the primate auditory system. *Science, 299,* 568–572.
Quinn, P. C., Yahr, J., Kuhn, A., Slater, A. M., & Pascalils, O. (2002). Representation of the gender of human faces by infants: A preference for female. *Perception, 31,* 1109–1121.
Reale, R. A., Calvert, G. A., Thesen, T., Jenison, R. L., Kawasaki, H., Oya, H., et al. (2007). Auditory–visual processing represented in the human superior temporal gyrus. *Neuroscience, 145,* 162–184.
Richards, J. E. (2005). Localizing cortical sources of event-related potentials in infants' covert orienting. *Developmental Science, 8,* 255–278.
Rosenblum, L. D., & Saldana, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 318–331.
Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics, 59,* 347–357.
Saint-Amour, D., De Sanctis, P., Molholm, S., Ritter, W., & Foxe, J. J. (2007). Seeing voices: High-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the McGurk illusion. *Neuropsychologia, 45,* 587–597.
Schwartz, J., & Tallal, P. (1980). Rate of acoustic change may underlie hemispheric specialization for speech perception. *Science, 207,* 1380–1381.
Sergent, C., Baillet, S., & Dehaene, S. (2005). Timing of the brain events underlying access to consciousness during the attentional blink. *Nature Neuroscience, 8,* 1391–1400.
Shestakova, A., Brattico, E., Huotilainen, M., Galunov, V., Soloviev, A., Sams, M., et al. (2002). Abstract phoneme representations in the left temporal cortex: Magnetic mismatch negativity study. *NeuroReport, 13,* 1813–1816.
Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex, 17,* 2387–2399.
Summerfield, Q., & McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *The Quarterly Journal of Experimental Psychology Section A, 36,* 51–74.
Tervaniemi, M., Medvedev, S. V., Alho, K., Pakhomov, S. V., Roudas, M. S., Van Zuijen, T. L., et al. (2000). Lateralized automatic auditory processing of phonetic versus musical information: A PET study. *Human Brain Mapping, 10,* 74–79.
Tremblay, K., Kraus, N., & McGee, T. (1998). The time course of auditory perceptual learning: Neurophysiological changes during speech–sound training. *NeuroReport, 9,* 3557–3560.

Tzourio, N., de Schonen, S., Mazoyer, B., Bor, A., Pietrzyk, U., Bruck, B., et al. (1992). Regional cerebral blood flow in two-month-old alert infants. *Society for Neuroscience Abstracts, 18,* 1121.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences, U.S.A., 102,* 1181–1186.

von Kriegstein, K., & Giraud, A. L. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biology, 4,* e326.

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., & Giraud, A. L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience, 17,* 367–376.

Walker-Andrews, A. S., Bahrick, L. E., Raglioni, S. S., & Diaz, I. (1991). Infants' bimodal perception of gender. *Ecological Psychology, 3,* 55–75.

Wallace, M. T., Carriere, B. N., Perrault, T. J., Jr., Vaughan, W., & Stein, B. E. (2006). The development of cortical multisensory integration. *Journal of Neuroscience, 26,* 11844–11849.

Watanabe, M. (1992). Frontal units of the monkey coding the associative significance of visual and auditory stimuli. *Experimental Brain Research, 89,* 233–247.

Wessinger, C. M., VanMeter, J., Tian, B., Van Lare, J., Pekar, J., & Rauschecker, J. P. (2001). Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. *Journal of Cognitive Neuroscience, 13,* 1–7.

Winkler, I., Cowan, N., Csépe, V., Czigler, I., & Näätänen, R. (1996). Interactions between transient and long-term auditory memory as reflected by the mismatch negativity. *Journal of Cognitive Neuroscience, 8,* 403–415.

Winkler, I., Kushnerenko, E., Horvath, J., Ceponiene, R., Fellman, V., Huotilainen, M., et al. (2003). Newborn infants can organize the auditory world. *Proceedings of the National Academy of Sciences, U.S.A., 100,* 11812–11815.

Yakovlev, P., & Lecours, A. R. (1967). The myelogenetic cycles of regional maturation of the brain. In A. Minkovski (Ed.), *Regional development of the brain in early life* (pp. 3–69). Oxford: Blackwell.

Zatorre, R. J., & Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cerebral Cortex, 11,* 946–953.

Zatorre, R. J., Evans, A. C., Meyer, E., & Gjedde, A. (1992). Lateralization of phonetic and pitch discrimination in speech processing. *Science, 256,* 846–849.