Published in final edited form as:
Neuropsychologia. 2007 January 7; 45(1): 93–106. doi:10.1016/j.neuropsychologia.2006.04.017.

Neural responses elicited to face motion and vocalization pairings

Aina Puce[*,†,‡], **James A Epling**[*], **James C Thompson**[*,†], and **Olivia K Carrick**[*]

[*] Center for Advanced Imaging, West Virginia University, Morgantown WV, USA

[†] Department of Radiology, West Virginia University, Morgantown WV, USA

[‡] Department of Neurobiology and Anatomy, West Virginia University, Morgantown WV, USA

Abstract

During social interactions our brains continuously integrate incoming auditory and visual input from the movements and vocalizations of others. Yet, the dynamics of the neural events elicited to these multisensory stimuli remain largely uncharacterized. Here we recorded audiovisual scalp event-related potentials (ERPs) to dynamic human faces with associated human vocalizations. Audiovisual controls were a dynamic monkey face with a species-appropriate vocalization, and a house with opening front door with a creaking door sound. Subjects decided if audiovisual stimulus trials were congruent (e.g. human face-human sound) or incongruent (e.g. house image-monkey sound). An early auditory ERP component, N140, was largest to human and monkey vocalizations. This effect was strongest in the presence of the dynamic human face, suggesting that species-specific visual information can modulate auditory ERP characteristics. A motion-induced visual N170 did not change amplitude or latency across visual motion category in the presence of sound. A species-specific incongruity response consisting of a late positive ERP at around 400 ms, P400, was selectively larger only when human faces were mismatched with a non-human sound. We also recorded visual ERPs at trial onset, and found that the category-specific N170 did not alter its behavior as a function of stimulus category – somewhat unexpected as two face types were contrasted with a house image. In conclusion, we present evidence for species-specificity in vocalization selectivity in early ERPs, and in a multisensory incongruity response whose amplitude is modulated only when the human face motion is paired with an incongruous auditory stimulus.

Keywords

biological motion; multisensory; faces; cross-modal integration; temporal cortex; mismatch

1. INTRODUCTION

Humans and other animals interpret social signals sent by body movements, facial gestures and vocalizations of conspecifics with little difficulty (Darwin, 1899/1998; Shettleworth, 2001; Tomasello & Call, 1997). For these multisensory social signals to make sense, the brain must rapidly process the movements of a facial or body gesture and integrate them with associated vocalizations (Partan & Marler, 1999). In macaques, auditory species-specific calls can strongly activate brain regions traditionally regarded to process visual category-specific information (Gil-da-Costa et al., 2004), and this multisensory social information appears to be already integrated in areas traditionally regarded as auditory cortex (Ghazanfar, Maier, Hoffman, & Logothetis, 2005). Recent neuroimaging studies in humans clearly indicate that

Address for correspondence: enter for Advanced Imaging, PO Box 9236, West Virginia University School of Medicine, Morgantown WV 26506-9236, USA, Email: apuce@hsc.wvu.edu, TEL: +1 304 293 5016, FAX: +1 304 293 4287.
Corresponding author: Aina Puce, Ph.D.

lateral temporal cortex responds selectively to hearing human vocalizations (Belin, Zatorre, & Ahad, 2002; Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; Binder et al., 2000). Activation in human lateral temporal cortex can also be differentiated across human versus animal vocalization domains, confirming species-specific processing for (verbal and non-verbal) human vocalizations (Fecteau, Armony, Joanette, & Belin, 2004). Additionally, animal vocalizations activate regions in temporal cortex anterior and superior to regions sensitive to tool sounds (Lewis, Brefczynski, Phinney, Janik, & DeYoe, 2005).

Neuroimaging studies in the healthy human brain also demonstrate discrete and separate activation in temporoparietal cortex to viewing static images of animals/humans as opposed to tools (Chao, Haxby, & Martin, 1999; Chao & Martin, 2000; Fang & He, 2005). Furthermore, animal images and sounds have been found to activate the neighboring regions of posterior superior temporal cortex (specifically superior temporal sulcus and gyrus) known to respond to viewing human face and body motion (Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004; Beauchamp, Lee, Haxby, & Martin, 2002; Bonda, Petrides, Ostry, & Evans, 1996; Grossman et al., 2000; Puce, Allison, Bentin, Gore, & McCarthy, 1998), indicating that these regions may well possess multisensory response properties.

The dynamics of human neural responses elicited to human facial motion and vocalization combinations are not understood and have not been well studied. In a very early study, neural correlates of audiovisual integration to viewing moving mouths and hearing syllables was said to occur within 180 ms post-stimulus over the temporal scalp (Sams et al., 1991). A comprehensive electrophysiological study to audiovisual stimulation consisting of simple auditory tones and visual shapes indicated that auditory unimodal responses occurred earlier than unimodal visual responses, and that neural activity associated with audiovisual integration (around 140–165 ms) preceded that of unimodal visual stimulation alone (Giard & Peronnet, 1999). To date, there are few studies that have investigated the dynamics of the audiovisual human neural responses elicited to dynamic facial expressions and concurrent vocalizations.

Visual ERPs elicited to viewing human face and body motion, defined here as a *motion-sensitive N170*, typically also show a consistent negativity over the posterior temporal scalp that occurs in the 190–220 ms post-stimulus time range (Puce & Perrett, 2003; Puce, Smith, & Allison, 2000; Wheaton, Pipingas, Silberstein, & Puce, 2001). The amplitude of this motion-sensitive N170 ERP can vary as a function of face movement type (e.g. mouth opening vs closing, eye aversion vs direct gaze (Puce & Perrett, 2003; Puce, Smith & Allison, 2000) or body movement type (e.g. hand closing vs opening). The subsequent ERP activity typically shows positive components, P350 and P500, which vary their behavior as a function of task demand and social context (Puce & Perrett, 2003; Puce et al., 2000). Similarly, category-specific human scalp and intracranial event-related potentials (ERPs) elicited to viewing *static* human faces consist of a negativity over lateral and ventral temporal cortex that typically occurs at around 170–200 ms post-stimulus which is most prominent to faces – defined here as a *category-specific N170* (Allison et al., 1994; Bentin, Allison, Puce, Perez, & McCarthy, 1996). The exact relationship between the *category-specific N170* and the *motion-sensitive N170* is not well understood.

There are many unanswered questions relating to the processing of human facial movements and associated vocalizations by the human brain. For instance, what are the dynamics of human neural responses elicited to viewing human facial movements in the presence of non-verbal vocalizations? Are they species-specific and therefore different from neural responses to facial movements and vocalizations of other primates? Are they different to multisensory neural responses elicited to audiovisual stimuli involving inanimate objects? Does the nature of these audiovisual neural responses change to incongruous or implausible combinations of these categories?

In this study, we sought answers to these four questions in a complex audiovisual experimental paradigm. Stimulus categories are outlined in Figure 1. Subjects viewed apparent visual motion stimuli (Fig. 1 column and Fig. 2) which were presented with concurrent sounds (Fig. 1 row and Fig, 2). With respect to the first question regarding the characterization of ERPs to human facial motion and an associated human vocalization, we were able to record clear and discernable ERPs to this complex stimulus. The most clearly seen components were a vertex-centered auditory component, N140, a visual motion-sensitive N170 which had a bilateral posterior temporal scalp distribution, a parietally centered P250 component that followed the N170, and a P400 component that had a broad distribution across the posterior scalp. With respect to the question about the potential species-specific nature of these audiovisual ERPs, our experimental paradigm included a condition where a non-human primate face was presented opening its mouth, similar to the human face, and with a scream vocalization that human subjects identified as being non-human and belonging to a primate (Fig. 2). The ERP components elicited to the monkey face movement and associated monkey vocalization were similar in morphology and topography to ERPs elicited to the human face and human vocalization, arguing against the idea of species-specific effects. The question as to whether the ERPs are modulated by animacy was investigated by including a third stimulus type in the experimental design – an image of a house, where the apparent motion stimulus consisted of the front door opening with an associated sound of a creaking door (Fig. 2). Again, this ERP was similar in morphology and topography to the human and monkey ERPs. Animacy *per se* did appear to modulate some of our ERPs - the N140 was larger for human and monkey relative to the house sound stimuli. The motion-sensitive N170 was not influenced by this manipulation – somewhat surprisingly, the amplitude and latency of the response were identical across the three stimulus categories. Finally, with respect to the fourth question regarding congruity of the stimulus across the auditory and visual modalities, we presented each of the three visual stimuli with different types of concurrent sound (Fig. 1), which were matched (e.g. human face-human sound) or mismatched (e.g. monkey face-house sound). A species-specific effect was noted, in that N140 amplitude was largest to the human face when paired with the human sound. Additionally, the P400 ERP was selectively larger when only the human face was mismatched with a non-human sound. In addition, our experimental design allowed us to record *category-specific N170s* in response to the initial onset of the visual stimulus as well as *motion-specific N170s* elicited to the apparent visual motion stimulus with the associated sound. The scalp distributions of the two types of N170 were similar: a predominant negativity was seen at the bilateral posterior temporal scalp, consistent with previous literature (e.g. Bentin et al., 1996;Puce et al., 2000). The latencies of the category-specific N170 elicited to stimulus onset were much shorter than those of the motion-sensitive N170 – in line with previous literature where the two responses have been studied separately.

2. METHODS

2.1 Subjects

Fourteen healthy volunteers (7 females, 7 males) ages 22–56 years (mean age 31.1 + 10.2 years) participated in the study. There were 12 right-handers and 2 left-handers. Subjects had no previous history of neurological abnormalities, had normal or corrected to normal vision, and reported having normal hearing, although this was not formally tested. Subjects gave informed consent for a protocol approved by the West Virginia University Internal Review Board for Human Research Subjects.

2.2 Stimuli

Three types of visual stimuli were used and consisted of single exemplars of grayscale images of a human face, a monkey face, and a residential house (Fig. 1). Each stimulus type had two configurations: the human and monkey faces were depicted with mouths open and closed, and the house was shown with its front door open and closed – each visual image pair was presented in rapid sequence to produce an apparent motion stimulus (Fig. 2). The grey scale images were corrected to each have comparable brightness and contrast using Photoshop 7.0 software (Adobe Systems Inc.).

Three types of auditory stimuli were used and consisted of a human burp, a monkey scream, and a door squeaking. The loudness and duration of each stimulus were equated to 600 ms using Creative Wave Studio 4.10 (Creative Technology Ltd.) software. Additionally, harmonic-to-noise ratios (HNRs) (Lewis et al., 2005; Riede, Herzel, Hammerschmidt, Brunnberg, & Tembrock, 2001) were calculated for each auditory stimulus. The HNRs were 11.59, 17.21, and 15.28, respectively for the human burp, the monkey scream and the creaking door sound. The HNRs for our stimuli were found to be comparable to one another when contrasted with a larger sample and wider set of natural sound categories (Lewis et al., 2005).

The visual and auditory stimuli were played concurrently to produce an apparent visual motion stimulus that occurred with a sound of identical duration.

2.3 Experimental paradigm

The experimental paradigm was made up of a series of 450 trials. A typical trial timeline is shown in Fig. 2. Each trial consisted of the presentation of an initial visual stimulus without sound (Stimulus 1) which remained on the screen for 1000 ms. This allowed neural activity that was explicitly associated with visual stimulus onset to proceed through its course. This was followed by Stimulus 2, where an instantaneous change in the visual display occurred, accompanied by a binaural audio stimulus. The duration of this second visual stimulus and its accompanying sound was 600 ms. At 1600 ms the visual stimulus then returned to its initial state (Stimulus 3, same as Stimulus 1) with no accompanying sound. At 3890 ms the stimulus display was replaced by a black background with no accompanying auditory stimulation. The black background was presented for a total period of 1000 ms prior to a new image being presented to start the next trial.

Various combinations of the audiovisual pairings in Stimulus 2 produced a set of nine trial types consisting of three Match (congruous) and six Mismatch (incongruous) trials (Fig. 1). Match trials were:

1. *Human-Human* i.e. human face and human vocalization. Here the subject saw the image of the face initially appear with closed mouth (1000 ms), then the mouth opened and a burping sound was heard (600 ms), and finally the mouth closed and remained in that position (1000 ms);

2. *Monkey-Monkey* i.e. monkey face and monkey vocalization. This trial type was similar to that of the human face, except that here when the monkey's mouth opened a monkey scream was heard;

3. *House-House* i.e. house and creaking door sound. Here the house was initially presented with a closed front door, and then the front door opened with a creaking sound. The trial ended with the house being shown with the front door closed using identical timing to the other two trial types.

Mismatch trials were:

4. *Human-Monkey:* human face and monkey vocalization (scream);

5. *Human-House:* human face and house sound (opening door);

6. *Monkey-Human:* monkey face and human vocalization (burp);

7. *Monkey-House*: monkey face and house sound (opening door);

8. *House-Human*: house and human vocalization (burp);
9. *House-Monkey*: house and monkey vocalization (scream).

Trial types were randomized and presented in random sequence in four individual runs of approximately six minutes each with rest periods in between. Over the entire four run experiment a total of 50 trials of each of the nine types were presented. Subjects were asked to fixate on the screen's center and indicate whether each presented audiovisual combination was a Match or a Mismatch with a two-button response following auditory stimulus offset.

Experiments were conducted in a darkened, quiet room. The subject sat in a comfortable chair approx. 5 feet away from a 13 inch computer monitor. All visual stimuli subtended a visual angle of 2.9° × 2.9°. Binaural auditory stimuli were presented using low-conductance ear tubing, at a level that was comfortable to each subject.

2.4 Electroencephalographic (EEG) Recordings

A high-density, continuous EEG was recorded for each subject using a 124-channel Electrocap during each of the four experimental runs with a band pass filter of 0.1–100 Hz and a gain of 5,000. The vertical electro-oculogram (vEOG) was recorded from bipolar electrodes above and below the left eye, and the horizontal EOG (hEOG) was recorded from bipolar electrodes at the outer canthi of both eyes. Electrode impedances were kept below 15 kOhms.

After the four experimental runs were completed, the positions of the recording electrodes in 3D space were digitized using a Polhemus 3 SpaceFastrak digitizer. The coordinates of the recording electrodes for each subject were stored as a separate file. A grand average of the recording electrode positions across the 14 subjects was made by averaging the electrode files for the individual subjects. This grand averaged electrode file was used when topographic voltage maps of ERP activity were generated.

2.5 EEG Analysis

In an offline analysis procedure, the continuous EEG was first cut-down into 4044 ms segments, spanning the presentation of Stimuli 1–3 and including the time at which subjects made a behavioral response. The EEG segment was further cut down to into epochs. Epoch 1 contained the visual stimulus that initiated the trial (Stimulus 1) (Fig. 2). This epoch had a duration of 1028 ms, and its baseline was corrected using the 100 ms pre-stimulus to the presentation of Stimulus 1. Epoch 2 contained the onset of the audiovisual stimulus pair (Stimulus 2) as well as the offset of the sound stimulus and the return of the visual stimulus to baseline (Stimulus 3) 600 ms after the presentation of Stimulus 2 (Fig. 2). Epoch 2 had a duration of 2144 ms and its baseline was corrected using the 100 ms pre-stimulus period prior to Stimulus 2, and included the ERPs elicited to Stimuli 2 and 3. We did not generate a third epoch to examine the ERPs to Stimulus 3 (onset of visual Stimulus 3, and offset of audiovisual Stimulus 2). For each EEG epoch, two separate artifact rejection and baseline correction procedures were made prior to averaging ERPs. Trials containing hEOG, vEOG and EMG artifacts were rejected prior to averaging by setting an amplitude exclusion criterion of greater than ±75 μVolts for each epoch.

For each subject, the artifact rejected and baseline corrected EEG data for Epochs 1 and 2 from the four runs for each of the nine conditions were averaged using Neuroscan 4 software (Compumedics USA, El Paso TX). This yielded a total of 18 ERP files per subjects. Digital filtering was performed on each ERP file using a low-pass cutoff frequency of 30 Hz (6dB/oct) and a zero phase shift. Grand average ERPs for Epochs 1 and 2 were then obtained from the average ERPs of the 14 subjects. The main focus of the experiment was on the audiovisual ERPs elicited in Epoch 2 with responses recorded to match and mismatch trials, and to a lesser extent on Epoch 1 and ERP differences to Stimulus 1 (onset of visual stimulation to three visual stimulus types).

Grand averaged ERP data were qualitatively inspected in two ways. First, pseudotopographic displays of the ERPs associated with each electrode were viewed with the Neuroscan 4 software. Main ERP component peak latencies and the latency range for peak duration were noted from inspection of individual subject ERPs and the grand average ERP. For later ERP activity that did not have a clear well defined peak, but instead consisted of a slow persistent potential, latency ranges for area under the curve (AUC) measures were determined. Second, topographic voltage maps at appropriate time points were created using the EMSE software suite (Source Signal Imaging, San Diego CA). Topographic voltage maps were displayed with the digitized electrode positions.

From the topographical voltage displays of both individual subject ERPs and the grand average ERP, electrodes of interest were chosen for further scrutiny and quantitative analysis based on their amplitude distributions in our high-density recording protocol. We chose to average ERP amplitudes (and latencies) across a cluster of electrodes in each component’s main amplitude distribution for each individual subject, so that for the same sensors we could ensure that the voltage maxima of each component were included. This is an important difference between our high-density study and previous studies using 10–20 and 10–10 system based electrode positions. An average latency amplitude and latency were generated from the ERPs of the selected electrodes and these data formed the input for subsequent statistical analysis (Puce et al., 2000).

Finally, surface Laplacians (Nunez et al., 1994; Perrin, Pernier, Bertrand, & Echallier, 1989), or the second spatial derivative of the topographic voltage map, were generated using EMSE software. The surface Laplacian has been postulated to be a good approximation to the potential at the underlying dural surface (Nunez, 1987). Surface Laplacians, also known as Scalp Current Density (SCD) plots can add in the identification of underlying neural sources. Typically, Laplacians are biased towards the display of activity of shallow versus deeper cortical generators (Giard & Peronnet, 1999).

Quantitative data from individual subject ERPs were determined for Epoch 2, where audiovisual stimulation was presented. Main ERP component peak latencies and amplitudes and AUC measures were calculated using a semi-automated procedure in the Neuroscan 4 software from latency ranges chosen from the qualitative data inspection described earlier. (Auditory) N140 and (visual) N170 peaks were detected across 100–200 ms and 145–220 ms post-stimulus time ranges, respectively, whereas AUC measures for P250 and for P400 were made over 268–424 ms and 424–664 ms post-stimulus time ranges, respectively. In Epoch 1, where visual only stimulation was present, P100, N170 and P250 were detected across these three separate time intervals: 80–140 ms, 150–190 ms, and 208–280 ms. Data files of ERP peak amplitude, latency and AUC measures were stored as ASCII files and imported into Microsoft Excel spreadsheets. Data from electrodes of interest were extracted and average measures of ERP peak latency, amplitude or AUC were created across selected electrode clusters, identified in the qualitative analysis.

2.6 Statistical Analysis

All statistical analyses were performed using standard SPSS V9 software (SPSS Inc, Chicago IL). Three sets of analyses were run in order to determine what differences there were between the: (i) general categories in the audiovisual Match conditions (Analysis 1); (ii) audiovisual Mismatch conditions within each visual stimulus type, where the audiovisual Match and the Mismatch conditions were compared within each visual stimulus type (Analysis 2); (iii) category differences in visual onset ERPs at the beginning of each trial (Analysis 3).

2.6.1 Analysis 1 (Audiovisual Match conditions)—ERP peak latencies, amplitudes, and AUC measures each were compared across the Match conditions using 2 × 3 ANOVAs with main effects of Condition (Human, Monkey, House) and Region. Contrasts were then run on data from significant main effects or interaction effects to identify which conditions were significantly different.

2.6.2 Analysis 2 (Audiovisual Match versus Mismatch conditions)—ERP peak latencies, amplitudes, and AUC measures each were compared with each set of Mismatch conditions e.g. Human (Human-Human, Human-Monkey, Human-House). A 2 × 3 ANOVA with main effects of Condition (Human-Human, Human-Monkey, Human-House) and Region was run. Similar ANOVAs were run for the Monkey and House conditions. Contrasts were then run on data from significant main effects or interaction effects to identify which conditions were significantly different.

2.6.3 Analysis 3 (Visual onset ERPs at the start of each trial)—Stimulus onset at the beginning of each experimental trial produced a visual ERP. These data were also analyzed using a similar ANOVA based analysis approach to that outlined above. ERP peak latencies, amplitudes, and AUC measures each were compared across the three visual stimulus onset conditions using 2 × 3 ANOVAs with main effects of Condition (Human, Monkey, House) and Hemisphere (Right, Left). Contrasts were then run on data from significant main effects or interaction effects to identify which conditions were significantly different.

3. RESULTS

3.1. Overall characteristics of ERP waveforms to combined audiovisual stimulation

Audiovisual Match conditions—Four prominent ERP components were elicited to the audiovisual stimulus combinations that comprised Stimulus 2: N140, N170, P250, and P400. The voltage distribution of the N140 was focused over the central scalp (Fig. 3B top row), and the Laplacian indicated a potential active source overlying the right centrotemporal region (Fig. 3C top row). This predominantly auditory related ERP was very similar across stimulus types – as shown by the similarity in the topographic voltage and Laplacian maps across conditions. The N140 appeared to be larger for the human and monkey stimuli, relative to the house stimulus (See Fig. 1, top row). The next prominent ERP component was the N170 – seen over the bilateral temporal scalp (Fig. 3B second row), with a putative active generator being confined to each temporal lobe (Fig. 3C, second row). It was also seen clearly in response to all three stimulus types. Two subsequent later positive potentials were seen later in the epoch – the P250 and the P400. Both potentials had broad positivities that were centered on the parietal scalp (Fig. 3B, third and fourth rows). The Laplacian maps for each respective potential suggested that multiple sources may contribute to this activity (Fig. 3C, third and fourth rows). There was a frontocentral midline source and another more posterior occipital source again on the midline seen for the conditions where a human or monkey audiovisual stimulus was delivered. For the house audiovisual stimulus, the occipital source was also present, but this time there appeared to two sources sites in the left and right lateral frontal regions.

Audiovisual Mismatch conditions—The morphology and scalp topography of the ERPs to the Mismatch conditions was similar to the ERPs to the Match conditions, and differed only in terms of amplitude and latency (presented in section 3.2.2).

Visual onset ERPs elicited at the start of each trial—The visual ERPs that were elicited to stimulus onset showed a morphology consisting of a P100-N170-P250 triphasic complex that was seen to all three visual stimulus categories. The initial P100 consisted of a substantial positivity elicited due the large contrast change in the display arising from a stimulus on a white background appearing on a black screen. This resulted in the N170 also being positive in amplitude, but of course being negative relative to its flanking P100 and P250 neighbors.

3.2. Statistical Analysis of ERP data

3.2.1 Analysis 1: Audiovisual Match Conditions—For each visual stimulus there was a single auditory stimulus that constituted its 'Match' e.g. Image-Sound combinations of Human-Human, Monkey-Monkey, and House-House. Separate analyses of ERP peak latency and amplitude data were performed for each of the respective most prominent ERP components seen to the congruous or Matched audiovisual stimulus categories.

N140: As already described, N140 showed a maximal distribution over the frontocentral scalp (Fig. 3). We selected 3 electrodes from the midline at the vertex, and 2 frontocentral electrodes in each hemisphere that were within the maximal amplitude distribution (highlighted as white circles in Fig. 3B). Average ERP peak latency and amplitude measures were made for each subject. For the group overall, average N140 amplitude and latency across both hemispheres for the Human, Monkey and House Match conditions were −7.34, −6.34, and −3.64 μV, and 148, 145, and 143 ms, respectively. A 2-way repeated measures ANOVA show a significant main effect of N140 amplitude for Stimulus Type (Human, Monkey, House) (F[2,26] = 8.63, p =0.001), but not for Region (Left, Midline, Right). There was no significant interaction effect. Contrasts indicated that the main effect of Stimulus type resulted from a significantly smaller response to the House condition with respect to both the Human (F[1,13] = 49.17, p < 0.0001) and the Monkey (F[1,13] = 6.38, p = 0.03) conditions. There was no significant difference between the Human and Monkey conditions. The ANOVA for N140 latency indicated that there were no significant effects of Stimulus Type or Region, nor was there a significant interaction effect.

N170: Four electrodes from each hemisphere showing maximal N170 amplitude over each temporal scalp were selected (highlighted as white circles in Fig. 3B) for the calculation of mean N170 amplitude and latency values in each subject. Mean N170 amplitudes and latencies across the Human, Monkey and House Match conditions were −3.92, −4.74, and −3.45 μV and 201, 207, and 213 ms. Separate 2-way repeated measures ANOVA for main effects of Stimulus (Human, Monkey, House) and Region (Left Temporal, Right Temporal) were performed for N170 amplitude and for N170 latency. There were no significant main effects or a significant interaction effect for either N170 peak amplitude or latency.

P250: This first of two later positivities occurred in a more widespread scalp distribution than N170, but included the posterior temporal scalp. Hence, the same 4 electrodes from each temporal scalp were used to generate average P250 area under the curve (AUC) measures. Mean P250 AUC in the left hemisphere were 721, 463, and 730 for the human, monkey and house Match conditions. In the right hemisphere AUC values of 733, 439, and 773 were obtained for the 3 conditions. A 2-way repeated measures ANOVA showed a significant main effect for Stimulus type (Human, House, Monkey) (F[2,26]=4.89, p = 0.025). Contrasts indicated that this effect was due to a significantly smaller response in the Monkey relative to the Human (F[1,13] = 12.40, p < 0.005) and House (F[1,13] = 5.47, p = 0.04) conditions, and no difference between the Human and House conditions. No significant main effect of Region was observed, nor was the interaction effect significant.

P400: AUCs from 4 electrodes over each temporal scalp were averaged together, and P400 AUCs from electrodes over the midline were also averaged together (electrodes highlighted in Fig. 3B). For the posterior temporal scalp, the mean P400 AUCs for the human, house and monkey Match conditions were 1365, 1462, and 1833, respectively. The 2-way repeated measures ANOVA indicated a significant main effect for Stimulus Type (F[2, 26] = 4.38, p =

0.023), but not for Region. Contrasts revealed that this difference was driven by a significantly larger AUC for the House condition relative to the Human condition (F[1,13]=8.78, p = 0.01), but not for the Monkey condition. The interaction effect was not significant.

For the vertex P400 data mean AUCs for the Human, Monkey and House Match conditions were 961, 966, and 1533. The 1-way repeated measures ANOVA trended toward a main effect for Stimulus Type (F[2, 26]=2.83, p = 0.097). Contrasts indicated that there was a difference between the Human and House conditions (F[1,13]=6.37, p = 0.025), but not between the Human and Monkey conditions (F[1,13]=0.00, p = 0.98) or between the Monkey and House conditions (F[1,13]=2.55, p = 0.134].

3.2.2 Analysis 2: Audiovisual Match vs Mismatch Conditions—Separate analyses for selected ERP attributes (see mean values in Tables 1 and 2 for each condition) were performed for the Match and associated Mismatch conditions paired with each visual image, given the observed differences between visual stimulus types in the Match conditions. A repeated measures ANOVA was performed for each of the three audiovisual conditions associated with a particular visual stimulus type, so that the effects of mismatch could be evaluated.

N140: For N140 amplitude there was a main effect of Stimulus Type (F[2,26] = 6.41, p = 0.005, and see also Table 2) when the audiovisual pairings for the Human image were compared. Contrasts revealed the Human-human condition N140s had much larger amplitudes than the Human-Monkey (F[1,13] = 9.59, p = 0.009), or Human-House (F[1,13] = 13.09, p =0.003) conditions. There was no main effect of Region, nor was there an interaction effect between Stimulus Type and Region. For the ANOVA performed for the Monkey image paired with different sounds there were no significant main effects or interaction effects for N140 amplitude. Finally, for the House image the ANOVA analysis revealed a significant main effect for Stimulus Type (F[2,26] = 4.57, p = 0.02), but no main effect for Region, or no interaction effect. Contrasts revealed that N140 amplitude was larger for the House-Human relative to the House-House condition (F[1,13] = 6.28, p = 0.03). There was a trend for N140 to be larger for the House-Human relative to the House-Monkey condition (F[1,13] = 3.21, p = 0.10), but this was caused by a larger N140 in the left frontotemporal scalp (F[1,13] = 5.10, p = 0.04). These effects can be seen in the group average ERPs shown in Figure 4.

For N140 latency, in general no significant main effects were observed in any of the ANOVAs which were performed on the ERP data elicited to each of the audiovisual stimulus pairings (see Table 2). There was a trend for a main effect of Region for the House image dataset (F [2,26] = 3.61, p = 0.08), and contrasts did not indicate significant differences between the conditions. There was a significant interaction effect between Stimulus Type and Region (F [2,26] = 4.19, p = 0.03). Contrasts revealed that this effect was caused by longer latencies in the House-Monkey pairing, and House-House pairings relative to the House-Human pairing in the left hemisphere (F[1,13] = 5.72, p = 0.03; F[1,13] = 7.10, p = 0.02).

N170: The ANOVAs performed on N170 latency and amplitude data revealed that there were no significant differences between the matched and mismatched sounds for any visual image type, nor were there any interaction effects. The lack of any statistical difference between the congruous and incongruous audiovisual pairings underscores the strong visual motion-sensitive bias of this ERP component (see Fig. 5, and also Table 1).

P250: For AUC data (Fig. 5 and Table 1) the only significant difference that was seen in the ANOVA analysis occurred for the Monkey image and the associated sounds (F[2,26] = 3.70, p = 0.04). Contrasts revealed that this effect was due to a significant difference between both mismatched Monkey-Human and Monkey-House conditions (F[1,13] = 4.80, p = 0.05) and the (matched) Monkey-Monkey and (mismatched) Monkey-House sounds (F[1,13] = 5.47, p = 0.04).

P400: For AUC data (Fig. 5 and Table 1), the ANOVA revealed a significant effect for the Human image and associated sounds. The main effects showed trends, however, the interaction effect between Stimulus Type and Hemisphere was highly significant (F[2,26] = 6.35, p = 0.006). Contrasts revealed that this effect was driven by a difference between the Human and House sounds (main effect, F[1,13] = 24.96, p < 0.0001), and that the response was larger in the left hemisphere (interaction term, F[1,13] = 8.61, p = 0.01). No significant effects of match vs mismatch sounds were observed for the monkey face or the house.

3.2.3 Analysis 3: Visual Onset ERPs at the start of each trial—The start of each trial began with an initial visual stimulus that remained on the screen to allow the neural activity associated with this transition to die away prior to introducing the audiovisual stimulus pair. The main visual ERP activity to stimulus onset was observed over the posterior scalp, with all stimulus categories elicited reliable P100, N170 and P250 activity and an additional later negativity, the N320 (Fig. 6). ERP data were analyzed using a 2-way repeated measures ANOVA for main effect of Stimulus Type and Region, using a similar approach to that used for the data of Epoch 2.

P100 amplitudes were 7.34, 8.33, and 7.68 μV, and P100 latencies were 125, 125, and 125 ms, for the Human, Monkey, and House conditions. There was a trend toward a significant main effect of Stimulus Type (F[2,26]) = 2.94, p = 0.08), and interaction effect (F[2,26] = 3.02, p = 0.07), but no significant main effect of Region. Contrasts indicated that P100 amplitudes for monkey faces were larger than those for human faces (F[1,13] = 10.12, p = 0.007), and that this effect was larger in the left hemisphere (F[1,13] = 7.06, p = 0.02). There were no significant main effects or interaction effects for P100 latency.

N170 amplitudes were 2.25, 2.51, and 2.84 μV for the Human, Monkey and House conditions. The stimulus on a black background was presented from a uniform white background – producing a high contrast change. The N170s observed here were therefore positive in polarity. They are effectively riding on a large overall positivity (see Fig. 6). N170 latencies were 168, 169, and 166 ms for the three stimulus conditions. There was a tendency for N170 to be larger in the right hemisphere (F[1,13] = 3.99, p = 0.07), and there were no significant effects of Stimulus Type, nor was there an interaction effect between Stimulus Type and Region. For N170 latency there were no significant main effects of Stimulus Type or Region, however there was a trend toward a significant interaction effect (F[2,26] = 3.17, p = 0.07). Contrasts did not reveal significant differences between the pairs of conditions.

P250 amplitudes were 9.23, 9.70, and 10.88 μV, and latencies were 230, 236, and 231 ms, for the Human, House, and Monkey conditions. There was a trend toward a main effect of Stimulus Type (F[2,26] = 3.70, p = 0.05), a significant main effect of Region (F[1,13] = 10.28, p = 0.007) with a larger P250 in the left hemisphere than the right, but no significant interaction effect. Contrasts indicated that these differences were attributable to larger P250s for the House condition relative to the Human condition (F[1,13] = 4.70, p = 0.05). For P250 latency, there were no significant main effects of Stimulus Type or Region, but there was a significant interaction effect (F[2,26] = 5.58, p = 0.02). Contrasts indicated that this effect was attributable to shorter P250s in the right hemisphere between the Human and House conditions (F[2.26] = 6.28, p = 0.03).

For the late negativity, N320, AUC measures were calculated in the latency range 260–400 ms post-stimulus. Figure 6 shows what appears to be a clear difference in the group average data between the three visual conditions. The amplitude of this ERP component however, was extremely variable from individual to individual and a 2 X 3-way ANOVA with main effects of Stimulus Type (Human, Monkey, House) and Region (Left, Right) confirmed that there were no significant main effects or interactions in the N320 AUC data.

3.3 Results Summary

Audiovisual Match Conditions—The main findings were that N140 amplitude was larger for both facial image-vocalization sound pairings (Human and Monkey), relative to the object-object sound (House) pairing. No effects were observed on N170 amplitude for any of the manipulations. Additionally, in the later potentials P250 was larger for the human context pairings (human and house), relative to the animal-animal vocalization pairing (monkey). P400 was largest for the incongruous condition where the human face image was paired with the house sound.

Audiovisual Match vs Mismatch conditions—N140 amplitude was largest to pairing the Human face image and its accompanying vocalization relative to the mismatched audiovisual pairings with the human face. Additionally, N140 amplitude was largest when the House image was paired with the human vocalization. P250 amplitudes were largest for the monkey face was paired with incongruous sounds such as the human vocalization and the house sound. Finally, P400 amplitudes were largest when the Human image was paired with the house sound, and this effect was most apparent in the left hemisphere.

Visual onset ERPs—Overall there were few significant differences in ERPs between the three experimental conditions. No significant differences in ERP component behavior was observed for either N170 or N320. Some isolated amplitude and latency differences were observed for P100 and P250 latencies.

4. DISCUSSION

Characteristics of perceptually-based ERP components elicited to complex sounds and visual motion

Our experimental design contained three stimulus transitions per trial, eliciting an ERP for each stimulus event. The time between successive stimulus transitions sufficiently long to allow ERP activity associated with each transition to be easily visualized (see Fig. 1). Our focus was on the ERP elicited to the second transition – when the audiovisual stimulus combination was presented. The audiovisual pairings of natural images and sounds elicited reproducible ERPs, with similarities in morphology across the various stimulus conditions (e.g. Figs. 4–5). Despite the complexity of the audiovisual stimulus and its long stimulus duration i.e. 600 ms, we were able to record early vertex-centered ERP activity consistent with that expected to auditory stimulation i.e. N140, and it preceded putative visual activity e.g. motion-sensitive N170. Our N140s were consistent with previous studies (Woods & Elmasian, 1986) describing a vertex maximal N1 or first prominent auditory ERP (our N140). The previous work suggests that as stimulus complexity increases larger N1 amplitudes are elicited, but scalp topography and the overall morphology of the ERP waveform are preserved. Additionally, as stimulus complexity increases e.g. from pure tone, to complex tone, vowel sounds and syllables N1 latency increases its duration to around 140 ms (Woods & Elmasian, 1986). The amplitudes and latencies of the N140 observed in this study were typically larger and longer than those reported previously with speech sounds consisting of single syllables (Woods & Elmasian, 1986).

The scalp topography, latency and amplitude of the visual *motion-sensitive* N170 elicited to moving faces described here is consistent with that reported previously (Puce & Perrett, 2003; Puce et al., 2000; Puce et al., 2003). Motion-sensitive N170s are typically largest to faces when compared to moving checkerboard patterns (Puce et al., 2000), however, their latencies do not appear to be influenced by stimulus complexity *per se* (Bach & Ullrich, 1994; Kubova, Kuba, Spekreijse, & Blakemore, 1995; Puce et al., 2000). In our study, motion-sensitive N170s did not vary their amplitude or latency as a function of the complex visual stimulus type.

The scalp topography and latencies of the *category-specific* N170 were also similar to that described previously (Bentin et al., 1996). However, in the current study N170 amplitude did not follow the amplitude gradient that has previously been described in terms of being larger for (human) faces relative to other stimulus categories (Bentin et al., 1996). The significance of our findings relating to motion-sensitive and category-specific N170 are discussed in detail in the next section.

Are there species-specific neural responses in the human brain and can these differences be related to effects of animacy?

These two important questions are related in part. Our data indicate that the answer to both of these questions can be ‘yes’ in that the behavior of sensory ERPs is dependent on the context of the concurrent sensory input. For example, when examining ERPs to congruous images and sounds i.e. in the Match conditions, N140 was significantly larger to both the Human and Monkey conditions relative to the House, consistent with an animacy effect. In contrast, when the images were paired with incongruous sounds i.e. the Mismatch conditions, N140 decreased its amplitude only when the human face was paired with a non-human sound, suggestive of a species-specific effect. Hence, the presence of visual stimulation clearly influences the behavior of this auditory ERP component. It would be interesting to further explore the effects of context provided by one modality on the neural responses elicited in another stimulus modality with these types of complex stimuli. MEG and/or ERP techniques with superior temporal resolution would probably answer these questions best, relative to other methods such as fMRI, as previous invasive electrophysiological studies in monkeys have documented changes in amplitudes and latencies of unimodal ERP components when combined with stimulation in another modality (Schroeder & Foxe, 2002).

We had previously recorded motion-sensitive N170s to motion of the face, hand, leg, and amplitude and latency differences were observed as a function of which body part was moved (Wheaton et al., 2001). The shortest latency response was observed for leg motion, and this was followed by face and hand motion, respectively. Similarly, N170 amplitude changes were seen to different motion types of the same human body part i.e. larger response occurred to hand closing relative to opening, a leg stepping forward relative to stepping back, and a mouth opening relative to mouth closing (Wheaton et al., 2001). Here we recorded motion-sensitive N170s to one type of face movement i.e. mouth opening, as well as to a front-door opening in a house. We predicted that this response could potentially be larger for the human face than for the monkey face (and the house also), if indeed human motion is processed differently from that of other animals and objects. If, on the other hand, the motion-sensitive N170 is a response that is elicited by viewing biological motion in the form of moving primate faces (human or non-human), then it would be larger for both human and monkey faces, relative to houses. A final potential outcome was that viewing motion of high-level visual stimuli could conceivably elicit motion-sensitive N170s that are potentially equal in size. Data from neuroimaging studies would argue that this latter alternative was unlikely. The human superior temporal sulcus (STS) shows preferential activation to human motion relative to other high-level visual motion, whereas other motion-sensitive regions in the lateral temporal cortex will activate more consistently to other types of complex motion (e.g. Pelphrey et al., 2003; Puce et al., 1998). From the point-of-view of an ERP study, volume-conducted activity from two anatomically separate, but nevertheless nearby regions might show similar topographic scalp ERP distributions and may not be separable spatially (for example see Puce et al., 2003). However, differences in N170 amplitude could be observed between moving high-level visual stimuli relative to lower-level controls (cf. faces vs checkerboards as in Puce et al., 2003), but might not occur between such high-level visual motion stimuli such as biological motion and the motion of objects. Future combined ERP and MEG studies conceivably might be able to better address the issue of temporally coincident and spatially separate but nearby active neural sources.

We predicted that that *category-specific* N170 would be modulated by the category of visual stimulus, and that human and monkey faces would elicit a larger N170 than would houses. We based our predictions on data from previous studies using static animal faces and other objects. N170 has previously been reported to be largest when observing static human faces (e.g. Bentin et al., 1996; Eimer & McCarthy, 1999) relative to many other object categories, but is similar to that elicited to ape faces (Carmel & Bentin, 2002). Typically static animal faces (cat or dog) elicit N170s that are around 70–80% of those seen to human faces (Bentin et al., 1996). Given the previous visual category-specific literature with N170 being elicited to static images, it was somewhat surprising therefore that we observed no differences in N170 amplitude across the stimulus categories. The previous studies have documented these N170 differences in stimuli that are task-irrelevant, with subjects detecting target stimuli that do not belong to these categories. In our task each stimulus presentation was task-relevant and the initial visual onset of the stimulus signaled to the subject that the audiovisual stimulus pair upon which they were required to make a judgment was imminent. Further work exploring effects on the amplitude and latency of N170 as a function of task type and also unisensory and polysensory stimulation might be able to explain some of these observed differences.

We did observe an apparent species-specific effect in our later ERP, the P400, with the largest P400 occurring when only the human face was mismatched with inappropriate sounds. We discuss the significance of this finding in more detail in the next section.

Audiovisual mismatched stimulus pairs elicit a neural response which may reflect a detector of physical incongruity

Previous ERP studies investigating different types of stimulus mismatch have documented different types of mismatch response. The mismatch negativity (MMN) is a negative potential that occurs at around 170–200 ms post-stimulus, and is observed when subjects explicitly ignore incoming auditory stimuli (Sams, Paavilainen, Alho, & Naatanen, 1985). Our mismatch response is unlikely to be an MMN given that our subjects were actively focused on the stimuli and our ERP effects of mismatch occurred around 200 ms later than the MMN. Another previously described mismatch potential, the N400, does occur at around the latency of the effects observed in our study. However, it is typically elicited over the centroparietal scalp to unimodal semantically incongruous information, such as when subjects parse sentences with semantically incongruous endings (Kutas & Hillyard, 1980). The N400, however, has also been described in tasks when incongruous stimuli are faces (Jemel, George, Olivares, Fiori, & Renault, 1999), and musical notes (Besson & Macar, 1987). The N400 potential has been called the semantic incongruity, to reflect the fact that it is elicited mainly in situations calling for the processing of semantic information. Our P400 mismatch response, however, had an opposite polarity to N400, and may more likely correspond to a late incongruity potential not associated with semantic incongruity *per se*. The P416 was originally seen when subjects heard the last word of a grammatically correct spoken sentence being uttered by a voice which changed its *gender* (McCallum, Farmer, & Pocock, 1984). McCallum and colleagues postulated that the (auditory) P416 was elicited to the *physical incongruity* that was produced by the change in the speaker's voice. In our multisensory congruity manipulation, the mismatched audiovisual stimuli were physically incongruous. Interestingly, our largest P400 occurred when the human face image was presented without the human voice. The largest P400 was indeed elicited when the face image was paired with the house sound – corresponding to potentially the most incongruous stimulus combination.

Interestingly, the mismatch manipulation produced modulations of both perceptually-based as well as cognitive ERP activity. N140 amplitude to the human vocalization was largest when the image of the human face was present, indicating that visual context has a modulatory effect even on earlier perceptual responses. Hence, there was a bias favoring the human face and vocalization, relative to either a face of a non-human primate (monkey) or to a non-living thing (house).

The current study put into the context of neuroimaging studies of human speech and non-speech

Our ERP data indicate that neural activity relating to species-specific auditory information is already differentiable at around 140 ms post-stimulus onset. The relatively short latency of the N140 elicited to the complex auditory stimuli could allow for the rapid processing of incoming auditory information at around 7 items/second. Continuously incoming vocal information for humans in every day life usually consists of speech, typically with an average rate of 172 words/ minute during regular speech (Walker, Roberts, & Hedrick, 1988). Words consist of multiple phonemes, and for example, the average phonemes/word in Swedish discourse is around 6.78 (Spens, 1995). Pauses are also inserted in natural speech in order to clump phrases, emphasize certain words, indicate endings of sentences, and also breathe (with average pause rates in normal discourse are around 776 ms) (Merlo & Mansur, 2004). Taking these figures into account suggests that N140 and its latency can allow for the processing of incoming continuous human vocalizations as would occur in a normal conversation.

Neuroimaging studies have demonstrated that human speech and non-speech vocalizations activate specialized regions of superior temporal cortex. Human vocalizations, including speech sounds such as words, pseudowords and speech played backwards have been shown to selectively activate bilateral regions of the superior temporal sulcus (STS) (Belin et al., 2000; Binder et al., 2000). Binder and colleagues (2000) suggested that it is the acoustic characteristics of the stimulus rather than linguistic factors that generate this activation. However, this activation selectivity persists when the acoustic characteristics of the stimuli have been controlled for (Belin et al., 2002; Belin et al., 2000). There is also a suggestion that non-speech human vocalizations activate the right STS more strongly than the left STS, whereas linguistic vocalizations tend to favor the left STS, although the activation that is seen can be bilateral (Belin et al., 2002; Scott, Blank, Rosen, & Wise, 2000).

These above studies have investigated responses to auditory human vocalizations in the absence of visual input. There have been a number of studies that have investigated the activation patterns that are produced when human speech is presented in the context of a visually presented facial stimulus (Calvert, Campbell, & Brammer, 2000; Olson, Gatenby, & Gore, 2002). In these studies, the effects of synchronizing and desynchronizing the visual and auditory components of the speech stimulus, and presenting the stimuli in one modality were investigated. The STS found to be active in response to these types if stimuli irrespective of whether the visual and auditory components were synchronized, desynchronized, or if the visual stimulus (face showing speech movements of the mouth) was presented in isolation. The strongest activation was seen in the condition where audio-visual components of stimulation were matched (Calvert et al., 2000; Olson et al., 2002). The stronger response in the STS in response to the matched audio-visual speech stimulus is thought to reflect crossmodal binding functions in this region of cortex (Calvert et al., 2000). However, it is possible that these regions might also play a role in the binding of audio-visual non-speech stimuli (Bushara et al., 2003), in a study where a high-level motion stimulus was presented. For example, Beauchamp and colleagues (2004) have found that different regions within the posterior temporal cortex respond selectively to audiovisual stimulation consisting of tool sounds (Beauchamp, Lee, Argall, & Martin, 2004). Taken together, these neuroimaging studies indicate that human temporal cortex has selective mechanisms for integrating audio-visual stimuli, and this is seen to work at its best for human vocalizations that occur in the presence of the visual image of the face.

While it is difficult to localize the sources for the ERP activity seen in this study, robust activity was elicited over the temporal scalp, consistent with date from the neuroimaging studies. Unimodal auditory studies have localized the source for N100 (or N1) to the superior temporal plane bilaterally (Eggermont & Ponton, 2002; Giard et al., 1994; Godey, Schwartz, de Graaf, Chauvel, & Liegeois-Chauvel, 2001). In one early experiment using audiovisual stimulation (Sams et al., 1991), MEG activity was sampled from the left hemisphere in response to viewing lips mouth a syllable, in addition to hearing an uttered syllable. This early investigation, devoted to the McGurk effect, indicated that neural activity seen at around 180 ms post-stimulus onset appeared to be crucial for blending the audiovisual stimulus (Sams et al., 1991). While there are neuroimaging data (described above) to these complex stimulus displays, ERP/MEG studies typically have not studied this issue. Part of the reason for this is that ERP/MEG component morphology can be complex in audiovisual tasks. Additional activity can be observed to multisensory stimulus presentation, relative to when neural activity is examined to stimulation in either sensory modality in isolation (see (Giard & Peronnet, 1999). In a study examining audiovisual integration effects using a basic visual shape (e.g. circle) and an auditory tone (e.g. a tone burst), and a combined audio-visual condition where the circle was distorted to an ellipse when accompanied by tone burst, a set of ERP data was recorded in response to visual stimulation alone, auditory stimulation alone, and combined auditory-visual stimulation (Giard & Peronnet, 1999). Their data indicate that audio-visual interaction effects commonly occur in ERP components elicited below 200 ms post-stimulus – which they showed by comparing summated unisensory ERP waveforms with their audiovisual ERP waveform. Furthermore, they also identified later cognitive ERP activity whose latency shortened as a function of multimodal stimulation. Bearing in mind Giard and Peronnet's pioneering audiovisual ERP study in 1999, we constructed the current study so that the experimental conditions of interest would be under combined audiovisual stimulation.

5. CONCLUSIONS

We used a mismatch manipulation, shown in previous ERP literature to be a powerful approach to study perceptual or cognitive systems. Our electrophysiological data add evidence to existing neuroimaging data showing that the human brain has specialized neural mechanisms for processing the actions and vocalizations of other humans, and that species-specific effects can occur at both early more perceptually-based (N140) and late more cognitively-based (P400) processing stages following stimulus delivery. In addition, our data show that the neural 'double-take' or incongruity effect occurs on a rather slow time scale, relative to the rate at which incoming utterances would be processed. This might explain why it takes so long for 'the penny to drop' when we are confronted with unexpected and conflicting visual and auditory input from our conspecifics.

Acknowledgments

We thank Jillian Hardee and Leor Zellner for assistance with data analysis, and Mary Pettit for editorial assistance. Nicki Hertel calculated the harmonic-to-noise ratios for our sound stimuli made available to us in the laboratory of Dr Jim Lewis.

References

Allison T, Ginter H, McCarthy G, Nobre AC, Puce A, Luby M, et al. Face recognition in human extrastriate cortex. J Neurophysiol 1994;71(2):821–825. [PubMed: 8176446]

Bach M, Ullrich D. Motion adaptation governs the shape of motion-evoked cortical potentials. Vision Res 1994;34(12):1541–1547. [PubMed: 7941362]

Beauchamp MS, Argall BD, Bodurka J, Duyn JH, Martin A. Unraveling multisensory integration: patchy organization within human STS multisensory cortex. Nat Neurosci 2004;7(11):1190–1192. [PubMed: 15475952]

Beauchamp MS, Lee KE, Argall BD, Martin A. Integration of auditory and visual information about objects in superior temporal sulcus. Neuron 2004;41(5):809–823. [PubMed: 15003179]

Beauchamp MS, Lee KE, Haxby JV, Martin A. Parallel visual motion processing streams for manipulable objects and human movements. Neuron 2002;34(1):149–159. [PubMed: 11931749]

Belin P, Zatorre RJ, Ahad P. Human temporal-lobe response to vocal sounds. Brain Res Cogn Brain Res 2002;13(1):17–26. [PubMed: 11867247]

Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. Voice-selective areas in human auditory cortex. Nature 2000;403(6767):309–312. [PubMed: 10659849]

Bentin S, Allison T, Puce A, Perez A, McCarthy G. Electrophysiological studies of face perception in humans. J Cogn Neurosci 1996;8:551–565.

Besson M, Macar F. An event-related potential analysis of incongruity in music and other non-linguistic contexts. Psychophysiology 1987;24(1):14–25. [PubMed: 3575590]

Binder JR, Frost JA, Hammeke TA, Bellgowan PS, Springer JA, Kaufman JN, et al. Human temporal lobe activation by speech and nonspeech sounds. Cereb Cortex 2000;10(5):512–528. [PubMed: 10847601]

Bonda E, Petrides M, Ostry D, Evans A. Specific involvement of human parietal systems and the amygdala in the perception of biological motion. J Neurosci 1996;16(11):3737–3744. [PubMed: 8642416]

Bushara KO, Hanakawa T, Immisch I, Toma K, Kansaku K, Hallett M. Neural correlates of cross-modal binding. Nat Neurosci 2003;6(2):190–195. [PubMed: 12496761]

Calvert GA, Campbell R, Brammer MJ. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Curr Biol 2000;10 (11):649–657. [PubMed: 10837246]

Carmel D, Bentin S. Domain specificity versus expertise: factors influencing distinct processing of faces. Cognition 2002;83(1):1–29. [PubMed: 11814484]

Chao LL, Haxby JV, Martin A. Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. Nat Neurosci 1999;2(10):913–919. [PubMed: 10491613]

Chao LL, Martin A. Representation of manipulable man-made objects in the dorsal stream. Neuroimage 2000;12(4):478–484. [PubMed: 10988041]

Darwin, C. The Expression of the Emotions in Man and Animals. Vol. 3. Oxford: Oxford University Press; 18991998.

Eggermont JJ, Ponton CW. The neurophysiology of auditory perception: from single units to evoked potentials. Audiol Neurootol 2002;7(2):71–99. [PubMed: 12006736]

Eimer M, McCarthy RA. Prosopagnosia and structural encoding of faces: evidence from event-related potentials. Neuroreport 1999;10(2):255–259. [PubMed: 10203318]

Fang F, He S. Cortical responses to invisible objects in the human dorsal and ventral pathways. Nat Neurosci 2005;8(10):1380–1385. [PubMed: 16136038]

Fecteau S, Armony JL, Joanette Y, Belin P. Is voice processing species-specific in human auditory cortex? An fMRI study. Neuroimage 2004;23(3):840–848. [PubMed: 15528084]

Ghazanfar AA, Maier JX, Hoffman KL, Logothetis NK. Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. J Neurosci 2005;25 (20):5004–5012. [PubMed: 15901781]

Giard MH, Peronnet F. Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. J Cogn Neurosci 1999;11 (5):473–490. [PubMed: 10511637]

Giard MH, Perrin F, Echallier JF, Thevenet M, Froment JC, Pernier J. Dissociation of temporal and frontal components in the human auditory N1 wave: a scalp current density and dipole model analysis. Electroencephalogr Clin Neurophysiol 1994;92(3):238–252. [PubMed: 7514993]

Gil-da-Costa R, Braun A, Lopes M, Hauser MD, Carson RE, Herscovitch P, et al. Toward an evolutionary perspective on conceptual representation: species-specific calls activate visual and affective processing systems in the macaque. Proc Natl Acad Sci U S A 2004;101(50):17516–17521. [PubMed: 15583132]

Godey B, Schwartz D, de Graaf JB, Chauvel P, Liegeois-Chauvel C. Neuromagnetic source localization of auditory evoked fields and intracerebral evoked potentials: a comparison of data in the same patients. Clin Neurophysiol 2001;112(10):1850–1859. [PubMed: 11595143]

Grossman E, Donnelly M, Price R, Pickens D, Morgan V, Neighbor G, et al. Brain areas involved in perception of biological motion. J Cogn Neurosci 2000;12(5):711–720. [PubMed: 11054914]

Jemel B, George N, Olivares E, Fiori N, Renault B. Event-related potentials to structural familiar face incongruity processing. Psychophysiology 1999;36(4):437–452. [PubMed: 10432793]

Kubova Z, Kuba M, Spekreijse H, Blakemore C. Contrast dependence of motion-onset and pattern-reversal evoked potentials. Vision Res 1995;35(2):197–205. [PubMed: 7839616]

Kutas M, Hillyard SA. Reading senseless sentences: brain potentials reflect semantic incongruity. Science 1980;207(4427):203–205. [PubMed: 7350657]

Lewis JW, Brefczynski JA, Phinney RE, Janik JJ, DeYoe EA. Distinct cortical pathways for processing tool versus animal sounds. J Neurosci 2005;25(21):5148–5158. [PubMed: 15917455]

McCallum WC, Farmer SF, Pocock PV. The effects of physical and semantic incongruities on auditory event-related potentials. Electroencephalogr Clin Neurophysiol 1984;59 (6):477–488. [PubMed: 6209114]

Merlo S, Mansur LL. Descriptive discourse: topic familiarity and disfluencies. J Commun Disord 2004;37 (6):489–503. [PubMed: 15450437]

Nunez PL. A method to estimate local skull resistance in living subjects. IEEE Trans Biomed Eng 1987;34 (11):902–904. [PubMed: 3692509]

Nunez PL, Silberstein RB, Cadusch PJ, Wijesinghe RS, Westdorp AF, Srinivasan R. A theoretical and experimental study of high resolution EEG based on surface Laplacians and cortical imaging. Electroencephalogr Clin Neurophysiol 1994;90(1):40–57. [PubMed: 7509273]

Olson IR, Gatenby JC, Gore JC. A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. Brain Res Cogn Brain Res 2002;14 (1):129–138. [PubMed: 12063136]

Partan S, Marler P. Communication goes multimodal. Science 1999;283(5406):1272–1273. [PubMed: 10084931]

Pelphrey KA, Mitchell TV, McKeown MJ, Goldstein J, Allison T, McCarthy G. Brain activity evoked by the perception of human walking: controlling for meaningful coherent motion. J Neurosci 2003;23 (17):6819–6825. [PubMed: 12890776]

Perrin F, Pernier J, Bertrand O, Echallier JF. Spherical splines for scalp potential and current density mapping. Electroencephalogr Clin Neurophysiol 1989;72(2):184–187. [PubMed: 2464490]

Puce A, Allison T, Bentin S, Gore JC, McCarthy G. Temporal cortex activation in humans viewing eye and mouth movements. J Neurosci 1998;18(6):2188–2199. [PubMed: 9482803]

Puce A, Perrett D. Electrophysiology and brain imaging of biological motion. Philos Trans R Soc Lond B Biol Sci 2003;358(1431):435–445. [PubMed: 12689371]

Puce A, Smith A, Allison T. ERPs evoked by viewing facial movements. Cog Neuropsychol 2000;17:221–239.

Puce A, Syngeniotis A, Thompson JC, Abbott DF, Wheaton KJ, Castiello U. The human temporal lobe integrates facial form and motion: evidence from fMRI and ERP studies. Neuroimage 2003;19(3): 861–869. [PubMed: 12880814]

Riede T, Herzel H, Hammerschmidt K, Brunnberg L, Tembrock G. The harmonic-to-noise ratio applied to dog barks. J Acoust Soc Am 2001;110(4):2191–2197. [PubMed: 11681395]

Sams M, Aulanko R, Hamalainen M, Hari R, Lounasmaa OV, Lu ST, et al. Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. Neurosci Lett 1991;127(1):141–145. [PubMed: 1881611]

Sams M, Paavilainen P, Alho K, Naatanen R. Auditory frequency discrimination and event-related potentials. Electroencephalogr Clin Neurophysiol 1985;62(6):437–448. [PubMed: 2415340]

Schroeder CE, Foxe JJ. The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. Brain Res Cogn Brain Res 2002;14(1):187–198. [PubMed: 12063142]

Scott SK, Blank CC, Rosen S, Wise RJ. Identification of a pathway for intelligible speech in the left temporal lobe. Brain 2000;123(Pt 12):2400–2406. [PubMed: 11099443]

Shettleworth SJ. Animal cognition and animal behaviour. Animal Behaviour 2001;61:277–286.

Spens, K-E. Evaluation of speech tracking results: Some numerical considerations and Examples. In: Plant, G.; Spens, K-E., editors. Profound Deafness and Speech Communication. San Diego, Calif: Singular Pub. Group; 1995. p. 417-437.

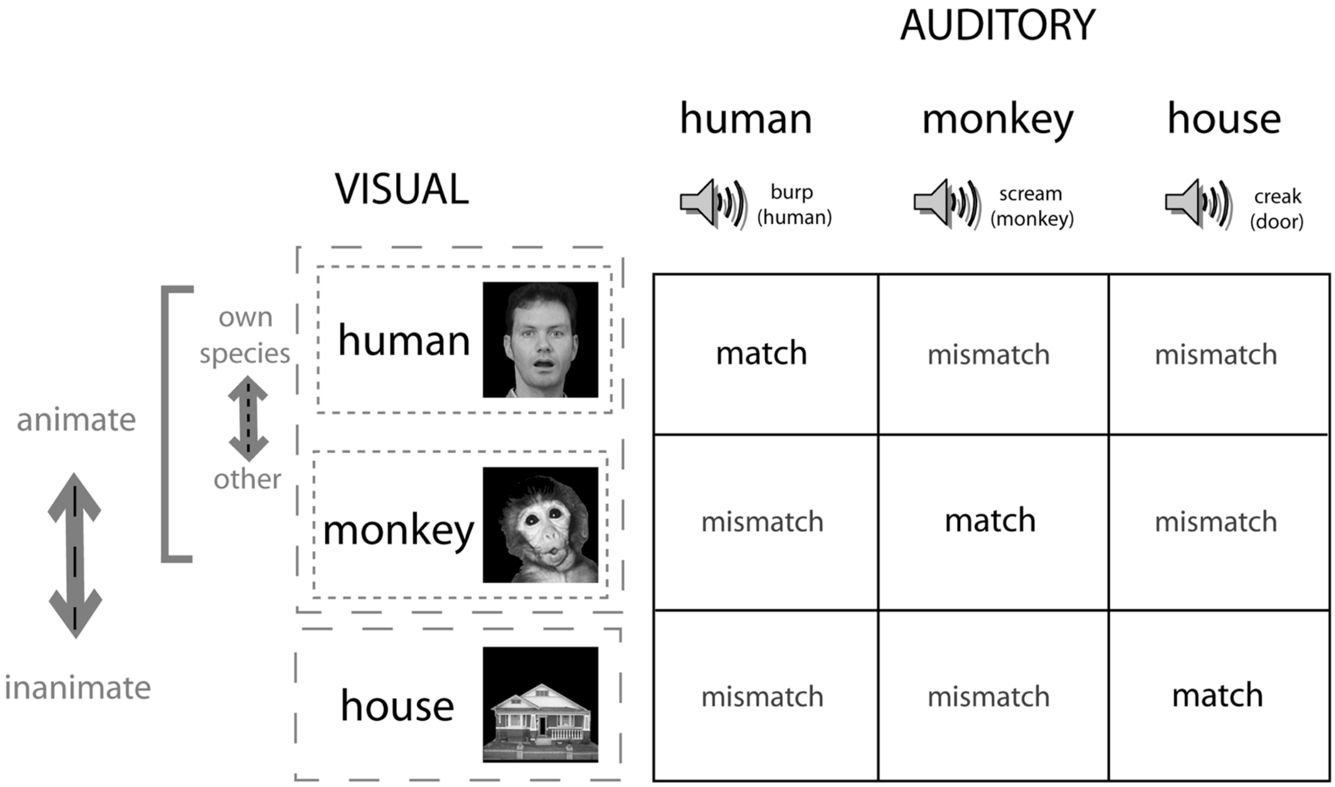
Tomasello, M.; Call, J. Primate Cognition. New York: Oxford University Press; 1997.

Walker VG, Roberts PM, Hedrick DL. Linguistic analyses of the discourse narratives of young and aged women. Folia Phoniatr (Basel) 1988;40(2):58–64. [PubMed: 3169657]

Wheaton K, Pipingas A, Silberstein R, Puce A. Neuronal responses elicited to viewing the actions of others. Vis Neurosci 2001;18:401–406. [PubMed: 11497416]

Wheaton KJ, Pipingas A, Silberstein RB, Puce A. Human neural responses elicited to observing the actions of others. Vis Neurosci 2001;18(3):401–406. [PubMed: 11497416]

Woods DL, Elmasian R. The habituation of event-related potentials to speech sounds and tones. Electroencephalogr Clin Neurophysiol 1986;65(6):447–459. [PubMed: 2429824]



	human	monkey	house
human	match	mismatch	mismatch
monkey	mismatch	match	mismatch
house	mismatch	mismatch	match

Figure 1.

Experimental design. The possible stimulus dimensions for audiovisual stimulation are presented in a 3 × 3 matrix. The three types of visual stimulus can be differentiated across animate/inanimate domains (human face, monkey face vs house), or human/non-human domains (human face vs monkey face). A similar parcellation also applies to the stimuli in the auditory modality with animate/inanimate domains (human burp, monkey scream vs creaking door), or human/non-human domains (human burp vs monkey scream). There are 9 potential combinations of audiovisual stimulus, which constitute matches or mismatches across sensory modality. The 3 ‘matched’ audiovisual stimulus pairs are a human face/human sound, monkey face/monkey sound, house image/house sound. Similarly, there are 6 possible mismatched audiovisual stimulus pairs.

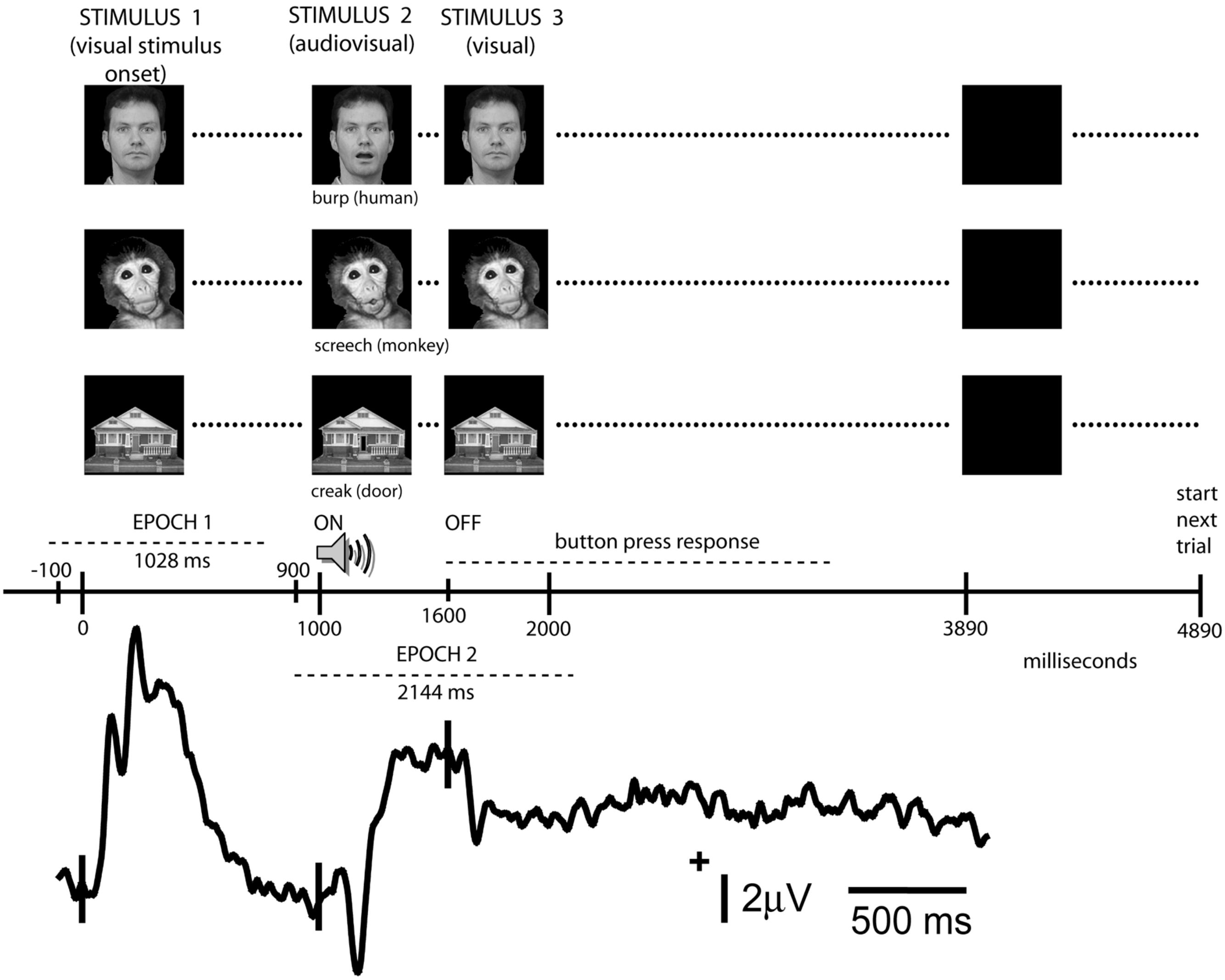


Figure 2.

Task timing and stimulus examples. The experimental trial begins with the presentation of one of the three types of image (human face, monkey face, house) which remains on for a period of 1 sec during which time a visual onset ERP is generated (Stimulus 1). After this time the image is replaced by another image which sets up an apparent motion condition –this change in the visual image is accompanied by the presentation of a sound (Stimulus 2). The duration of the image/sound combination is 600 ms, following which the original image is once again presented for a period of 2290 ms (Stimulus 3). A black screen, shown for 1 sec, separates one trial from another. Subjects are required to response with a button press following the occurrence of Stimulus 2. ERP data were analyzed in two separate epochs designed to highlight visual stimulus onset (Epoch 1) and the presentation of the audiovisual stimulus pair (Epoch 2) as shown on the experimental time line.

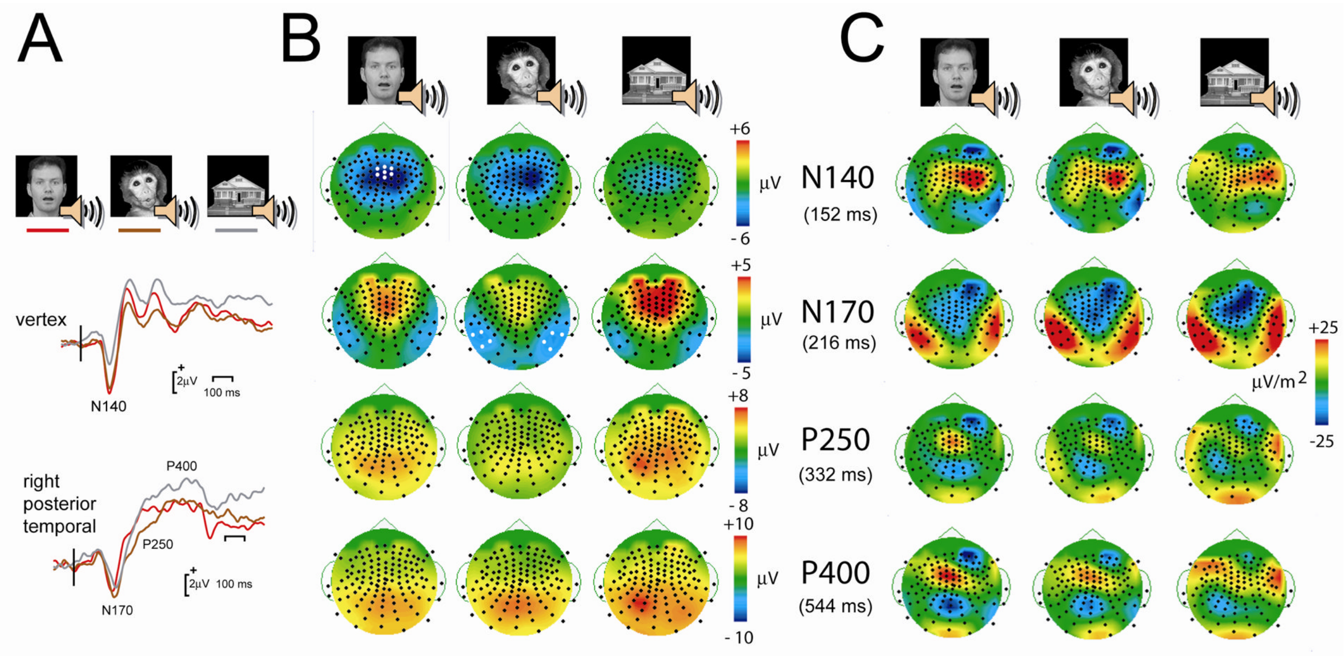


Figure 3.

Audiovisual Match conditions: Examples of ERP waveforms, topographic voltage maps and Laplacians for N140, N170, P240 and P400. **A.** ERP waveforms from the vertex, highlighting N140, and right posterior temporal scalp, displaying N170, P250 and P400, are shown. Vertical bar shows the onset of Stimulus 2 i.e. an apparent visual motion and associated sound stimulus. **B.** Topographic voltage maps show the scalp distribution of each ERP component at its peak latency. N140 has a focal centrofrontal distribution, N170 a bilateral posterior temporal distribution, whereas the remaining components are seen mainly in the posterior scalp. Electrode positions are shown as black circles, and electrodes chosen for statistical analyses are shown as white circles. L = left, and R = right. **C.** Laplacian maps for each ERP component at its peak latency show sources (warm colors) and sinks (cool colors) of neural activity.

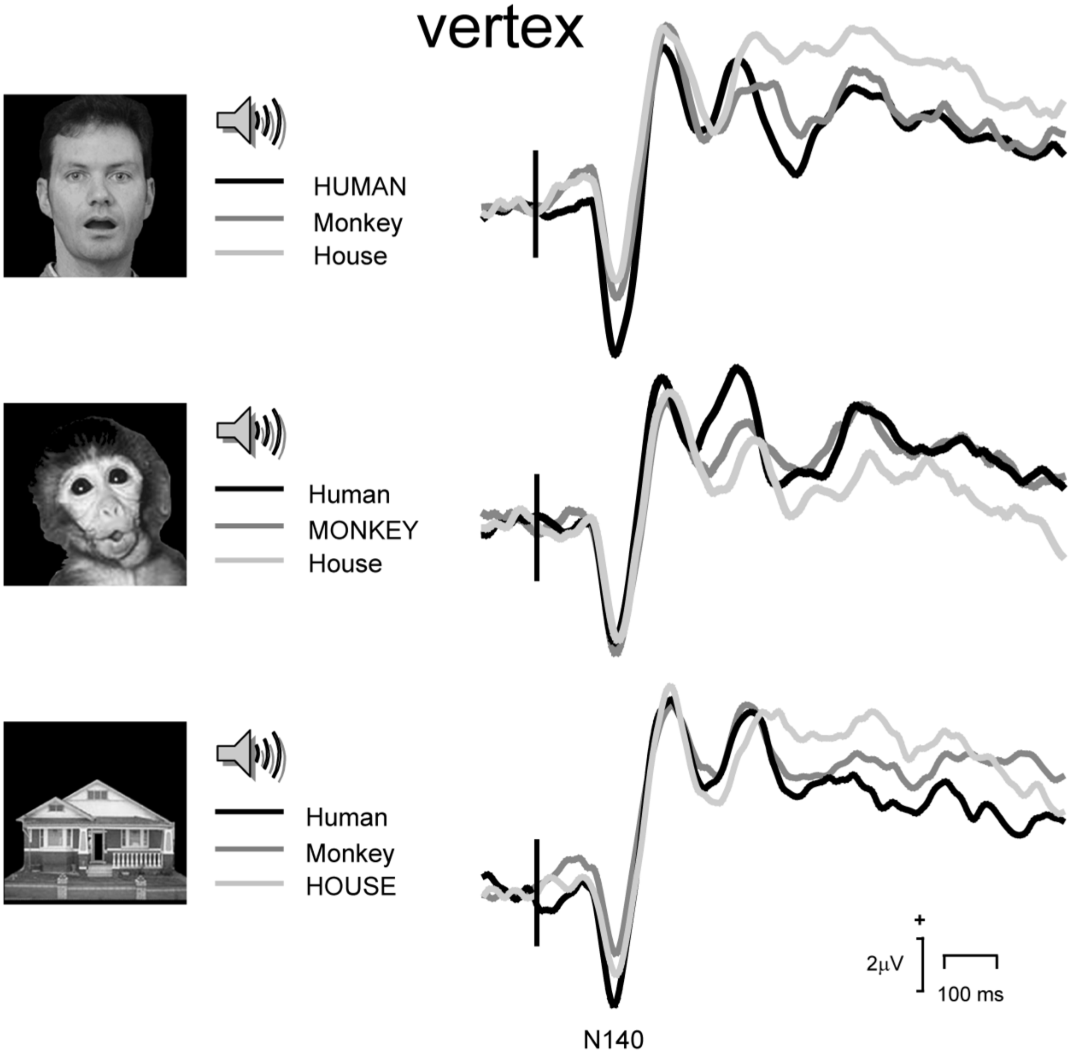


Figure 4.
Audiovisual Match vs Mismatch conditions: ERP waveforms obtained from the vertex region showing N140. Vertical bar depicts audiovisual stimulus onset. Legend shows each respective Match condition in upper case.

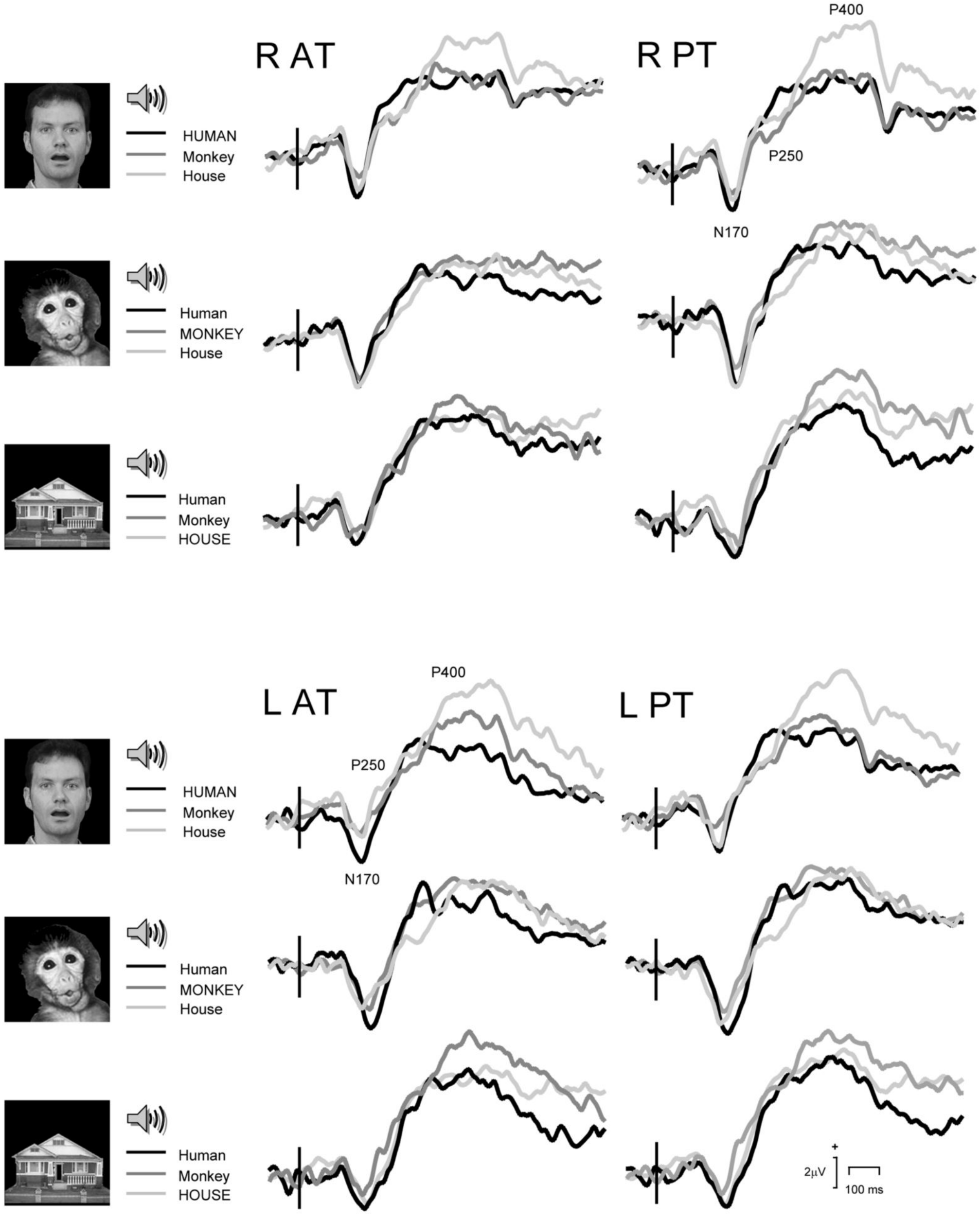


Figure 5.
Audiovisual Match vs Mismatch conditions: ERP waveforms showing N170, P250 and P400. **Top panel.** Right (R) hemisphere anterior (AT) and posterior (PT) temporal scalp ERPs. **Bottom panel.** Left (L) hemisphere anterior (AT) and posterior (PT) temporal scalp ERPs. Vertical bar depicts audiovisual stimulus onset. Legend shows each respective Match condition in upper case.

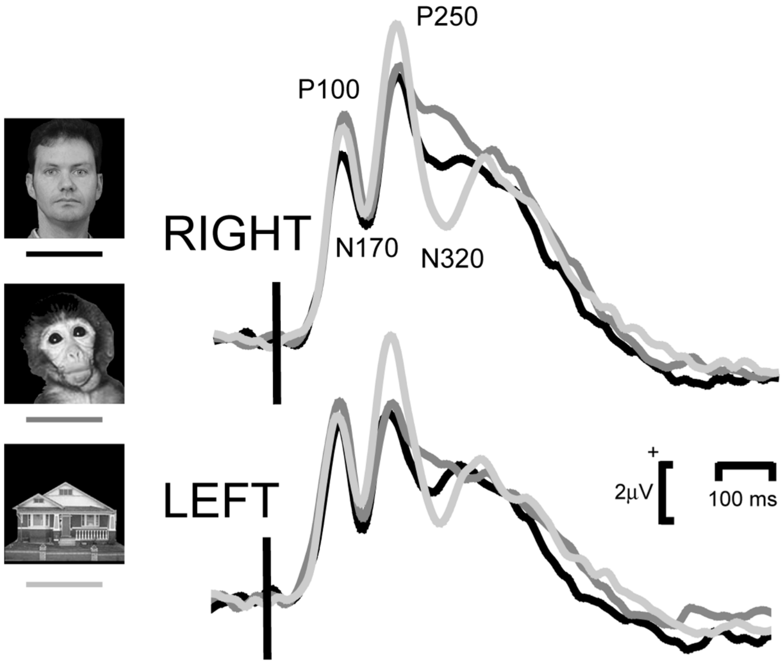


Figure 6.
Visual onset ERPs elicited to visual stimulation from the left and right posterior temporal scalp. Vertical bar depicts visual stimulus onset.

Table 1

Audiovisual Mismatch conditions

Posterior temporal scalp ERP latency, amplitude and area under the curve (AUC) measures were analyzed using 2-way repeated measures ANOVAs with main effects of Stimulus Type (Human, Monkey, House) and Region (Left, Right).

					Stim Type × Region	
	Stim Type F[2,26]	P	Region F[1,13]	P	F[2,26]	P
N170 lat						
Human	1.52	ns	0.21	ns	0.03	ns
Monkey	2.08	ns	0.00	ns	0.06	ns
House	2.01	ns	0.32	ns	1.29	ns
N170 ampl						
Human	1.47	ns	0.90	ns	2.52	0.10
Monkey	2.56	0.10	1.54	ns	1.22	ns
House	0.65	ns	0.16	ns	0.55	ns
P250 AUC						
Human	1.40	ns	0.02	ns	0.24	ns
Monkey	**3.70**	**0.04**	0.04	ns	0.28	ns
House	1.08	ns	0.02	ns	1.19	ns
P400 AUC						
Human	3.11	0.09	3.12	0.10	**6.35**	**0.006**
Monkey	0.57	ns	0.27	ns	0.21	ns
House	2.61	0.10	0.23	ns	0.37	ns

Table 2

Audiovisual Mismatch conditions

Midline P400 area under the curve (AUC) measures were analyzed using 1-way repeated measures ANOVAs with a main effect of Stimulus Type (Human, Monkey, House). Lateral frontocentral N140 ERP measures were analyzed using 2-way repeated measures ANOVAs with main effects of Stimulus Type and Region (Left, Midline, Right).

	Stim Type		**Region**		**Stim Type × Region**	
Midline	**F[2,26]**	**P**				
P400 AUC						
Human	3.37	0.05				
Monkey	1.34	ns				
House	1.23	ns				
Frontocentral	**F[2,26]**	**P**	**F[1,13]**	**P**	**F[2,26]**	**P**
N140 lat						
Human	1.52	ns	0.21	ns	0.03	ns
Monkey	2.08	ns	0.00	ns	0.06	ns
House	2.01	ns	0.32	ns	1.29	ns
N140 ampl						
Human	1.47	ns	0.90	ns	2.52	0.10
Monkey	2.56	0.10	1.54	ns	1.22	ns
House	0.65	ns	0.16	ns	0.55	ns