Human brain responses to non-verbal audiovisual dynamic stimuli

Aina Puce

Department of Psychological and Brain Sciences
Indiana University,
Bloomington IN, USA
Email: ainapuce@indiana.edu

ABSTRACT

When interacting with others we focus on the spoken word, but also 'read' non-verbal cues from the face and voice. Our recordings of the electrical activity of human brain (event-related potentials, or ERPs) indicate that these audiovisual cues are integrated as early as 140msec post-stimulus (sensory ERP components: auditory N140 and visual N170). When multisensory inputs were congruent, and potentially redundant, N170 lost its usual category-sensitivity, whereas auditory N140 showed selectivity to primate and human vocalizations. A late ERP, P400, was significantly larger when a human face was paired with an incongruent sound. This cross-modal incongruity is similar to a previously described auditory (physical) incongruity potential (McCallum et al 1984). When audiovisual and unisensory stimulation were contrasted, auditory N140 and visual N170 exhibited underadditivity, whereas later ERPs showed more complex effects. Our data indicate that the human brain possesses specialized circuitry for rapidly processing information from the face and voice.

'When you speak to a man, look on his eyes; when he speaks to you, look on his mouth.'
Benjamin Franklin, Poor Richard's Almanack (1732-1757).

The human face attracts our attention, as the eyes signal where the focus of the attention is (Kleinke, 1986; Puce and Perrett, 2003; Itier and Batty, 2009) and the mouth is the source of human verbal communication – as suggested by Franklin's quotation. Indeed, when the eyes of an observer are tracked when they gaze on the face of another, the eyes and mouth of the viewed face gets the lion's share of the observer's gaze – there is a triangle of visual focus where the apex is the mouth and the base consists of the eyes (Yarbus, 1967). Processing what the eyes do is rapid and unisensory – changes in the eyes are seen and not heard. The human brain responds reliably to changes in the gaze of other individuals – with changes in the electrical activity of brain occurring around 1/5 of a second after the observed gaze change (Puce et al., 2000; Puce et al., 2003). These original studies used unisensory stimuli, and an identical neural response was observed to an opening and closing mouth (Puce et al., 2000). Yet, in everyday life when we interact with others, mouth movements are usually associated with speech. Sometimes this oral output can be non-verbal e.g. a sneeze, cough, yawn, or burp, but in all cases the actions of the face and voice form a congruent multisensory stimulus. These non-verbal forms of multisensory behavior are not unique to human primates. Non-human primates and other animals utilize these multisensory cues from their conspecifics (e.g. alarm calls) and other animals for survival (Seyfarth and Cheney, 2003; Templeton and Greene, 2007; Papworth et al., 2008; Stephan and Zuberbuhler, 2008).

The human neuroimaging and neurophysiological literature has focused on language and verbal communication in studies which have typically relied on unisensory stimuli

(auditory or visual) (Hagoort, 2003; Friederici, 2004; Van Petten and Luka, 2006; Chakraborty and McEvoy, 2008; Hagoort, 2008). The human voice has been called an 'auditory face' (Belin et al., 2004) and not surprisingly the human brain has developed neural systems for decoding this important stimulus category (Belin et al., 2000; Belin et al., 2002), which may be similar in monkeys (Petkov et al., 2008). Human and non-human primates show selective brain activation for the vocalizations of their own conspecifics, with different brain regions becoming active when vocalizations of other animals and one's conspecifics are contrasted within the same experiment (Fecteau et al., 2004; Petkov et al., 2008). These brain responses are separable from those to other types of complex sounds such as tools or environmental sounds (Engel et al., 2009).

Non-verbal cues in humans are critical to human communication (Hari and Kujala, 2009), yet tend to be underemphasized due to the emphasis that we place on the spoken word. It has been argued that very little of the intended message actually comes from the spoken word - movements of the face, hands and body, and pauses and inflexions in the voice form valuable and interpretable sources of information (Mehrabian and Ferris, 1967). These displays are multisensory, yet until only recently neuroimaging investigations have begun study what the neural correlates for multisensory stimuli, and how they might be differentially affected in neuropsychiatric disorders (Doehrmann and Naumer, 2008; Zupan et al., 2009).

In our laboratory we recently begun investigating the neural correlates of evaluating the dynamic face and voice (Puce et al., 2007; Brefczynski-Lewis et al., 2009), as a logical extension of our visually based studies of face, hand and body actions (Puce et al., 2000; Wheaton et al., 2001; Puce et al., 2003; Thompson et al., 2004; Wheaton et al., 2004; Thompson et al., 2005; Carrick et al., 2007; Thompson et al., 2007). Here I describe these investigations and integrate their collective findings. One experiment investigated the potential specificity for human faces and voices and the influence that congruity plays on audiovisual integration (Puce et al., 2007). The second experiment studied neurophysiological responses to the presentation of audiovisual, visual only and auditory only stimuli which consisted of a dynamic face and non-verbal vocalizations (Brefczynski-Lewis et al., 2009).

Incongruity as a means to highlight potential species specificity for face and voice

We presented healthy subjects with 3 different visual animations and 3 auditory sounds. Audiovisual stimulus pairings could be congruous or incongruous and subjects made a forced-choice button press to indicate whether each stimulus pair was congruous or incongruous. We used a human face, a monkey face and an image of a house for visual stimuli. The mouth on either face, or the front door of the house, opened concurrent with a presented sound, which could be a (human) burp, a monkey screech or a creaking door sound. Each trial consisted of an initial visual stimulus onset (1000ms duration) followed by a change in the visual display (mouth or door opening) and the presentation of the sound (400ms duration). The trials ended with sound offset and the visual stimulus reverting to its initial position (closed mouth or door). Stimulus pairings were randomly mixed, so that on any given trial subjects did not know if a congruous or incongruous stimulus would be presented. There were 3 congruous stimulus pairings (human mouth opening/human burp, monkey mouth opening/monkey screech, house front door opening/door squeaking) and 6 incongruous pairings (e.g. front door opening/monkey screech, human mouth opening/door squeaking). We measured electrical activity of the human brain non-invasively from a 128 channel scalp array. We averaged responses to the various congruous and incongruous types and generated event-related potentials (ERP) in the group of subjects.

Congruous conditions produced ERPs consisting of a prominent central scalp auditory potential – the N140 (Giard et al., 1994; Giard and Peronnet, 1999; Eggermont and Ponton,

2002), and a bilateral posterior temporal scalp visual potential – the N170 (Bentin et al., 1996; Puce and Perrett, 2003). Auditory N140 was significantly larger to the human and monkey vocalizations relative to the squeaking door sound, despite all sounds having comparable harmonic-to-noise ratios. N170 did not change with stimulus category in this experiment.

ERP component amplitudes and latencies to incongruous stimulus conditions were compared to the congruous conditions and several interesting findings were noted. First, N140 was significantly attenuated when the human face was paired with the two non-human sounds. This effect was not seen for the monkey face or house images, suggesting that human face and vocalization pairings enjoy augmented processing in the human brain. Second, 'incongruity' ERPs occurred at around 400ms post-audiovisual stimulus onset, manifesting as a broad midline parietal positivity (P400). P400 was significantly enhanced only when the human face was paired with the most incongruous auditory stimulus – the inanimate object sound. This multisensory incongruity ERP is reminiscent of the so-called 'physical incongruity' reported by McCallum and colleagues in 1984. The incongruity was generated to a mismatch in the gender of a speaker's voice (McCallum et al., 1984).

Audiovisual integration of human non-verbal vocalizations and associated facial movements

We recorded ERPs and changes in brain's blood flow (using functional magnetic resonance imaging, fMRI) in healthy human volunteers in response to viewing a synthetic human face and associated real human vocalizations. I will not report on the fMRI data in detail here, suffice it to say that the studies produced complex, but convergent, data sets.

Stimuli were presented to subjects in blocks consisting of auditory presentation only, visual presentation only, and combined audiovisual presentation. Stimuli were commonly occurring non-verbal facial movements associated with non-verbal vocalizations such as a sneeze, cough, burp, yawn, sigh, and singing a note. We asked subjects to detect two randomly presented unisensory targets – one was auditory and the other was visual, in an attempt to avoid subject bias towards one sensory modality. The two targets consisted of an eyeblink (visual target) and the vocalization 'mmmm' where the face remained motionless (auditory target). Subjects pressed a response button on target detection. We recorded 128 channel ERPs, as in the previous experiment.

The main finding in the ERP data set was a significant decrease in both auditory N140 and visual N170 amplitudes in the audiovisual condition relative to the unisensory conditions. These data indicate that sensory processing related to stimulus categories may be scaled back in response to the presence of incoming sensory information that is congruous and potentially redundant. This was further reinforced by earlier N170 latencies that were observed in the multisensory condition. Later ERPs occurring subsequent to N140 and N170 showed either maximal, but not superadditive, amplitudes in the audiovisual stimulation condition, or alternatively had amplitudes equal to their unisensory counterparts. Somewhat surprisingly and unexpectedly, early ERP activity at around 80ms post-stimulus onset showed superadditive effects, indicating that congruous multisensory stimulation conveys advantages in sensory processing that occur rapidly.

The fMRI data, in a separate group of subjects, showed underadditivity in the multisensory condition in occipitotemporal cortex, in the middle temporal gyrus and in the lateral occipital/fusiform cortices. These sites are potentially consistent with generators for the sensory ERPs (auditory N140 and visual N170) described earlier. Blood flow responses that were equal in multisensory and unisensory conditions, and maximal, but not superadditive, responses were seen in the temporo-parietal junction, superior temporal sulcus and inferior temporal gyrus. These areas have been associated with visual or auditory actions of a non-

social (Binkofski and Buccino, 2006; Lewis, 2006) or social nature (Haxby et al., 2002; Decety and Lamm, 2007; Hari and Kujala, 2009; Van Overwalle, 2009).

Synthesis and conclusions

Our electrophysiological data indicate that audiovisual information from a dynamic face and voice is integrated quickly, as shown by changes in the behavior of sensory ERP components such as the auditory N140 and visual N170. Indeed, changes in electrical activity of brain occurring as early as 80 ms post-stimulus onset were observed in this study. Figure 1 (below) shows a schematic summary of our findings across two different experiments. I have differentiated sensory from cognitive/affective ERP components using a temporal demarcation point at 250 ms post-stimulus onset.

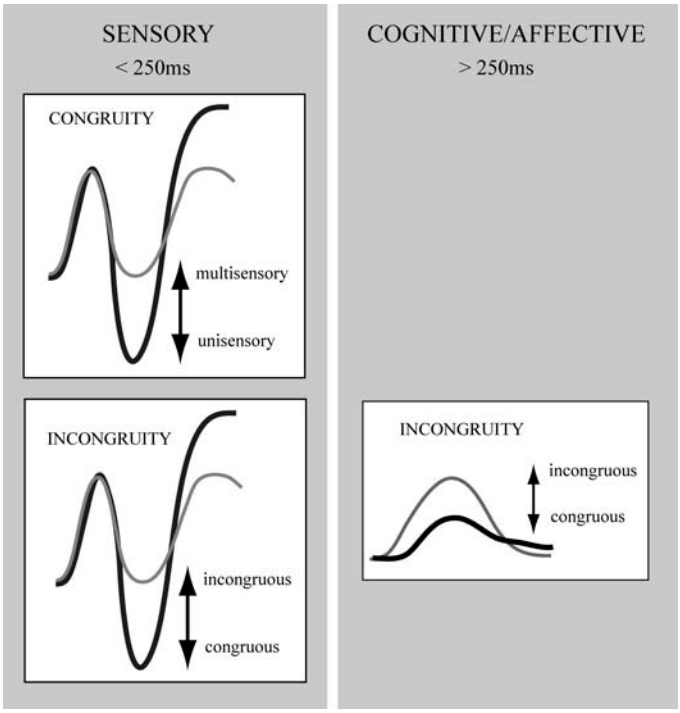

Figure 1: *Synthesis of ERP data recorded across two multisensory experiments.*

For congruous multisensory stimulation, smaller sensory ERP components were observed relative to their unisensory counterparts (see Fig. 1, left panel). Presumably, given that there is a degree of redundancy in the presented information the brain does not have to work as hard to make sense of the incoming stimulus.

However, incongruous stimulation *involving only human faces* produced significant differences in both sensory and cognitive/affective ERP components (Fig. 1 bottom part of left and right panels). Notably these effects were not observed when non-human face or non-face stimuli were presented.

In sum, recordings of the electrical activity can clarify the time-scales over which multisensory input is integrated and identify which stimulus categories might enjoy privileged processing by the human brain. Multisensory processing of human face and voice information appears to have its own specialized neural mechanisms based not only on the data presented here, but also on the neuroimaging studies reviewed earlier in this Chapter. These specialized neural mechanisms are key to our ability to rapidly interact with our conspecifics in everyday professional and personal interactions.

REFERENCES

Belin P, Zatorre RJ, Ahad P (2002) Human temporal-lobe response to vocal sounds. Brain Res Cogn Brain Res 13:17-26.

Belin P, Fecteau S, Bedard C (2004) Thinking the voice: neural correlates of voice perception. Trends Cogn Sci 8:129-135.

Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B (2000) Voice-selective areas in human auditory cortex. Nature 403:309-312.

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

Binkofski F, Buccino G (2006) The role of ventral premotor cortex in action execution and action understanding. J Physiol Paris 99:396-405.
Brefczynski-Lewis J, Lowitszch S, Parsons M, Lemieux S, Puce A (2009) Audiovisual non-verbal dynamic faces elicit converging fMRI and ERP responses. Brain Topogr 21:193-206.
Carrick OK, Thompson JC, Epling JA, Puce A (2007) It's all in the eyes: neural responses to socially significant gaze shifts. Neuroreport 18:763-766.
Chakraborty A, McEvoy AW (2008) Presurgical functional mapping with functional MRI. Curr Opin Neurol 21:446-451.
Decety J, Lamm C (2007) The role of the right temporoparietal junction in social interaction: how low-level computational processes contribute to meta-cognition. Neuroscientist 13:580-593.
Doehrmann O, Naumer MJ (2008) Semantics and the multisensory brain: how meaning modulates processes of audio-visual integration. Brain Res 1242:136-150.
Eggermont JJ, Ponton CW (2002) The neurophysiology of auditory perception: from single units to evoked potentials. Audiol Neurootol 7:71-99.
Engel LR, Frum C, Puce A, Walker NA, Lewis JW (2009) Different categories of living and non-living sound-sources activate distinct cortical networks. Neuroimage 47:1778-1791.
Fecteau S, Armony JL, Joanette Y, Belin P (2004) Is voice processing species-specific in human auditory cortex? An fMRI study. Neuroimage 23:840-848.
Friederici AD (2004) Event-related brain potential studies in language. Curr Neurol Neurosci Rep 4:466-470.
Giard MH, Peronnet F (1999) Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. J Cogn Neurosci 11:473-490.
Giard MH, Perrin F, Echallier JF, Thevenet M, Froment JC, Pernier J (1994) Dissociation of temporal and frontal components in the human auditory N1 wave: a scalp current density and dipole model analysis. Electroencephalogr Clin Neurophysiol 92:238-252.
Hagoort P (2003) How the brain solves the binding problem for language: a neurocomputational model of syntactic processing. Neuroimage 20 Suppl 1:S18-29.
Hagoort P (2008) The fractionation of spoken language understanding by measuring electrical and magnetic brain signals. Philos Trans R Soc Lond B Biol Sci 363:1055-1069.
Hari R, Kujala MV (2009) Brain basis of human social interaction: from concepts to brain imaging. Physiol Rev 89:453-479.
Haxby JV, Hoffman EA, Gobbini MI (2002) Human neural systems for face recognition and social communication. Biol Psychiatry 51:59-67.
Itier RJ, Batty M (2009) Neural bases of eye and gaze processing: the core of social cognition. Neurosci Biobehav Rev 33:843-863.
Kleinke CL (1986) Gaze and eye contact: A research review. Psychol Bull 100:78-100.
Lewis JW (2006) Cortical networks related to human use of tools. Neuroscientist 12:211-231.
McCallum WC, Farmer SF, Pocock PV (1984) The effects of physical and semantic incongruities on auditory event-related potentials. Electroencephalogr Clin Neurophysiol 59:477-488.
Mehrabian A, Ferris SR (1967) Inference of attitudes from nonverbal communication in two channels. J Consult Psychol 31:248-252.
Papworth S, Bose AS, Barker J, Schel AM, Zuberbuhler K (2008) Male blue monkeys alarm call in response to danger experienced by others. Biol Lett 4:472-475.
Petkov CI, Kayser C, Steudel T, Whittingstall K, Augath M, Logothetis NK (2008) A voice region in the monkey brain. Nat Neurosci 11:367-374.

Puce A, Perrett D (2003) Electrophysiology and brain imaging of biological motion. Philos Trans R Soc Lond B Biol Sci 358:435-445.
Puce A, Smith A, Allison T (2000) ERPs evoked by viewing facial movements. Cog Neuropsychol 17:221-239.
Puce A, Epling JA, Thompson JC, Carrick OK (2007) Neural responses elicited to face motion and vocalization pairings. Neuropsychologia 45:93-106.
Puce A, Syngeniotis A, Thompson JC, Abbott DF, Wheaton KJ, Castiello U (2003) The human temporal lobe integrates facial form and motion: evidence from fMRI and ERP studies. Neuroimage 19:861-869.
Seyfarth RM, Cheney DL (2003) Meaning and emotion in animal vocalizations. Ann N Y Acad Sci 1000:32-55.
Stephan C, Zuberbuhler K (2008) Predation increases acoustic complexity in primate alarm calls. Biol Lett 4:641-644.
Templeton CN, Greene E (2007) Nuthatches eavesdrop on variations in heterospecific chickadee mobbing alarm calls. Proc Natl Acad Sci U S A 104:5479-5482.
Thompson JC, Clarke M, Stewart T, Puce A (2005) Configural processing of biological motion in human superior temporal sulcus. J Neurosci 25:9059-9066.
Thompson JC, Abbott DF, Wheaton KJ, Syngeniotis A, Puce A (2004) Digit representation is more than just hand waving. Brain Res Cogn Brain Res 21:412-417.
Thompson JC, Hardee JE, Panayiotou A, Crewther D, Puce A (2007) Common and distinct brain activation to viewing dynamic sequences of face and hand movements. Neuroimage 37:966-973.
Van Overwalle F (2009) Social cognition and the brain: a meta-analysis. Hum Brain Mapp 30:829-858.
Van Petten C, Luka BJ (2006) Neural localization of semantic context effects in electromagnetic and hemodynamic studies. Brain Lang 97:279-293.
Wheaton KJ, Pipingas A, Silberstein RB, Puce A (2001) Human neural responses elicited to observing the actions of others. Vis Neurosci 18:401-406.
Wheaton KJ, Thompson JC, Syngeniotis A, Abbott DF, Puce A (2004) Viewing the motion of human body parts activates different regions of premotor, temporal, and parietal cortex. Neuroimage 22:277-288.
Yarbus A (1967) Eye movements and vision. New York: Plenum Press.
Zupan B, Neumann D, Babbage DR, Willer B (2009) The importance of vocal affect to bimodal processing of emotion: implications for individuals with traumatic brain injury. J Commun Disord 42:1-17.