ELSEVIER

Neuroscience Letters 260 (1999) 133–136

Neuroscience Letters

The combined perception of emotion from voice and face: early interaction revealed by human electric brain responses

Beatrice de Gelder*, Koen B.E. Böcker, Jyrki Tuomainen, Menno Hensen, Jean Vroomen

Cognitive Neuroscience Laboratory, Tilburg University, P.O. Box 901535000, LE Tilburg, The Netherlands

Received 1 September 1998; received in revised form 24 November 1998; accepted 2 December 1998

Abstract

Judgement of the emotional tone of a spoken utterance is influenced by a simultaneously presented face expression. The time course of this integration was investigated by measuring the mismatch negativity (MMN). In one condition, the standard stimulus was an angry voice fragment combined with a (congruous) angry face expression. In the deviant pair, the voice expression was kept the same and only the face expression changed to an (incongruous) sad face. The pairs with a deviant visual item evoked a negative electric brain response showing the characteristics of the MMN, which is usually evoked only by auditory deviations. Similar results were obtained by employing incongruous standard and congruous deviant pairs. These findings provide compelling evidence of an early integration of face with voice information in the processing of affect.

Keywords: Emotional expression; Face expression; Voice expression; Crossmodal integration; Electric brain response; Mismatch negativity

Organisms are equipped with multiple sensory systems that each provide inputs via separate routes. Well known cases include ventriloquism (visuo-spatial integration) and audiovisual speech (illustrated most dramatically by the fact that a heard /ba/ and a seen /ga/ are reported as a /da/ [12]. In this last case multiple input routes create redundancy as the same information is provided twice, once by the ears and once by the eyes. Does this actually lead to crosstalk between the modalities? If so, one would expect a faster and more efficient response [17]. Crosstalk should be particularly important for domains of information like those of language or emotions that are of vital importance for the organism.

Simultaneous inputs represent a familiar situation for organisms thriving in natural environments, yet currently only a few instances of multisensorial perception have been studied in the laboratory. An entirely novel case was discovered recently and concerns the concurrent recognition of emotion in the voice and the face [3,11]. With concurrent presentation of a face expression and a voice expression, the voice has a strong impact on recognition of the face expression. Likewise, a face expression has a strong impact on the recognition of the emotion expressed in the voice. When the expression of the face is congruent with that of the voice subjects are faster in recognizing the information presented in the two channels that in one channel only. While shorter response latencies indeed suggest that organisms exploit multiple inputs in the interest of faster and better behavioral response, they do not by themselves provide evidence in support of early integration. Shorter latencies are equally compatible with a race model, that is, each input is processed independently and the fastest process determines the outcome. However, it seems to be in the interest of time and accuracy of behavioral response that redundancy should be exploited on line by a mechanism that integrates the two input channels as they come in, rather than juxtaposing them and letting intrinsic speed decide the outcome. Based on these conjectures, we asked whether already in the early stages of processing the emotional tone of a vocal expression, information from a face expression would

* Corresponding author. Tel.: +31 13 4662167; fax: +31 13 4662370; e-mail: b.degelder@kub.nl

0304-3940/99/$ - see front matter
PII: S0304-3940(98)00963-X

have an impact. The most appropriate methodology for investigating this issue is provided by electrophysiology.

The method of recording electric brain responses (ERPs or event-related potentials) has been applied successfully to address issues of time course of cognitive processes at stake in understanding language. Different components of the brain's electrical response to stimuli have a characteristic profile corresponding to different aspects of the stimulus situation and response (for a recent analysis see [10]. The ERP methodology has been applied to understanding processing of faces [8] and of spoken words with emotional valence [5]. One specific electric brain response is known to be sensitive to stimulation in the auditory modality only, that is, the mismatch negativity (MMN) [14]. The MMN is elicited by a low-probability deviant stimulus in a repetitive train of standard auditory stimuli. It is not under attentional control (but see [20]), and is taken to reflect the functional content of the auditory representations.

Paradoxically, the auditory basis of the MMN makes it also suitable for studying crosstalk between audition and vision. Recently, the MMN has been employed to gather evidence for processing of auditory inputs presented in combination with visual stimuli, and it was found that the properties of the visual stimulus have an impact on the auditory representation and the MMN. Surakka et al. [18] studied the effect of the emotional valence of a picture on the MMN evoked by sine tones. Results indicated that the amplitude of the MMN was significantly attenuated when pictures of positive valence (and low arousal) were viewed as compared to pictures of neutral or negative valence. In another study with a similar type of experimental design as in the current experiment, Sams et al. [16] recorded over the left hemisphere the magnetic counterpart of the MMN (MMN_m) evoked by the auditory presentation of a Finnish syllable /pa/ that was contingent upon a low probability change in the concurrent visual presentation of face and jaw-movements from those belonging to /pa/ to those belonging to /ka/. Subjects perceived either /ka/ or /ta/ (cf. the McGurk effect). Sams et al. [16] observed a MMN_m which started at 180 ms and was localized in the supratemporal auditory cortex.

We exploited the MMN for tracing the early influence of face expressions on processing of the affective tone of the voice. Seven male participants (25 ± 1 years old) received concurrent affective voice and face stimulation. In the EEG experiment the subject was instructed to pay attention to the faces and ignore the auditory stimuli. The faces were taken from the well known Ekman-Friesen set and the voice fragments were short sentences produced by semi-professional actors. Six congruous and six incongruous audio-visual pairs were each presented 140 times in random order to the participants. A congruous pair (angry voice, angry face) served as standard on 85% of the trials, and the same voice fragment combined with a incongruous expression (angry voice, sad face) served as the deviant on 15% of the trials. In the second experimental condition standard and deviant pairs were exchanged. The voice fragment (duration 980 ± 216 ms) started with a variable delay (750–1250 ms) after the onset of the presentation of the face for 2500 ms to reduce interference of the brain response elicited by the faces with measurement of the MMN. The variable intertrial interval (measured from the offset of the visual stimulus) was 0.5–1 s. The brain responses were recorded from Ag/AgCl electrodes placed at F3, Fz, F4, C3, Cz, C4, P3, Pz and P4, all referenced to the nose [14]. The EEG was sampled at 250 Hz (amplification 25000) using an analog bandpass of 0.03–70 Hz (rolloff –12 dB/oct). The impedances were kept below 5 kΩ. Eye movements and blinks were monitored by six EOG electrodes (in two vertical and one horizontal bipolar derivation), and were corrected offline [19]. Trials with other artifacts were discarded. AEPs were averaged (a minimum of 91 trials, range 91–120 per condition) time-locked to the auditory stimulus. A 125 ms pre-stimulus time interval was used as baseline. Digital high-pass filtering (6 Hz, rolloff –24 dB/oct) further eliminated the interference of the visual response. We showed behaviorally that the -system is tuned to combine this dual input [3], and see also Table 1 for separate behavioral results with the current stimulus set. If integration takes place very early during processing of the vocal expression, this will be reflected in an MMN. If not, only the visual ERP components will show evidence of processing of the face.

The results showed that an early auditory brain response is elicited at F3, Fz and Cz (latency 178 ms; Fig. 1a, bold line) when after a number of presentations of a voice-face pair with congruent expressions, a pair consisting of the same voice expression but a different face expression is presented. The same result was found when a congruous voice-face pair follows after a number of incongruous pairs (Fig. 1a, thin line). The MMN was measured from these difference waves which were obtained by subtracting the standards (preceding the deviant stimuli) from the deviants. The average amplitudes in the interval from 172 to 184 ms were entered into a repeated measures Analysis of Variance (ANOVA) with the factors Congruence (congruent or incongruent), Anterior-Posterior Electrode Position (frontal,

Table 1

	Angry face (ms)	Sad face (ms)	Paired t (d.f. = 7)
Angry voice	532 ± 148	575 ± 144	3.87***
Sad voice	545 ± 157	501 ± 109	2.28

**$P < 0.05$; ***$P < 0.001$. In a separate behavioral experiment, eight different participants from the ones in the EEG experiment (21 ± 2 years old) were presented five times with 24 congruent (four sad faces paired with three sad voice fragments and four angry faces paired with three angry voice fragments) and 24 incongruent stimulus pairs (4 sad faces paired with three angry voice fragments, and vice versa). The participants were instructed to label the voice expression as fast as possible. Button press Reaction Times, measured from the onset of the stimuli, were slower for incongruous pairs.

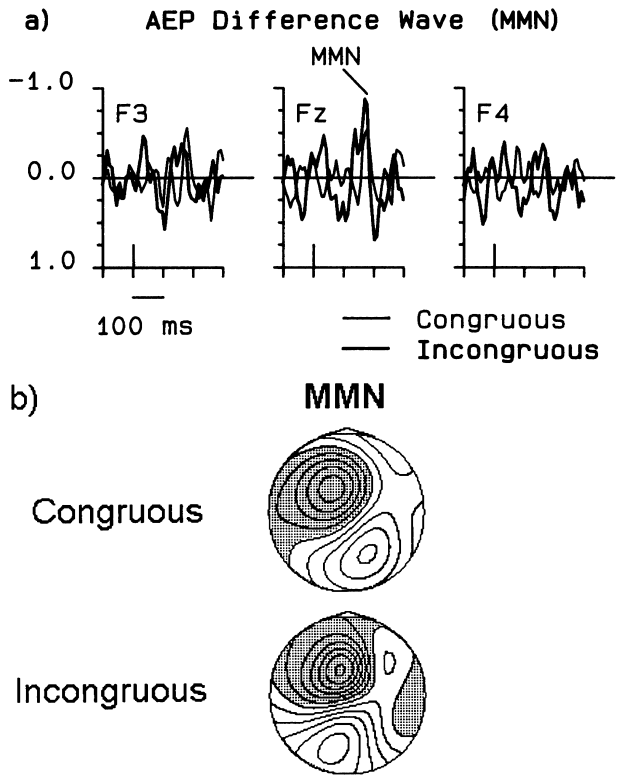

Fig. 1. (a) The grand average ($n = 7$) deviant-standard difference wave of the auditory electric brain potential (AEP) for both congruous (thin line) and incongruous (thick line) face-voice pairs. Surplus negativity evoked by the deviant stimulus pair is plotted upwards. The vertical bar on the horizontal axis indicates the onset of the auditory stimulus. The *y*-axis depicts amplitude (in mV). b) The isopotential map (0.1 μV between lines; shaded area negative) at 178 ms, showing the scalp distribution of the MMN for voices with congruous and incongruous faces.

central or parietal), and Laterality (left, mid-line or right). Results indicated that the MMN was larger at left and mid-line sites than over the right hemisphere, at frontal ($F_{2,12} = 7.33$, $P < 0.05$, Greenhouse-Geisser $e = 0.67$) and central ($F_{2,12} = 4.15$, $P < 0.05$, $e = 1.0$) but not at parietal ($F_{1,12} < 1$, n.s.) chains of electrodes (Fig. 1b). The MMN was maximal at Fz, where it (tended to) differ from zero voltage baseline for both congruous (tested separately by one-way ANOVA, $F_{1,6} = 5.23$, $P = 0.06$, mean = –0.62 ± 0.27 mV (SEM)) and incongruous ($F_{1,6} = 21.81$, $P < 0.01$, mean ∴– = – 1.0 ± 0.22 mV (SEM)) deviants, which do not differ significantly from each other ($F_{1,6} = 2.93$, n.s). Neither a main effect, nor any interactions were found for Congruity. This brain wave strongly resembles the MMN typically associated with detection of a change in the auditory input [14]. The experimental paradigm, and the polarity, distribution (Fz maximum) and latency (178 ms) of the response are entirely compatible with those of the MMN. The absence of any subsequent slow positivity (P3 or P3a), in the unfiltered wave forms, indicates that the influence of the voice processing is independent of attention and that the observed peaks are not an instance of N2b [14].

When time-locked to the presentation of the face, the brain responses evoked by the deviant faces (Fig. 2a) showed a large positivity at 535 ms that is clearly later than the face-specific positive ERP component, which usually occurs at around 200 ms [8]. A second repeated measures ANOVA on the average amplitudes of the visual difference wave between 485 and 585 ms showed that this positivity was significantly different from zero voltage baseline (tested separately by one-way ANOVA, $F_{1,6} = 16.07$, $P < 0.01$). Its amplitude was larger at central and parietal than at frontal electrodes ($F_{2,12} = 6.03$, $P < 0.05$, $e = 0.61$) and it was also larger over midline than over lateral electrodes ($F_{2,12} = 12.91$, $P < 0.01$, $e = 0.90$; Fig. 2b). Neither a main effect, nor any interactions were found for Congruity. The oddball paradigm (the series of standards and deviants), the instruction to attend to the faces (that were the changing stimuli), and the distribution (Pz maximum) and the latency of the peak (535 ms) are all arguments that favor an interpretation of this positive deflection as an instance of P3 [6]. This in turn suggests that the subjects were indeed attending to the faces.

These ERP results extend significantly our previous behavioral evidence of crossmodal bias [3]. They highlight that

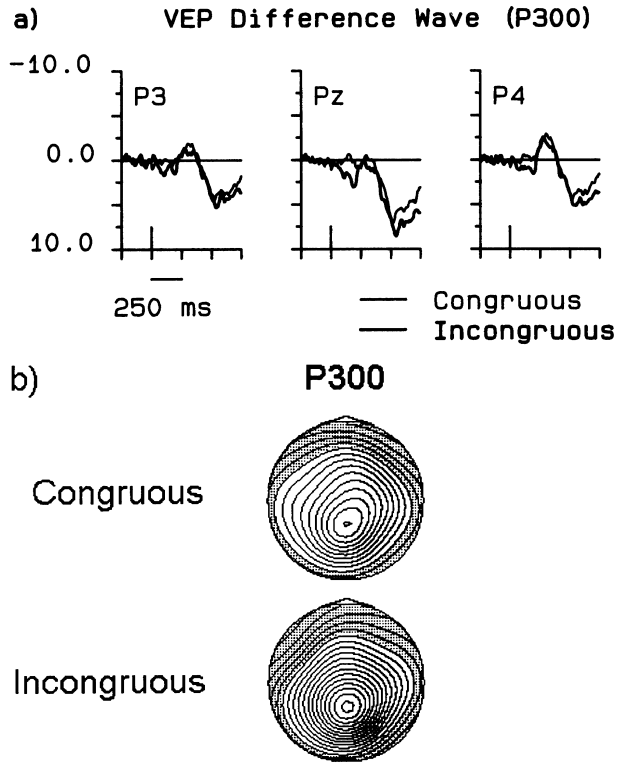

Fig. 2. (a) The grand average ($n = 7$) deviant-standard difference wave of the visual electric brain response for both congruous (thin line) and incongruous (thick line) face-voice pairs. The vertical bar on the horizontal axis indicates the onset of the visual stimulus to which these averages were time-locked. The y-axis depicts amplitude (in mV). Note the different scaling of both axes compared to Fig. 1 that should be consulted for further legends. (b) The isopotential map (0.5 μV between lines; shaded area negative) at 535 ms, showing the scalp distribution of the P300 for congruous and incongruous faces.

this is a mandatory process, a claim that was so far only based on behavioral results showing the same pattern as for the McGurk illusion and the ventriloquism effect [2]. Most importantly, this study provides unique information about the time course. Audiovisual integration takes place early on during the perceptual process, at the latest at 178 ms after voice onset, thereby strongly suggesting that the integration of face and voice processing occurs before both have been (fully) processed independently. Finally, the fact that the type of the stimulus pair combination was not relevant in evoking the MMN (that is, no main effect nor interactions involving Congruity) seems to indicate that the integration involves low-level perceptual features.

The ERP method was adopted because of its potential for addressing issues of timing, but with its low spatial resolution it cannot contribute directly to the identification of brain regions implicated in this crossmodal interaction. Indirectly though our result offers some indications. With regard to the neural pathways responsible for the early audio-visual combination, it should be noted that the primary neuroanatomical source of the MMN is usually localized in the supratemporal cortex [14]. For language stimuli, it is in the left supratemporal cortex [15], which is compatible with the present lateralization of the MMN. Furthermore, neurons in the temporal lobe (especially in the posterior areas of the superior and medial temporal lobe) respond selectively to faces and facial emotions as shown by single-cell recordings of monkey brains [1] and intraoperative recordings of awake humans [7]. Information from the facial expression might be routed via direct cortico-cortical pathways or be relayed via subcortical structures such as the amygdala so far known for its important role in facial emotion recognition [13]. However, there is some evidence that amygdala does not (directly) participate in the generation of the MMN to simple tones. Kropotov et al. [9] studied six patients who had intracranial electrodes implanted in several different anatomical sites. No MMN could be recorded from electrodes placed in amygdala, hippocampus, basal ganglia nor the ventrolateral nucleus of the thalamus. The only sites eliciting MMNs were Brodmann areas 21 and 42 (in the temporal cortex).

Our study suggests that crossmodal processes involved in perception of multiple emotional cues are a topic of investigation in their own right extending the scope of emotion research beyond the more familiar studies of either face or voice processing in normals and their associations or dissociations in brain damage. Our results also point to the possibility that some brain disorders might specifically affect audiovisual perception even with spared processing in each separate modality [4].

[1] Baylis, G.C., Rolls, E.T. and Leonard, C.M., Selectivity between faces in the responses of a population of neurons in the cortex in the superior temporal sulcus of the monkey, Brain Res., 342 (1985) 91–102.

[2] Bertelson, P., Vroomen, J., de Gelder, B. and Driver, J., The ventriloquist effect: it does not depend on the direction of overt visual attention, Percept. Psychophysiol., (1999) in press.

[3] de Gelder, B. and Vroomen, J., Perception of emotions by ear and by eye, Cognition and Emotion, submitted.

[4] de Gelder, B., Bachoud-Levi, A.-C., Vroomen, J. and Popelier, T., Impaired perception of face and emotion in Huntington's disease, J. Cog. Neurosci. Suppl., (1998) 113.

[5] de Pascalis, V., Morelli, A. and Montirosso, R., Word recognition time and event-related potentials during emotional processing, J. Psychophysiol., 4 (1990) 229–240.

[6] Donchin, E., Ritter, W. and McCallum, W.C., Cognitive psychophysiology: the endogenous components of the ERP. In Callaway, E., Tueting, P. and Koslow, S.H. (Eds.), Event-Related Brain Potentials in Man, Academic Press, New York, 1978, pp. 349–411.

[7] Fried, I., Mateer, C., Ojemann, G., Wohns, R. and Fedio, P., Organization of visuospatial functions in human cortex, Brain, 105 (1982) 349–371.

[8] Jeffreys, D.A., Evoked potential studies of face and object processing, Visual Cognit., 3 (1996) 1–38.

[9] Kropotov, J.D., Näätänen, R., Sevostianov, V., Alho, K., Reinikainen, K. and Kropotova, O.V., Mismatch negativity to auditory stimulus change recorded directly from the human temporal lobe, Psychophysiology, 32 (1995) 418–422.

[10] Kutas, M. and Dale, A., Electrical and magnetic readings of mental functions. In Rugg, M.D. (Ed.), Cognitive Neuroscience. Studies in Cognition, MIT Press, Cambridge, MA, 1997, pp. 197–242.

[11] Massaro, D.W. and Egan, P.B., Perceiving affect from the voice and the face, Psychon. Bull. Rev., 3 (1996) 215–221.

[12] McGurk, H. and MacDonald, J., Hearing lips and seeing voices, Nature, 264 (1976) 746–748.

[13] Morris, J.S., Frith, C.D., Perrett, D.I., Rowland, D., Young, A.W., Calder, A.J. and Dolan, R.J., A differential neural response in the human amygdala to fearful and happy facial expressions, Nature, 383 (1996) 812–815.

[14] Näätänen, R., Attention and Brain Function, Lawrence Erlbaum, Hillsdale, NJ, 1992.

[15] Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vainio, M., Alku, P., Ilmoniemi, R.J., Luuk, A., Allik, J., Sinkkonen, J. and Alho, K., Language-specific phoneme representations revealed by electric and magnetic brain responses, Nature, 385 (1997) 432–434.

[16] Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O.V., Lu, S.-T. and Simola, J., Seeing speech: visual information from lip movements modifies activity in the human auditory cortex, Neurosci. Lett., 127 (1991) 141–145.

[17] Stein, B.E. and Meredith, M.A., The Merging of the Senses, MIT Press, Cambridge, MA, 1993.

[18] Surakka, V., Tenhunen-Eskelinen, M., Hietanen, J.K. and Sams, M., Modulation of human auditory information processing by emotional visual stimuli, Cognit. Brain Res., 7 (1998) 159–163.

[19] Van den Berg-Lenssen, M.M.C., Brunia, C.H.M. and Blom, J.A., Correction of ocular effects in EEG's using an autoregressive model to describe the EEG: a pilot study, Electroenceph. clin. Neurophysiol., 73 (1989) 72–83.

[20] Woldorff, M.G., Hackley, S.A. and Hillyard, S.A., The effects of channel-selective attention on the mismatch negativity wave elicited by deviant tones, Psychophysiology, 28 (1991) 30–42.