Perceiving Emotions by Ear and by Eye

Beatrice de Gelder

Cognitive and Affective Neurosciences Laboratory, Tilburg University, Tilburg
MGH-NMR Center, Charlestown, MA
degelder@nmr.mgh.harvard.edu

Abstract

Affective information is conveyed through visual as well as auditory perception. The present paper considers the integration of these channels of information, that is, the multisensory processing of emotion. Findings from behavioral, neuropsychological and imaging studies are reviewed.

1. Introduction

The overwhelming majority of studies on human emotion recognition have used face expressions (for a recent overview, see [1]) but only a few studies have studied how emotion in the voice is perceived (see [2] for a review). A few researchers have looked into possible correspondences between the two different sensory processes in search of common processing resources shared by visual and auditory affective processes alike [3]. This issue is different from that of multisensory processing of emotion, which we recently started to explore.

Traditionally, multisensory research has addressed very simple phenomena like for example the way a light flash and a sound beep are combined or the enhancement of localization of a weak auditory stimulus by the simultaneous presentation of a weak light flash. In contrast, phenomena like audio-visual speech and audio-visual emotion perception are complex cases which represent situations that are much more typical of the rich environment in which the brain operates and reflect constraints the brain faced in the course of evolution. Of course, it should be possible to apply similar methodological imperatives to the study of complex cases as has been done to the simple ones in the past [4].

2. Behavioral experiments measuring cross-modal bias between face and voice expressions

In our initial behavioral studies of inter-sensory perception of emotion we combined a facial expression with a short auditory voice fragment and instructed participants to attend to and categorize either the face or the voice depending on the condition. These experiments provided clear evidence that a vocal expression which is irrelevant for the task at hand, can influence the categorization of a facial expression presented simultaneously [5]. Specifically, participants categorizing a happy or fearful facial expression were systematically influenced by the expression of a voice (e.g., the face was judged as less fearful if the voice sounded happy). Massaro and Egan [6] obtained similar results using a synthetic face expression paired with a voice expression. The effect of the face on voice recognition disappeared when face images were presented upside down adding further proof that the face expression was the critical variable [7]. Subsequently we explored the situation where subjects were asked to ignore the face but had to rate the expression of the voice. A very similar cross-modal effect was observed, this time consisting in an influence of the facial expression on recognition of the emotional expression in the voice ([8], Experiment 3).

One might ask whether attention rather than perceptual integration itself is the binding factor of audio-visual perception of emotion [9]. Participants judged whether a voice expressed happiness or fear, while trying to ignore a concurrently presented static facial expression. As an additional task, participants were instructed to add two numbers together rapidly (Experiment 1), or count the occurrences of a target digit in a rapid serial visual presentation (Experiment 2), or judge the pitch of a tone as high or low (Experiment 3). The face had an impact on judgments of the perceived emotion in all Experiments. This cross-modal effect was independent of whether or not subjects performed a demanding additional task indicating that the integration of visual and auditory information about emotions is a mandatory process, unconstrained by attentional resources.

3. Electrophysiological studies of the time course of multisensory affect perception

The phenomenon of Mismatch Negativity (MMN, [10]) can be used as a means of tracing the time course of the combination of the affective tone of voice with information provided by the expression of the face [11]. In the standard condition subjects were presented with concurrent voice and face stimuli with identical affective content (a fearful face paired with a fearful voice). On the anomalous trials the voiced expression was accompanied by a face containing an incongruent expression. We reasoned that if the system were tuned to combine these inputs, as was suggested by our behavioral experiments, and if integration is reflected by an influence of the face on how the voice is processed, this would be apparent in some auditory ERP components. Our results indicated that when following repeated presentations of a voice-face pair with the same expression, a pair is presented where the voice stimulus is the same but the expression of the face is different, an early (170 ms) deviant response is elicited at the scalp level (frontal topography). This response

strongly resembles the MMN, which is typically associated with a change (whether in intensity, duration or location) in a train of standard-repetitive auditory stimuli [10]. Our results are consistent with previous EEG results showing that the pitch MMN may be influenced by the simultaneous presentation of positive non-facial stimuli [12].

4. Neuro-anatomy of audio-visual perception of emotion

Using brain-imaging methods (fMRI) an important element of a mechanism for cross-modal binding of fearful face-voice pairs could be found in the amygdala [13]. In this study, subjects heard auditory fragments paired with either a congruent or an incongruent facial expression (happiness or fear) and were asked to judge the emotion from the face. When fearful faces were accompanied by short sentence fragments spoken in a fearful tone of voice an increase in activation was observed in the amygdala and the fusiform gyrus suggesting binding of face and voice expressions and the role of the amygdala in this process [14]. Unlike observed in our behavioral studies, no such advantage was observed for happy pairs. This could suggest that the rapid integration across modalities is not as automatic as it is for fear signals.

An intriguing possibility is that presentation in one modality activates areas typically associated with stimulation in the other modality. One would expect this pattern to obtain in cases where a naturalistic, as opposed to an arbitrary, association obtains between auditory and visual stimulus, as is the case for speech and emotion. For example, using fMRI we investigated auditory sadness and observed strong and specific orbitofrontal activity. Moreover, in line with the possibility just raised, among the observed foci there was strong activation of the left fusiform gyrus, an area typically devoted to face processing, following presentation of sad voices [15].

5. Selectivity in audio-visual affect pairing

How selective is the pairing mechanism that accounts for event identity? So far we have mentioned studies of audio-visual affect where the visual stimuli consisted of facial expressions. Such pairings are based on congruence in stimulus identity, i.e., emotional meaning, across the two input modalities. However, other visual stimuli such as objects and pictures of visual scenes also carry affective information and are equally conspicuous in the daily environment. Similarly, there are other sources of auditory affect information besides affective prosody, the most obvious candidates being word meaning and non-verbal auditory signals. If semantic relationship was the only determinant of identity-based pairings, either of those visual inputs should combine with either of those two alternative auditory messages. In fact, a similar issue has been raised for audio-visual speech where the issue is not settled [16]. Selectivity is an important issue for identity pairings and we believe that learning more about it will reveal important insights into the biological basis of multisensory perception. On the other hand, we do not know to what extent the familiar boundaries of our own species are the limits of our biological endowment. For example, for human observers facial expressions of higher apes might bind more easily with vocalizations than do visual scenes because the latter share more biologically relevant properties with human faces [17].

We subsequently addressed the same issue using a different technique, single-pulse transcranial magnetic stimulation (TMS, [18]). Two types of stimulus pairs were compared, one consisting of arbitrary paired stimuli where the pairing was learned and the other of natural pairings as described above. Participants were trained on the two types of pairs to ensure that the same level of performance was obtained for both. Our question was whether TMS would interfere with cross-modal bias obtained with meaningless shape/tone pairs (Learned condition) but not with voice/face pairs (Natural condition). Single pulse TMS applied over the left posterior parietal cortex at 50, 100, 150 and 200 ms disrupted integration at 150 ms and later but only for the learned pairs. Our results suggest that content specificity as manipulated here could be an important determinant of audio-visual integration. Of course, without further comparative investigation of the different parameters, we cannot be sure that this distinction reflects some other difference between the two categories of stimuli.

6. Qualitative differences between conscious and non-conscious audio-visual perception

The question we asked earlier about the role of attention can also be raised about awareness. An important aspect of audio-visual integration is the role of stimulus awareness. Understandably, if we can provide evidence that unseen stimuli or stimuli the observer is not aware of, still exert a cross-modal bias, then the case for dealing with an automatic, mandatory perceptual phenomenon is even stronger. By the same token the requirement that audio-visual bias should be studied in situations that are minimally transparent to the observer, is equally met when observers are unaware of the second element of the stimulus pair. Patients suffering from visual agnosia including an inability to recognize facial expressions do present a unique opportunity for investigating this issue. Patient AD has severe face recognition problems due to bi-lateral occipito-temporal damage [19]. Since her recognition of facial expressions is almost completely lost but recognition of emotions in the voices is intact we could investigate spared covert recognition of facial expressions with a cross-modal bias paradigm. With this indirect testing method we found clear evidence of covert recognition as her recognition of emotions in the voice was systematically affected by the facial expression that accompanied the voice fragment she was rating.

Patients suffering from hemianopia sometimes have residual visual abilities of which they are not aware (see [20]) and thus offer an opportunity to study audio-visual integration under conditions of subjective visual blindness. In the hemianopic patient GY we found behavioral and electrophysiological evidence for a cross-modal bias of facial expressions that were not perceived consciously on

processing of the emotion in the voice [21]. By manipulating awareness this way, we could also look at a possible interaction between awareness and type of audio-visual pairing, for example whether combinations of emotional stimuli other than the human face equally influenced perception of emotional voices. For this purpose we designed two types of pairs each with a different visual component, one consisting of facial expression-voice pairs and the other on emotional scene-voice pairs [21]. ERPs were measured in two hemianopic patients and we compared the pattern obtained in the intact hemisphere where patients were conscious of the visual stimuli with that obtained in the blind hemisphere were there was no visual awareness. We found support for the hypothesis that unlike naturalistic pairings, semantic pairings might require conscious perception and mediation by intact visual cortex. These results are in line with previous studies that have provided evidence in favor of qualitatively different processing systems for conscious and non-conscious perception.

7. Multisensory perception of emotion and motor theory

The preceding results raise many theoretical questions about the selectivity of the multisensory pairing mechanism and its organismic basis. Unlike the case of audio-visual speech, our daily environment provides a great variety of emotional cues visual (a multitude of emotionally laden images assault our senses) as well as auditory ones (whether verbal or not verbal, whether human or from other species). Yet, the voice-face pairs seem to stand out (at least that is what our present data indicate). Would a special status of voice-face pairs indicate a specific functional underpinning for this kind of pairs and might this be the basis for selectivity providing scene-voice pairs, scene-word pairings or face-word pairs with different characteristics than face-voice pairs? An interesting possibility is that emotional face and voice perception might activate similar motor schemes like for example the ones also underlying the production of these voice and face expressions. An analogy that comes to mind is with the motor theory of speech perception [22]. There are at present some findings in support of this view. As shown by Dimberg and collaborators [23], facial expressions are automatically imitated even when one is not aware of them. Sensorimotor cortex plays a role in the perception of emotional expressions of the face [24]. While these findings are intriguing, they only provide part of the solution to audio-visual perception

8. References

[1] Adolphs R., "Neural systems for recognizing emotion", *Curr. Opin. Neurobiol.,* 12:69-77, 2002.

[2] E. D. Ross, "Affective prosody and the aprosodias", in *Principles of behavioral and cognitive neurology*, M. M. Mesulam, Ed. London: Oxford University Press, 2000, pp. 316-331.

[3] Borod, J.C., Pick, L.H., Hall, S., Sliwinski, M., Madigan, N., Obler, L.K., Welkowitz, J., Canino, E., Erhan, H.M., Goral, M., et al., "Relationships among facial, prosodic, and lexical channels of emotional perceptual processing", *Cognition Emotion,* 14:193-211, 2000.

[4] Bertelson, P. and de Gelder, B., "The psychology of multisensory perception", in *Crossmodal Space and Crossmodal Attention*, C. Spence and J. Driver, Eds. Oxford: Oxford University Press, in press

[5] de Gelder, B., Vroomen, J., and Teunisse, J.-P. , "Hearing smiles and seeing cries: The bimodal perception of emotions", *B. Psychonomic Soc.*, 30, 1995.

[6] Massaro, D.W. and Egan, P.B., "Perceiving affect from the voice and the face", *Psychon. B. Rev.*, 3:215-221, 1996.

[7] de Gelder, B., Vroomen, J., and Bertelson, P., "Upright but not inverted faces modify the perception of emotion in the voice", *Curr. Psychol. Cogn*, 17:1021-1031, 1998.

[8] de Gelder, B. and Vroomen, J., "The perception of emotions by ear and by eye", *Cognition Emotion,* 14:289-311, 2000.

[9] Vroomen, J., Driver, J., and de Gelder, B. "Is cross-modal integration of emotional expressions independant of attentional resources?", *Cogn. Affect. Behav. Neurosci.*, 1:382-387, 2001.

[10] Näätänen, R. *Attention and Brain Function*. Lawrence Erlbaum, Hillsdale, 1992.

[11] de Gelder, B., Böcker, K.B., Tuomainen, J., Hensen, M., and Vroomen, J., "The combined perception of emotion from voice and face: early interaction revealed by human electric brain responses", *Neurosci. Lett.*, 260:133-136, 1999

[12] Surakka, V., Tenhunen, E., Hietanen, J.K., and Sams, M., "Modulation of human auditory information processing by emotional visual stimuli", *Cognitive Brain Res.*, 7:159-163, 1998.

[13] Dolan, R., Morris, J., and de Gelder, B., "Crossmodal binding of fear in voice and face", *Proc. Natl. Acad. Sci. U.S.A.,* 98:10006-10010, 2001.

[14] Goulet, S. and Murray, E.A., "Neural substrates of crossmodal association memory in monkeys: The amygdala versus the anterior rhinal cortex", *Behav. Neurosci.*, 115:271-284, 2001.

[15] Malik, N., de Gelder, B., and Breitner, J.C.S., (manuscript in preparation).

[16] Vroomen, J. and De Gelder, B., "Crossmodal integration: a good fit is no criterion", *Trends Cogn. Sci.*, 4:37-38, 2000.

[17] de Gelder, B., van Ommeren, B., and Frissen, I., "Feelings are not specious. Recognition of facial expressions and vocalizations of chimpanzees (Pan troglodytes) by humans", Cognitive Neuroscience Society Annual Meeting, 2003.

[18] Pourtois, G. and de Gelder, B., "Semantic factors influence multisensory pairing: a transcranial magnetic stimulation study", *Neuroreport*, 13:1567-1573, 2002.

[19] de Gelder, B., Pourtois, G., Vroomen, J., and Bachoud-Levi, A.C., "Covert processing of faces in prosopagnosia is restricted to facial expressions: evidence from cross-modal bias", *Brain Cognition*, 44:425-444, 2000.

[20] Weiskrantz, L., *Consciousness Lost and Found,* Oxford University Press, Oxford, 1997

[21] de Gelder, B., Pourtois, G., and Weiskrantz, L., "Fear recognition in the voice is modulated by unconsciously recognized facial expressions but not by unconsciously recognized affective pictures", *Proc. Natl. Acad. Sci. U.S.A.*, 99:4121-4126, 2002.

[22] Liberman, A.M. and Mattingly, I.G., "The motor theory of speech perception revised", *Cognition*, 21:1-36, 1985.

[23] Dimberg, U,. Thunberg, M., and Elmehed, K., "Unconscious facial reactions to emotional facial expressions", *Psychol. Sci.*, 11:86-89, 2000.

[24] Adolphs, R., Damasio, H., Tranel, D., Cooper, G., and Damasio, A.R., "A role for somatosensory cortices in the visual recognition of emotion as revealed by three-dimensional lesion mapping", *J. Neurosci.* 20:2683-2690, 2000.