Rejoinder

Bimodal emotion perception: integration across separate modalities, cross-modal perceptual grouping or perception of multimodal events?

Beatrice de Gelder
Tilburg University, The Netherlands and Université Libre de Bruxelles, Belgium

Jean Vroomen
University of Louvain, Belgium

The combined perception of affective information provided by seeing a facial expression and hearing a tone of voice presents us with a phenomenon that has received little systematic attention in the laboratory. We argue that although the FLMP appears good at modelling the presently available data, more complex theoretical questions must be raised. We discuss the theoretical shortcomings of the FLMP and envisage two other theoretical frameworks that could guide future research.

Biological organisms are equipped with sense organs that receive inputs simultaneously. In many cases, behaviour results from processing of multisensorial inputs and perception is about multimodal events (de Gelder, 1999). Recent studies on the processing of affective information provided by seeing a facial expression and hearing an affective tone of voice signify the start of a more naturalistic approach to information processing in the new field of affective neuroscience.

We are pleased that our data fit so well the model argued for by Massaro and Cohen. This clearly indicates convergence between two independent

Requests for reprints should be sent to Beatrice de Gelder, Department of Psychology, University of Tilburg, Warandelaan 2, PO Box 90153, 5000 LE, Tilburg, The Netherlands; email: b.degelder@kub.nl.

http://www.tandf.co.uk/journals/pp/02699931.html

sets of findings concerning the same phenomenon. However, unlike Massaro and Cohen suggest, we are unwilling to let this convergence settle the theoretical debate that looms behind the fit between data and one way of modelling them. We believe that alternative theoretical accounts must be envisaged in order to bring into focus the richness and complexity of these phenomena. On a previous occasion we have argued against the confounding in Massaro's approach between questions that concern the input modality and issues of modularity of information (de Gelder & Vroomen, 1989). More recently we have taken issue with the basis on which Massaro argues the symmetry in the impact of visual speech and written language on processing spoken language (Vroomen & de Gelder, in press). Here, we envisage three different frameworks for the study of bimodal emotions. The first of these, the FLMP defended by Massaro and collaborators is effectively capable of modelling the presently available data. In the FLMP, inputs are processed separately and then combined and integrated. A second alternative is to view cross-modal integration as resulting from perceptual grouping. A third, more speculative proposal is that perceptual grouping points to a class of natural multimodal events as the basic perceptual objects.

1. *Bimodal emotion perception as integration.* Massaro is among the few researchers to have developed a theory of perception that focuses on situations where the system is exposed to multiple inputs. Recently he has argued that the FLMP offers a unique model of cross-modal integration taking place in perception of affect by ear and by eye (Massaro & Egan, 1996). Basically, the FLMP assumes that inputs provided by a voice and by a facial expression are evaluated independently against the corresponding prototype of emotion in long-term memory (e.g., anger). Following evaluation, a general integration algorithm operates and combines representations. Perceptual output or the resulting response category is a function of the information in the constituent representations. Given its pattern-matching approach to perception/recognition the model offers no a priori constraints on more or less natural, or possible/impossible multimodal pairings. So far, the FLMP has not faced up to the costs of such an "ultra-light" theoretical approach to multimodal cognition because all the evidence reported so far by Massaro and collaborators has supported the model. But so far only very few aspects of the phenomenon have been studied and the FLMP provides no heuristics and offers no cues about the underlying theoretical architecture of the sensory processing system that would constrain intersensory integration. Just picture a research programme that systematically needs to look at any combination between any kind of inputs (linguistic, pictorial, etc.) of whatever sensory origin (visual, auditory, tactile, gestural, etc.) in whatever cognitive domain (language, affect, reasoning, memory, etc.). Moreover, given this absence

of theoretical framework, there is nothing typically cross-modal about integration as envisaged in the FLMP. For this model integration is insensitive to content or input modality. For example, it assumes that the impact of an angry voice on an angry or fearful face presents the same phenomenon as the impact one facial expression could have on processing another facial expression. But in the face/voice case we are dealing with cross-modal integration and in the other (so far not explored) case, the process takes place within the visual modality and an advantage in latencies for rating the expression of a face in the latter situation might be due to statistical summation.

2. *Bimodal perception as cross-modal perceptual grouping.* Because its focus is limited to integration metrics no other aspects of multimodality matter for the FLMP than the connection between input and stored prototypes and the temporal constraints binding the two are not sensitive to the nature of the phenomena. But spatial and temporal constraints of the kind that are critical in ventriloquism may be equally important for bimodal emotion. And the fact that we know that audiovisual speech and ventriloquism dissociate illustrates the interdependency between content based groupings (strongest in audiovisual speech) and spatio-temporal grouping principles at stake in ventriloquism (Bertelson, Vroomen, Wiegeraad, & de Gelder, 1994). Characteristically, the FLMP ignores the existence of domain-specific boundaries as argued for in more modular approaches to perception. In the latter perspective, the fact that perceiving speech by ear is influenced by seeing lip movements fits a framework which takes *spoken language* to be a relatively autonomous input domain to which dedicated perceptual mechanisms are selectively tuned. As a consequence, speech-relevant inputs, whether delivered by hearing or by seeing are grouped together but inputs provided by *written language* (combinations of visual or auditory speech with written words) are not, or only indirectly (see Vroomen and de Gelder, in press, for more detailed discussion).

3. *Bimodal events and event-related perception.* Our current research aims at uncovering the constraints on inputs to bimodal emotion perception by also looking for negative evidence or for instances where integration does not appear to obtain. For example, the way the FLMP sees it, an angry face presented together with the written word "afraid" will affect the rating of the face in the same way a word or sentence spoken in an angry tone does. Likewise, the word "angry" spoken in a neutral tone of voice will presumably activate the corresponding prototype but will this affect rating of the face? Preliminary evidence indicates that this is not the case. This suggests that integration is a function of other aspects of the input than just the availability of a stored prototype. To account for both positive and

negative evidence of integration, one may envisage an approach that is theoretically more ambitious and traces the origin of the perceptual grouping to multimodal events. Underlying the cases like audiovisual speech or affect there may be neural and functional mechanisms that target certain types of bimodal events. For example, although the association of a written word and a facial expression may well have been stored and automatised, it is not a natural one in the sense that it does not correspond to a naturally occurring multimodal event. Which multimodal combinations correspond to natural kinds of bimodal events and which go back to learned associations and are natural and automatised pairings subject to different sets of constraints? These are at present empirical questions and one can be confident that future theories will benefit just as much from the cases where integration does not obtain than from the ones where it does.

Manuscript received 18 September 1999
Revised manuscript received 24 September 1999

REFERENCES

Bertelson, P., Vroomen, J., Wiegeraad, G., & de Gelder, B. (1994). Exploring the relation between McGurk interference and ventriloquism. *Proceedings of the International Congress on Spoken Language Processing* (pp. 559–562). Yokohama, Japan: ASJ.

de Gelder, B. (1999). Recognizing emotions by ear and by eye. In R. Lane et al., (Eds.), *Cognitive neuroscience of emotions.* New York: Oxford University Press.

de Gelder, B., & Vroomen, J. (1989). Models in the mind, modules on the lips. *Behavioural and Brain Sciences, 124,* 762–763.

Massaro, D.W., & Egan, P.B. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin and Review, 3,* 215–221.

Vroomen, J., & de Gelder, B. (in press). Crossmodel integration: A good fit is no criterion. *Trends in Cognitive Sciences.*