Crossmodal integration: a good fit is no criterion

Massaro's Fuzzy Logical Model of Perception (FLMP) is one of the dominant approaches to intersensory integration[1]. It was originally developed for the situation where listeners hear tokens from a /ba/-/da/ continuum while viewing a face articulating /ba/ or /da/. More recently, FLMP has been extended to other situations, such as ventriloquism and the bimodal perception of emotion. Massaro has argued that FLMP provides an adequate and universal model of perception, in particular about how information across modalities is combined.

In this letter, however, we present several examples showing that FLMP does not enlighten the underlying perceptual processes. Massaro makes a strict distinction between information and information processing. FLMP is concerned only with the latter. The model makes specific assumptions about how information is combined and Massaro's extensive work suggests that, in almost all cases, FLMP fits the data more accurately than do alternative models that rely on different assumptions (additive or categorical). For Massaro, this is the signature of a universal law by which information is integrated: it is the crux of what he refers to as information processing. However, with this exclusive emphasis on information processing, a disregarded issue is whether there are content-based constraints on what sources of information do or do not integrate.

Reading versus lip-reading

The prime example is the case of reading versus lip-reading. From a behavioral, developmental and neurological perspective, there are many reasons why reading is unlike lip-reading. And of course, Massaro is well aware of them, and he would agree that the information is different. The main point for Massaro is that presenting speech with read or lip-read information follows the same principles of information processing in both cases, and these principles are captured by FLMP. In support of this view he presents data showing that reading the letters B or D has the same impact on an auditory /bi/-/di/ continuum as lip-reading /bi/ or /di/ (Ref. 1, Fig. 1, p. 311). He also refers to data in a previous paper, which showed that both lip-read /ba/ or /da/ and read BA or DA had a similar impact on an auditory /ba/-/da/ continuum (although the effect of lip-reading was in this case nine times that of reading[2]). Whatever the differences between these two experiments, in both cases FLMP was superior to other models in fitting the number of fusion responses (/di/ or /da/).

However, one issue arising from this work that deserves further consideration is what happens with combination responses (/bdi/ or /bda/)? When lip-read /bi/ is presented together with auditory /di/ subjects should perceive /bdi/, because of the McGurk effect (similarly, lipread /ba/ with auditory /da/ should be perceived as /bda/). However, we are not aware of any demonstration that the letter B (read) when combined with auditory /di/ is perceived as /bdi/, as would be expected if reading and lip-reading were equivalent. The absence of such a demonstration may in fact suggest that the congruence between reading and lip-reading only holds for fusion responses, which would restrict the scope of the FLMP considerably.

A more difficult problem is to determine at which processing level crossmodal interactions take place. Massaro has shown that there is an impact from reading and lip-reading on speech identification, but the question remains of how one can be sure that the interactions occurred at the same perceptual level in both cases and not at different stages. As a parallel example, consider the well-known Stroop phenomenon, in which subjects who are asked to say aloud the color of the ink that a color word (e.g. blue) is written with are heavily influenced by whether the word itself is congruent or incongruent with the color of the ink. Nobody would suggest that this occurs because the word itself changes the perception of the color of the ink. Rather, in the Stroop task there is competition at a response stage. In the current case, then, the issue is not whether reading and lip-reading interact at all with speech, but whether they interact at the same processing level, and whether FLMP allows one to distinguish between the various forms that this interaction may take.

At present, this is difficult to evaluate because FLMP has only been tested with three or, recently, four rather abstract processing levels (evaluation, integration, assessment and response selection). It is clear that the cognitive processes underlying vision and audition are much more complex than that, and that crossmodal interactions might therefore take place at levels not envisaged by the model (e.g. crossmodal interactions at the level of scene analysis or attention[3]). One possibility is that lip-reading interacts with speech at a perceptual level, while reading interacts with speech at a decision stage. There is some intuitive appeal to this proposal because it fits the observation that some subjects report 'hearing' something different when an auditory token is combined with a different lip-read token, while other researchers have failed to obtain perceptual effects (finding only biases) when written text is combined with speech[4].

Ventriloquism

Massaro also claims that FLMP can be applied to the ventriloquist scenario in which subjects are asked to judge the apparent origin of a sound when presented with a visual signal that originates from a different location. Subjects tend to underestimate the distance between the auditory and visual signals and sometimes even fuse them. It has been argued that this kind of crosstalk involves a decision about what is variously called, 'pairing'[5], 'unity'[6], or 'object-identity'[7]. The basic idea is that the perceptual system is required to decide whether auditory and visual information originate from a single source. In order to make this decision, the spatial proximity of the information sources and the similarity of the temporal pattern are thought to be used[8]. As a consequence, as the distance between an auditory and a visual stimulus is increased there will be two opposing effects on the overall bias of the perception of the auditory stimulus being displaced towards the visual attractor.

Firstly, the proportion of trials on which an interaction occurs will decrease with increasing separation because the pairing decision is supported in fewer trials. On the other hand, the size of the attraction on those few trials in which there is pairing will increase with increasing separation[9,10]. The overall effect of this might be that the crossmodal bias decreases when the distance between sound and light is increased. For example, when Bermant and Welch measured the visual bias on audition using separations between stimuli of 10, 20 and 30 degrees, they obtained a decrease in the bias from 57 to 17 to 12%, respectively[11]. Bertelson and Radeau also found that the visual bias decreased as the distance increased[12]. Models that do have a pairing decision predict that the size of the ventriloquist effect should decrease when the sound and light sources move so far apart that the pairing decision can no longer be supported. Sound and vision are then treated as separate events with no crossmodal influence. However, FLMP does not include a process similar to a pairing decision because in FLMP there is always integration. Thus, in FLMP the ventriloquist effect should *increase* when distance increases, because the farther apart the auditory and visual stimuli, the more the visual signal supports a distant location. The (weighted) average of the auditory and visual location should thus move away from the auditory location.

How then is it possible to conclude that the FLMP provides a good description of the data? The answer appears to be that the FLMP does not make a prediction. While FLMP is a very flexible tool, it fits data retrospectively by adjusting truth values until there is a satisfying fit.

Thus, when the visual signal moves away from the sound source and the bias decreases, one may obtain a good fit by decreasing the visual support for the more distant location. The crucial point, however, is that the truth values are meaningless because there is no guarantee that there is a correspondence with the perceptual mechanisms that lead to these truth values. Thus, the fact that fuzzy sets of mathematics can describe aspects of results does not enlighten us about the underlying mechanism: a good fit is therefore no criterion to accept FLMP as an adequate theory of perception.

Acknowledgement
We would like to thank Paul Bertelson for his insightful comments on a previous version of this paper.

Jean Vroomen and Beatrice de Gelder
Tilburg University, Department of Psychology, Warandelaan 2, PO Box 90153, 5000 Le Tilburg, The Netherlands.
tel: +31 13 466 2394
fax: +31 13 466 2370
e-mail: j.vroomen@kub.nl

References

1 Massaro, D.W. (1999) Speechreading: illusion or window into pattern recognition. *Trends Cognit. Sci.* 3, 310–317
2 Massaro, D.W. *et al*. (1998) Visible language in speech perception: lipreading and reading. *Visible Lang*. 22, 9–31
3 Driver, J. and Spence, C. (1998) Attention and the crossmodal construction of space. *Trends Cognit. Sci*. 2, 254–262
4 Frost, R. *et al*. (1988) Can speech perception be influenced by simultaneous presentation of print? *J. Mem. Lang*. 27, 741–755
5 Radeau, M. and Bertelson, P. (1977) Adaptation to auditory–visual discordance and ventriloquism in semirealistic situations. *Percept. Psychophys*. 22, 137–146
6 Welch, R.B. and Warren, D.H. (1980) Immediate perceptual response to intersensory discrepancy. *Psychol. Bull*. 88, 638–667
7 Bedford, F.L. (1999) Keeping perception accurate. *Trends Cognit. Sci*. 2, 4–11
8 Radeau, M (1994) Auditory–visual spatial interaction and modularity. *Cahiers de Psychologie Cognitives* 13, 3–51
9 Vroomen, J. Ventriloquism and the nature of the unity assumption. In *Cognitive Contributions to the Perception of Spatial and Temporal Events* (Aschersleben, G. *et al*., eds), Elsevier (in press)
10 Bertelson, P. Ventriloquism: A case of crossmodal perceptual grouping. In *Cognitive Contributions to the Perception of Spatial and Temporal Events* (Aschersleben, G. *et al*., eds), Elsevier (in press)
11 Bermant, R.I. and Welch, R.B. (1976) Effect of degree of separation of visual–auditory stimulus and eye position upon spatial interaction of vision and audition. *Percept. Mot. Skills* 43, 487–493
12 Bertelson, P. and Radeau, M. (1981) Crossmodal bias and perceptual fusion with auditory–visual discordance. *Percept. Psychophys*. 29, 578–584

Reply to Vroomen and de Gelder

Given that I am sympathetic to Vroomen and de Gelder's commentary[1], I can only hope that they have failed to read my lips (or my research papers) rather than misunderstood what they have read. Admittedly, my short review article[2] could be read out of context and the reader could easily believe that I have gone beyond the evidence given (in the same way that our perception often goes beyond the information given). We use and promote our information-processing framework primarily because it encourages the investigator to determine the stage (level in Vroomen and de Gelder's terms) of processing responsible for various behaviors.

I will show that their two main points can be easily pursued within our framework, after a short qualification of the origins of the FLMP. Vroomen and de Gelder state that, 'Originally, it (the FLMP) was developed for the situation where listeners hear tokens from a /ba-/da/ continuum while viewing a face articulating /ba/ or /da/.'[1] It is important to assure the reader that the FLMP was with us well before McGurk and MacDonald published their McGurk effect[3]. The model was originally developed to account for the integration of several auditory cues in speech perception and for various sources of information in sentence processing[4–6]. In assessing the model, it is important to note that the FLMP was not derived simply to describe speech perception by ear and eye, but rather to describe pattern recognition more generally.

Lip-reading versus reading

First, Vroomen and de Gelder question whether written text is operating at the same stage as visible speech when these sources are separately combined with auditory speech. It is intuitive, somehow, to believe that the influence of visible speech is more real than the influence of written text. However, it is worth noting a couple of caveats. First, Vroomen and de Gelder should not dismiss the positive finding of Frost *et al.*[7] as simply a bias because we now know that biases can be truly perceptual. This possibility was pointed out long ago by Paul Bertelson[8] when investigators tried to dismiss his ventriloquism effect as a response bias when analysed within the context of signal detection theory (for further discussion of the important distinction between perceptual bias and decision bias, see Ref. 9). How would one test whether the two types of visual input operate differently? Our experiment is simply a first step along that road. Contrasting different models of performance should then follow. It is straightforward to formulate a model based on the interpretation proposed by Vroomen and de Gelder in which the visual input has its influence on decision rather than perception (Ref. 10, Chapter 2). The outcome of these tests would speak to the issue of analogous processes in reading and lip-reading.

Vroomen and de Gelder[1] state that 'the issue is not whether reading and lip-reading interact at all with speech, but whether they interact at the same processing level, and whether FLMP allows one to distinguish between the various forms that this interaction may take'. This question has always been of central interest to us and is why I argue for the formalization and testing of alternative models. The post-perceptual guessing model, the auditory dominance model, and the 'Race' model have all been tested as alternatives to the FLMP, primarily because they assume different forms of interaction of the two sources of information.

Additional experiments can be generated to distinguish between various theoretical explanations. One important source of evidence comes from the nature of the judgments that are given. Specifically, Vroomen and de Gelder are interested in combination responses, such as /bdi/. Before discussing the story of /bdi/, it should be noted that when and how often these combinations occur is highly variable and unpredictable. In an early study with open-ended alternatives, Repp *et al.* found no combination responses (Ref. 11, and see Ref. 12, pp. 52–54). In our studies with /bd/ as one of the specified response alternatives, we have found up to 80% (Ref. 13) and as low as 10% combination responses (Ref. 10, p. 146) when a visual /ba/ is paired with an auditory /da/. In order to understand whether these combination responses should be equivalent in the reading and lip-reading conditions, however, it is first necessary to understand why they occur in lip-reading. Our interpretation has been that a visual /b/ paired with an auditory /d/ provide two sources of information that are consistent with /bd/. A visual /b/ looks a lot like a visual /bd/, and an auditory /d/ is somewhat similar to auditory /bd/. Thus, /bd/ can be a reasonable percept given these two sources. This explanation also predicts very few /db/ judgments when a visual /d/ is paired with an auditory /b/. In this case, a visual /d/ is very different from a visual /db/. These types of constraints probably do not occur in the reading situation, however, because the written letter activates some speech-like representation without actually providing a speech stimulus. This situation is more analogous to the