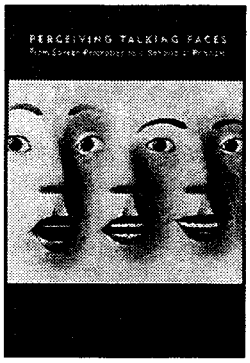
A Universal Algorithm for Pattern Recognition?



Perceiving Talking Faces
by Dominic W. Massaro
Cambridge, MA: MIT Press, 1998. 494 pp. ISBN 0-262-13337-7. $55.00

Review by Carol A. Fowler

Perceiving Talking Faces charts its author's wide-ranging and career-spanning research program. As an overview of Massaro's research enterprise, it is a fine book that should be of interest to many researchers. Although audiovisual speech perception remains the domain in which Massaro has done much of his research, *Perceiving Talking Faces* reveals that he has extended his model of perception to many other domains. Indeed, he makes the provocative claim in Chapter 4 that his model represents a universal law of pattern recognition (in company, he proposes, with such laws as Miller's [1956] seven plus or minus two and Weber's Law).

Perceiving Talking Faces summarizes two major enterprises. One enterprise is the development of a synthetic talking head, dubbed Baldi, that has been used in Massaro's research on audiovisual speech and emotion perception and for which Massaro and the coauthors of his last three chapters, Michael Cohen and Michael Berger, foresee a number of speech-related applications. The other enterprise, and the one on which I will focus, is Massaro's ambitious program of research in which he has developed and tested his Fuzzy Logical Model of Perception (FLMP). I have a much deeper appreciation and understanding of Massaro's work having read *Perceiving Talking Faces* than I did having read many of his articles and his related book (Massaro, 1987). His accomplishments, as outlined in *Perceiving Talking Faces*, are impressive.

As an overview of Massaro's research enterprise, it is a fine book that should be of interest to many researchers.

Delightfully and usefully, *Perceived Talking Faces* comes with a CD that allows the reader to meet Baldi and to see and hear the stimuli that Massaro and his colleagues have used in their work on speech perception and perception of emotion.[1] If readers wonder how natural or effective the speech movements or emotional expressions of Baldi are, they can judge for themselves. Helpfully, for readers who cannot read *Perceiving Talking Faces* near a computer, there is an appendix briefly summarizing what they should look for in each demonstration. However, for the next edition of this book, I recommend also that CD band numbers be printed in the margins of pages on which demonstrations are described so that backtracking readers can easily remind themselves of the context in which each demonstration was discussed.

The FLMP

The FLMP is a quantitative model of pattern recognition intended to capture the character of pattern recognition at an "algorithmic" (fairly abstract) level of description. Perceivers are assumed to have available multiple cues for perceptible events. Because environments are noisy, the entire set of cues that might specify an event is unlikely to be present on any given occasion. The pattern recognition algorithm enables recognition under such noisy conditions. It includes three successive (or in a dynamic version of the model, cascading) stages. The first is evaluation, in which fuzzy logical truth values are assigned to features associated with a relevant set of prototypes in memory. Truth values range between 0 and 1 and express the degree to which cues in the input match feature specifications associated with the prototype. Products of truth values assigned to each prototype are computed in the integration stage. Finally, in a decision stage, the prototype associated with the largest product is selected.

Massaro has repeatedly tested the FLMP against alternative quantitative models, and it almost invariably fits the data better than do its competitors. In the domain of speech perception, its success rate is not as impressive as that may sound, because, with the exception of TRACE, a connectionist model of perception (McClelland & Elman, 1986), competitors have been models that Massaro has made up to provide a comparison with FLMP, and the made-up models are generally implausible. Certainly it is good that FLMP fits data better than implausible models do, but one wonders how stiff the competition they provide can be. An absolute measure of fit is preferable, and Massaro provides two in Chapter 10, an innovation of which I had not been aware.

In articles and synthesized in *Perceiving Talking Faces*, Massaro shows that the FLMP provides excellent (and superior) fits to other kinds of data than to speech. It fits aspects of word reading, audiovisual perception of emotional expression, categorization, cognitive decision making and more. Within the domain of speech perception itself, Massaro shows that FLMP fits data obtained from speakers of a variety of languages, even though response patterns show cross-linguistic differences, and it fits the variety of response patterns obtained from individual young children, young adults, and older adults. FLMP's fitting these diverse data sets reveals that the critical differences underlying the data patterns are differences in information extraction, not in information processing, and the successful fits underlie Massaro's claim that the FLMP represents a universal law of pattern recognition.

The FLMP and Power

The extraordinary success of FLMP's fits does raise the question whether the model has excessive power (see Cutting, Bruno, Brady, & Moore, 1992). Massaro argues that it does not, showing that FLMP fails to fit data that it should not fit. Syllables (provided on the CD) were audiovisually presented with a sufficient temporal offset that one hears a syllable and then sees a mouth gesture that clearly could not be the cause of the acoustic syllable. Even so, participants were instructed to choose the one response that best reflected both what they saw and heard. Accordingly, responses were influenced by information in both modalities, however, there was no integration of auditory and visual information in pattern recognition, and FLMP fit the response patterns badly.

The FLMP is too powerful; however, it does fit data that it should not.

[1] *The delightfulness of the demonstrations is dimmed slightly by some misnumbering errors in the text. In Chapter 1, the text refers to CD bands 1.5, 1.6, and 1.7, when bands 1.4, 1.5, and 1.6 are intended. In Chapter 14, band 14.5 is mentioned twice, the second time when 14.6 is intended.*

DOMINIC W. MASSARO, Department of Psychology, University of California, Santa Cruz.

CAROL A. FOWLER, Haskins Laboratories, New Haven, Connecticut.

Tests involving the McGurk effect (e.g., McGurk & MacDonald, 1976), in which visible and audible speech are cross-dubbed, constitute much of Massaro's research on speech perception. Demand characteristics of some versions of these experiments can be substantial, because participants readily intuit that the experimenter is interested in whether what they see will influence what they report hearing. For the classic McGurk type of trial (discrepant acoustic and visible syllables, with the latter's place of articulation clearly visible), in which speechreading changes the phenomenal experience of hearing the speech, demand characteristics must play only a small role in response choices. Indeed, Massaro (1987; see also Chapter 9) has shown that, under selective attention instructions (report only what you heard), response patterns differ only a little from those generated under more liberal instructions that invite perceivers to show a visual influence. However, some demonstrations on the CD verify that selective attention to the acoustic signal is easy. In these cases, sight and sound do not integrate in perception; accordingly, the FLMP should not fit the data.

One example is a test ostensibly showing that manual gestures accompanying speech can give rise to a McGurk effect on speech identification (e.g., Thompson & Massaro, 1994; see also Chapter 6). A videotape showed a hand pointing either to a doll or a ball. On audiovisual trials, simultaneous with the pointing gesture, a syllable was presented from a ball-to-doll acoustic continuum. In the CD demonstration, the pointing gestures did not affect what I experienced or identified hearing (comparing identifications with eyes open and closed), but they did affect children's responses. Here are the instructions that children were given: "Sometimes the woman just points to the thing you should choose, . . . sometimes she just says it, and sometimes she both points and says which thing you should choose. Each time, though, she wants you to choose one thing, either the ball or the doll" (p. 171). These instructions urge participants to provide responses on audiovisual trials that combine visual and audible information, and participants complied, but not by means of perceptual integration if my performance on the demonstration is indicative. The FLMP should not provide any better fits to these data than to the temporally offset syllables, but it does.

The auditory and visual expressions of emotion on the CD likewise do not integrate. The sight of a smiling face does not change the way an angry rendition of "Please" sounds. FLMP should not fit these data well, but it does. Had McGurk and MacDonald (1976) tried dubbing manual gestures and speech, or acoustic and visual expressions of emotion instead of speech, there would be no phenomenon known as the McGurk effect.

FLMP and the Speech Community

The FLMP has not been influential in the community of speech scientists, and *Perceiving Talking Faces*, whatever its other accomplishments may be, it is unlikely to effect change in that realm. One reason is Massaro's lack of attention to work by other investigators in the field. In *Perceiving Talking Faces*, some theories of speech perception are addressed fleetingly (and are dismissed casually); most receive no mention at all.

Another reason why the speech community has not embraced FLMP is because it either evades or offers transparently wrong solutions to problems of central concern to speech scientists. It evades addressing the issue of what the "features" are that perceivers use to identify speech segments; it offers clearly inadequate placeholder features: (static) distal visible features—for example, lips closed—but (dynamic) proximal audible features—for example, a rise in the second formant—for /ba/. (My proposal that objects of perception are uniformly distal is dubbed unparsimonious.)

FLMP offers a transparently wrong answer to the question of the nature of perceptual objects, which it identifies as prototypes (in speech, "summary descriptions of perceptual units of the language", p. 61). Massaro does not address or even allude to the challenges that Pisoni and colleagues have mounted to ideas that perceptual objects of speech perception are appropriately identified as prototypes (e.g., Goldinger, 1998; Lively & Pisoni, 1997; Palmieri, Goldinger, & Pisoni, 1993).

In the FLMP, prototypes for speech represent consonant-vowel (CV) syllables; accordingly, CVs are perceptual units. There are many reasons why this has to be wrong. An obvious one is that English is not a CV language; many of its syllables are not CVs (e.g., "strength" (/strɛŋθ/ is a CCCVCC). In any case, a proposal that CVs are perceptual units leaves fundamental questions unanswered and, I believe, unanswerable in the FLMP framework. Why do perceivers judge that the CVs, /di/ and /du/, for example, begin with the same consonant, but with a different consonant from the beginning of /gi/? Why do no language communities anywhere contrast the beginning sounds in /di/ and /du/? Why do we use the same letter to spell the beginning letters of /di and /du/? (For that matter, why do we use an alphabet at all, rather than a syllabary?)

Then there is /bda/. Visible /ba/ dubbed onto acoustic /da/ frequently leads perceivers to report hearing /bda/. So they must have a prototype, /bda/. But where would it come from? Infants would not develop it in utero, where Massaro hazards that other speech prototypes may be acquired. It violates the phonotactics of English and of most languages, and so it does not occur in listening experience, Massaro's references to such words as "abdicate" (where /b/ and /d/ are in different syllables) notwithstanding. Massaro has a solution. Indeed, he chastises Larry Rosenblum and me for writing (Rosenblum & Fowler, 1991) as if we were not aware of it. We were, but I, for one, never could make sense of it. As Massaro points out, "in several venues" he has written: "Prototypes are generated for the task at hand" (p. 355).[2] But how would a perceiver know that he or she needed to generate a prototype, /bda/, for the task at hand? Massaro emphasizes that FLMP's decision rule picks the prototype best supported by the evidence; it does not require an absolute degree of good fit between evidence and prototype. So how can perceivers know that none of their existing prototypes is right for visual /ba/, acoustic /da/? (More than that, how do perceivers know what their prototypes' names are; what makes a cluster of features amount to /bda/ or /di/ or /du/?)

The Readership for PTF

Perceiving Talking Faces is generally very clearly and interestingly written, but there is some ambiguity regarding the intended audience. Early on, Massaro defines "level" of an independent variable for the reader, and so I thought (very briefly) that *Perceiving Taking Faces* was intended as a popular science book. Compatibly, the book helpfully provides an appendix of phonetic symbols for English consonants and vowels (confusingly using two different symbols for the vowel in "book"). But it does not help readers out on the Dutch phonetic symbols that they require in Chapter 5. Chapter 6 discusses the "infamous" Linda problem from the decision-making literature, without telling us what it is. Dutch phoneticians will breeze through the book, particularly if they have a side interest in decision making, and even if they do not know what levels of independent variables are. Other readers may have more difficulty, but they will find the struggle largely worthwhile. ❑

---

[2] *Massaro has, indeed, written that "Prototypes are generated for the task at hand" in many, many venues. In most of those venues, he follows that exact wording with this exact wording: "In speech perception, for example, we might envision activation of all prototypes corresponding to the perceptual units of the language being spoken" (e.g., Massaro, 1987, p. 17). But activation is not an "example" of generation. These sentences are contradictory.*

References

Cutting, J. E., Bruno, N., Brady, N. P., & Moore, C. (1992). Selectivity, scope and simplicity of models: A lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General, 121,* 364–381.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105,* 251–279.

Lively, S., & Pisoni, D. B. (1997). On prototypes and phonetic categories: A critical assessment of the perceptual magnet effect in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 1665–1679.

Massaro, D. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, NJ: Erlbaum.

McClelland, J., & Elman, J. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1–86.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97.

Palmieri, T., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 309–328.

Rosenblum, L. D., & Fowler, C. A. (1991). An audio-visual investigation of the loudness/effort effect for speech and nonspeech events. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 976–985.

Thompson, L. A., & Massaro, D. W. (1994). Children's evaluation and integration of pointing gestures during comprehension. *Journal of Experimental Child Psychology, 57,* 327–354.