

Review of Massaro, D.W. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle* (MIT Press, Cambridge, Mass.).

Noël Nguyen

to appear in the *Journal of Phonetics*

Send all correspondence to:

Noël Nguyen
Institut de Phonétique
Université de Provence
29 avenue Robert Schuman
13621 Aix-en-Provence, France
e-mail: Noel.Nguyen@lpl.univ-aix.fr

1 SYNOPSIS

1.1 General outline

This book is concerned with how multiple sources of information are processed in speech perception and, more generally, in pattern recognition. It is based upon an important research programme conducted by Massaro and his colleagues over the last two decades. The book focuses on the perception of so-called bimodal speech, addressing a wide range of issues about the way in which visual information (as provided by the speaker's face) and auditory information are combined with each other by the perceptual system. The scope of the book is much larger, however, as Massaro's purpose here is to describe and defend a new psychological law relevant to a wide variety of domains. In contrast to already well-established laws of the same kind (e.g. Weber's law of perception), which are all unidimensional, the new principle is multidimensional, in that it describes how several factors impact behaviour. This principle is embodied in a computational model of pattern recognition, the Fuzzy Logical Model of Perception (FLMP), whose latest version is presented and discussed in detail. The FLMP is systematically contrasted with alternative computational models, using a broad perceptual database as benchmark throughout the book. In a separate part, the book also deals with methods for synthesizing talking faces in experiments on bimodal speech perception, and introduces Baldi, the talking face developed by Massaro and his coworkers. The book is accompanied by a CD-ROM which contains a series of demonstrations relating to many of the topics dealt with.

The book is divided into four main sections. Section 1, "Perceiving Talking Faces", focuses on the perception of speech by ear and eye. Massaro reviews the most significant empirical findings in that domain, discusses the main methodological issues, and presents a general classification of the existing computational models of bimodal speech perception. Central to this section is the idea that speech perception obeys a general behavioural principle of integration between different sources of information. Section 2, "Broadening the Domain", aims at assessing how well this principle holds up across broad individual and situational variability. The author demonstrates that inter-

individual variations in how bimodal speech is perceived, depending on the listener's age or native language for instance, can be accounted for within the FLMP framework. Using examples taken from different perceptual and cognitive situations, Massaro also defends the idea that the FLMP adequately describes information processing irrespective of these situational differences. Section 3, "Broadening the Framework", opens with a presentation of an extended and more explicit version of the FLMP, designed in particular to account for the dynamics of speech perceptual processing. The section also includes a detailed analysis of the methodological issues involved in assessing quantitative predictions in psychology, along with a discussion of the critiques expressed by other investigators about the FLMP over the years. Finally, Section 4, "Creating Talking Faces", is specifically dedicated to the synthesis of visual speech.

1.2 The new behavioural principle

Although many readers may already be familiar with Massaro's Fuzzy Logical Model of Perception, I shall here assume the contrary, and proceed to present a brief outline of the model.

A central assumption of the FLMP is that pattern recognition involves a common set of processes regardless of the specific nature of the patterns. Speech is not seen as being associated with a dedicated processing module, as in the motor theory of speech perception (Liberman, 1996) for instance. On the contrary, the sensory information is assumed to be processed in the same way whether our brain is recognizing speech sounds, letters, or manual gestures, to take but a few examples. In any of these cases, the FLMP postulates that mapping a stimulus into a unique perceptual category entails going through three main stages of processing, the feature evaluation stage, the feature integration stage, and the decision stage.

The evaluation stage consists of converting the available sources of information into a set of properties referred to as features. Each feature is given a continuous (fuzzy truth) value, and represents the degree to which the stimulus corresponds to each of a set of internal prototypical patterns, along a particular perceptual dimension. Thus, one important visual feature in the perception

of CV syllables is the degree of opening of the lips. The model therefore assumes that the internal prototypes available to the perceptual system will specify that the lips are open at the onset of the syllable for /da/, closed for /ba/, etc. In a second stage, the features are integrated with each other, so as to determine the overall degree of match of the sensory input with each of the prototypes (e.g. each of the syllables known to the receiver). In the third and final stage, a decision is taken, on the basis of the relative goodness of match of the input with each prototype.

The FLMP makes a number of specific assumptions at each stage in this process. First, it hypothesizes that all of the available sources of information are simultaneously brought into play in pattern recognition. Thus, visible speech and auditory speech are both assumed to have an influence on how bimodal speech is perceived. Second, different sources of information are assumed to be evaluated independently of each other. This means for example that visible speech does not have any effect on how auditory speech is converted into a set of features, the two sources of information being combined at a later stage of processing only. The model also makes specific assumptions about how sources of information are integrated with each other (multiplicative rule), and about how decisions are taken (relative goodness rule).

A major prediction of the model is that "the influence of one source of information is greatest when the other source is neutral or ambiguous" (19). This prediction is best illustrated by an experiment whose results served as a database for testing models of pattern recognition on several occasions in the book (chapters 2 and 11). In this experiment, synthetic auditory stimuli ranging on a continuum between /ba/ and /da/ were crossed with visual stimuli also varying between /ba/ and /da/. The bimodal stimuli were presented to subjects in a forced-choice identification task, along with each of the unimodal stimuli. (This expanded factorial design is shown by Massaro to be the most appropriate experimental design for determining how two sources of information are combined with each other in pattern recognition.) For the bimodal stimuli, the main results are typically depicted as a two-factor plot, with the proportion of /da/ responses on the ordinate, the

levels of the auditory source of information on the abscissa, and a different curve for each of the levels of the visual source of information. When represented in that way, the results clearly show a statistical interaction between the two sources of information. Specifically, the influence of one source of information proves to be larger in the middle, ambiguous range of the other source. This interaction graphically takes the shape of an American football, which is for this reason presented throughout the book as the hallmark of the the Fuzzy Logical Model of Perception.

In summary, Massaro proposes a universal principle of perceptual cognitive performance to explain pattern recognition. According to this principle, "people are influenced by multiple sources of information in a diverse set of situations. In many cases, these sources of information are ambiguous and any particular source alone does not usually specify the appropriate interpretation. The perceiver appears to evaluate the multiple sources of information in parallel for the degree to which each supports various interpretations, integrate them together to derive the overall support for each interpretation, assess the support of each alternative based on all of the alternatives, and select the most appropriate response." (p. 291).

2 CRITICAL EVALUATION

2.1 General evaluation

This book is clearly a major contribution to the study of speech perception and, more generally, to cognitive psychology. It is admirably clear and is written in quite an elegant manner.

I do not doubt that the book will be read with great interest by research scientists from many different fields. This work is the result of an ambitious intellectual endeavour aimed at introducing a new behavioural law, which is placed by Massaro on an equal footing with Weber's law of perception, or the power law of learning. Speech scientists are presented with an extensive series of experiments on the perception of bimodal speech. Whatever stance they take in that domain, they should find quite challenging Massaro's view that speech perception constitutes but one as-

pect of a much more general form of cognitive processing, namely pattern recognition. Computer scientists working in the field of speech technology should be particularly interested in the book's final section about the synthesis of visual speech.

Regardless of their background, readers should also find the book worth using as a tutorial on the experimental methods available for investigating speech perception. A great variety of experimental paradigms and tasks are discussed at length by Massaro, who also extensively discusses the methods for assessing computational models of pattern recognition and, in particular, for fitting these models to observed results. In that respect, using the results of the experiment described above as a reference database was quite a good initiative in my view, as this allows the reader more easily to understand Massaro's point as new issues are raised, without having again to go through the details of the experimental design each time.

The book should also prove an invaluable resource for teaching. Care was taken to select prototypical results, as well as to set this work in its historical context. A number of rather fascinating anecdotes and historical references are given, going from McGurk's personal account of the discovery of the McGurk effect, to an audio-visual rendition of the introduction to George Miller's seminal article on the ubiquitousness of the number 7 plus or minus 2, with Miller's face texture-mapped onto Baldi's wire-frame head. The CD-ROM that accompanies the book enables the reader directly to experience the psychological illusions associated with the perception of bimodal speech, and constitutes as such a most useful research and teaching tool.

On the negative side, Massaro's use of the /ba-/da/ experiment as a leading strand throughout obviously results in the book being focused on the perception of non-sense syllables. Although the interaction of visible speech and audible speech in word recognition is mentioned on a number of occasions (e.g. pp. 21–23 and pp. 181–182), little attempt is made to demonstrate how the model could be applied to more complex situations presented by whole word recognition, let alone

connected speech. I also was surprised by the fact that little place was devoted to presenting other current theories and models of speech perception. Although models such as TRACE are mentioned on several occasions in the book, I think it is fair to say that the FLMP is still given the lion's share.

The book also has some minor defects such as the absence of a list of figures, and the fact that some of the CD-ROM bands (1.4, 1.5 and 1.6) are referred to incorrectly in the text. The list of the CD-ROM selections should have pointed to the pages where each band is referred to. In another domain, it would have been quite interesting to have the perceptual database used in the book made available on the CD-ROM. Although this would have probably required a substantial amount of additional work, I should also have found it useful to be provided with an interactive version of the main computational models discussed in the book (FLMP, the RACE model, the Single Channel model, etc.). The FLMP model can be downloaded from Massaro's laboratory Web site at Santa Cruz (<http://mambo.ucsc.edu>), but it is currently distributed in FORTRAN code which has to be modified and recompiled for each new set of data, an operation which is probably out of reach of many students in psychology or linguistics.

2.2 Specific comments

I now would like to comment in greater detail on two themes from the book which more specifically relate to the perception of auditory speech, namely the role of features in speech processing and the time course of speech processing.

2.2.1 Features

Most useful are the extensive comments made by Massaro about the status of features in his model (see in particular Chapter 2 and Chapter 10). A key question is how close are the features postulated in the FLMP model to what is classically referred to as distinctive features in standard phonetics (cf. Stevens & Blumstein, 1978, to take but an example). The book makes it clear to me that there is no direct relation between the former and the latter.

As indicated above, the FLMP postulates that there are three main stages of processing in pattern recognition: the feature evaluation stage, the feature integration stage, and the decision stage. Specific assumptions are made in the model about how features are integrated with each other, and how a decision is taken depending on the outcome of this integration. From a set of feature values, therefore, the model will predict the probability of occurrence of each possible response (e.g. "ba" and "da").

However, attention should be paid to the fact that these feature values are in no way derived from the stimulus. They are actually determined in an *posteriori* manner, from the subjects' observed responses, using an algorithm (STEPIT) which allows the deviation between these responses and the predicted ones to be minimal. Features are seen in the model as *free parameters*, whose values are set on the basis of the actual performance of the subject in the pattern recognition task, so as to make the model perform at its best, i.e. to maximize its goodness of fit. According to Massaro, "[the model is] *predicting* the exact *form* of the results, but *postdicting* the actual quantitative *values* that make up the overall predictions" (p. 294, his emphasis).

In other words, the stimulus is on no occasion explicitly mapped onto the internal features of the FLMP model. In that respect, features as defined in the FLMP look markedly different from phonetic features. Let us take for example the opposition between /ba/ and /da/, on which much emphasis is put in the book. Acoustically, /b/ and /d/ are said to differ from each other according to the feature grave-acute, /b/ being classified as grave and /d/ as acute. As is the case with FLMP features, grave and acute can be viewed as target values referring to prototypical stops. However, the grave-acute feature is explicitly defined in acoustical terms (e.g. slope of the short-term spectrum at the release of the stop, see Stevens & Blumstein, 1978). On the contrary, the exact nature of the FLMP features remains undetermined, their values being subject to one main constraint which is to make the model account for the subjects' responses as accurately as

possible. Thus, the acoustic structure of the stimulus is not directly taken into consideration in the estimation of the feature values.

In the experiments using audible speech, FLMP features do lend themselves to an acoustic interpretation. In the /ba/-/da/ experiment for example, the prototypes for /ba/ and /da/ are assumed to include one auditory feature, namely the variations in frequency of the second (F2) and third (F3) formants at the onset of the vowel (slightly falling F2-F3 for /da/, rising F2-F3 for /ba/). However, this interpretation stems from the fact that F2 and F3 onset frequencies were precisely the acoustic parameters manipulated by the experimenters to synthesize the auditory continuum between /ba/ and /da/. In other words, the acoustic significance of the FLMP features is derived from the way in which the experiment has been designed. The model does rely on a particular system of acoustic features, but this system is embodied in the experimental design, and is as such external to the model itself.

In practice, therefore, the issue of how speech sounds are mapped onto features is not addressed in the model. Why this is so is not clear to me. On several occasions, Massaro suggests that determining in advance how a given individual will convert a given stimulus into a set of feature values is simply out of our reach. This stimulus-to-feature mapping shows a variability which is said to be analogous to the variability of the weather: there are just too many previous contributions and influences to allow quantitative prediction (135). A fundamental distinction is in fact established in the FLMP between the intake of *information*, i.e. the stimulus-to-feature mapping, and *information processing*, i.e. how features are combined with each other and mapped into a response (cf. p. 135). While the FLMP predicts that the information will be processed in the same way from one individual to the other, regardless of whether it relates to speech sounds, facial movements, manual gestures, etc., it is assumed that the way in which this information is extracted from the stimulus is on the contrary subject to too many sources of variations to be accurately characterized ahead of time. In my understanding, this means that the so-called evaluation stage

cannot be accounted for by the model, or at least not with much accuracy.

However, at least on one occasion Massaro does suggest that this limitation is not consubstantial with every model of perception and pattern recognition, and could be circumvented in some way. According to him, one could indeed “easily hypothesize functions relating the feature values to the stimulus levels, [although] that would represent a *model of information* in addition to one of information processing" (294, my emphasis). This suggests that building such a model of information is feasible. Whether there is a possibility of the FLMP being completed with a model of this kind, i.e. an explicit stimulus-to-feature mapping stage, is an issue which remains to be addressed.

2.2.2 The time course of speech processing

Time plays quite a central role in different ways in the book. First, Massaro shows how the FLMP can be explicitly formalized to account for the dynamics of perceptual processing (chap. 9). This formalization is presented in reply to criticisms expressed by a number of investigators (e.g. McClelland, 1991), who have pointed out that the FLMP accurately characterizes the asymptotic outcome of the perceptual system (e.g. the probability for a particular response to occur), but has little to say about the time course of processing. The dynamic version of the FLMP is intended to address these reactions. In this version, the stimulus-to-feature mapping is assumed to take a certain amount of time. During this interval, the information about the stimulus gradually accumulates, and becomes increasingly accurate. It is assumed that accuracy increases as a negatively accelerated function of processing time, so that more information is gleaned early than late in the processing of the stimulus. One further assumption is that “integration of the separate features [is] updated continuously as the featural information is being evaluated. Similarly, decision [can] occur at any time after the stimulus presentation" (259). Thus, there is a partial temporal overlap between the different stages of processing, in the sense that one process can begin before a previous process is finished (see also Figure 2.1, p. 41).

These assumptions about the time course of information processing are supported by a number of experiments concerned with the effect of backward masking in the recognition of pure tones, and in the recognition of letters. Speech obviously raises a number of specific issues in that domain, however. Unlike written words, speech is a temporal phenomenon, it is continuous (i.e. there are no systematic acoustic boundaries between phonemes, syllables, or words) and, furthermore, time per se serves as a source of information in speech, as pointed out by Massaro (e.g. vowel duration is a major cue to the voicing of the following obstruent, to take but one example). However, few indications are given about how the model could be assessed in the speech domain (see remarks p. 194 and p. 263).

In addition to discussing the dynamics of processing, Massaro examines how the temporal relations between sources of information are dealt with in pattern recognition. Chapter 3 focuses on our sensitivity to temporal asynchronies between visible and audible speech. In the experiments reported in this chapter, bimodal CV syllables with various degrees of onset asynchrony between the auditory synthetic speech and the visible synthetic speech were presented to subjects in a forced-choice identification task. The results show that integration between the two sources of information still occurs when these sources of information are made asynchronous, provided that the time shift does not exceed a certain duration.

One major challenge for phoneticians and psycholinguists alike is to characterize the relationship between what could be called the *external* dynamics of speech, i.e. the temporal organization of the speech signal, and the *internal* time course of speech processing. Both play a role in the perception of speech, and it is most difficult to tell apart their respective influences on the listener's behaviour (Samuel, 1996). For example, in a gating study investigating the role of vowel duration as a cue to the voicing of the post-vocalic stop in CVC syllables, Warren and Marslen-Wilson (1988) found that the proportion of voiced-coda responses increased as the listeners were

presented with increasingly long portions of the initial CV sequence. One obvious interpretation is that longer vowels were perceived as being associated with voiced coda rather than voiceless ones. In keeping with Massaro's dynamical FLMP, however, it may also be assumed that evaluating the information provided by the vowel takes time, and that the evidence pointing to a voiced coda gradually accumulates as more processing time is made available to the listener, all other things being equal. Thus, the above finding raises the issue of how to differentiate the effect of vowel duration per se on the listener's response, from that of the internal dynamics of processing. Although this issue is not directly addressed in the book, there is no doubt that the FLMP would constitute a most appropriate framework for further investigations in this domain.

2.3 General Conclusion

This book provides us with quite an extensive review of the work carried out by the author and others on the use of multiple cues in speech perception and, more generally, pattern recognition. It is aimed at a very large audience, and constitutes a most useful tool both for teaching and research purposes. I do not doubt that it will soon become a major reference for researchers in phonetics, psycholinguistics, and cognitive psychology.

3 Bibliography

- Liberman, A.M. (1996) *Speech: A Special Code*. Cambridge, Mass.: MIT Press.
- McClelland, J.L. (1991) Stochastic interactive processes and the effect of context on perception, *Cognitive Psychology*, **23**, 1–44.
- Samuel, A.G. (1996) The role of time during lexical access, *Journal of the Acoustical Society of America*, **100**, 4/2, 2572.
- Stevens, K.N., and Blumstein, S.E. (1978) Invariant cues for place of articulation in stop consonants, *Journal of the Acoustical Society of America*, **64**, 1358–1368.
- Warren, P., and Marslen-Wilson, W. (1988) Cues to lexical choice - discriminating place and voice, *Perception and Psychophysics*, **43**, 21–30.