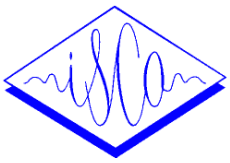

AVSP 2003 – International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4-7, 2003

Why the FLMP should not be applied to McGurk data
…or how to better compare models in the Bayesian framework

J.-L. Schwartz
Institut de la Communication Parlée, CNRS UMR 5009, INPG-Université Stendhal
INPG, 46 Av. Félix Viallet, 38031 Grenoble Cedex 1, France - schwartz@icp.inpg.fr

ABSTRACT

We come back on two criticisms addressed to the "Fuzzy-Logical Model of Perception" (FLMP): its assumed ability to fit any data, and the choice of a "fitting all" procedure rather than a "predicting AV from A and V" approach. We confirm the FLMP ability to fit random data in the "McGurk region", that is in all conditions including conflicting stimuli, thanks to the "0/0 trick". This is associated to a high instability of the *rmse* value in the optimal region. Then we introduce a Bayesian approach replacing fit by global likelihood. In this sounder framework, *rmse* instabilities decrease likelihood, hence the 0/0 trick is problematic for the FLMP. We conclude by indicating when the FLMP can be used in sound conditions: apart from the "McGurk" region.

1. Introduction

1.1. Twenty-five years of debate around the FLMP in the AV speech literature

Since the middle of the 70s, Dominic Massaro and his colleagues have consistently advocated the so-called "Fuzzy-Logical Model of Perception" (FLMP) as a pivot of a number of discussions about categorical perception in a first period, and, since the emergence of the famous "McGurk" paradigm [25], about audio-visual (AV) interactions and fusion in speech perception [11, 13]. In this field, the research by Massaro and colleagues about the FLMP has generated three main research achievements.

- A large number of experimental data dealing with vowels or consonants [6], natural or synthetic speech [15], adults or children [11], multilingual variability [20], attention [23], AV asynchrony [16], affect [21] etc.

- A contribution to a number of key questions about the fusion process, such as the independence vs. dependence, early vs. late, additive vs. multiplicative issues [11, 12].

- A systematic assessment campaign of quantitative models, comparing the FLMP with various other competitors, such as Trace [24], the TSD (Theory of Signal Detection), LIM (Linear Integration Model), CMP (Categorical Model of Perception), MDS (Multi-Dimensional Scaling), IAC (Interaction Activation and Competition), AMP (Additive Model of Perception), PM (Prelabelling Model, [1]): see [6, 13, 17, 22].

This last point is rare enough to deserve sincere admiration. Indeed, it is the objective of any model builder to compare his/her "baby" with those of concurrent researchers, and to attempt to demonstrate that it is indeed nicer (better) than others; but it is rare, if not unique, to find such a long-term continuity – the author of the present paper must admit that he himself did not exhibit such constancy and professionalism! But "la médaille a son revers", and the superiority of the FLMP, demonstrated over and over by Massaro and colleagues, raised some suspicion. More precisely, two major criticisms have been developed against the FLMP – and toughly discussed by Massaro.

1.2. The "fitting vs. prediction" argument

The original focus of the development of the FLMP was not AV fusion, but the integration of auditory cues in speech perception, and of multi-level cues in sentence processing. This has a profound technical consequence. In a problem such as assessing the combination of variable auditory cues in a phonetic identification task, e.g. F0 and VOT in the /zi/ vs. /si/ contrast [14], the values of the two cues are varied, but of course both of them are present all along the experiment: it is not possible to prepare a stimulus with no F0 or no VOT. This is exactly the same in the "G vs. Q" task [22], varying two geometric cues, both of them being present in the figure (obliqueness of the straight line and closedness of the gap in the oval). In consequence, they were led to *fit all the perceptual data at the same time*, and *compare the model predictions on this complete set of data*.

The situation is different with AV speech, where the experimenter can select one channel (A or V conditions) or both (AV condition). In consequence, researchers dealing with AV fusion for speech in noise [1, 4, 27, 28], AV perception by hearing-impaired [9] or the McGurk effect [2], generally felt that they should begin with tuning their model on unimodal data (pure A or V conditions) and then *assess or compare fusion models on their ability to predict AV performances from A-only or V-only data*.

The difference in strategy is understandable. But it is striking to notice that this difference happens to be crucial in comparison results. Indeed, all comparisons based on the "fitting A, V and AV at the

same time" procedure display a clear FLMP superiority [6, 12, 13, 15, 18]. On the other side, comparisons based on the "predicting AV from fitted A and V" procedure generally lead to FLMP performances somewhat lower than those of concurrent models [1, 9]. Hence, this methodological difference based on historical reasons became a theoretical topic! As a matter of fact, preferring the "fitting" or the "prediction" strategy could lead either to select or to reject FLMP as the race winner. The debate can be summarised by the two following quotations: "*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*" [32]; vs. "*To determine the truth of a theory, we believe that it is necessary to measure how well it accounts for the entire pattern of results, rather than how well some conditions predict the others (…); fitting the FLMP on the basis of the unimodal judgement makes the necessarily inaccurate assumption that the unimodal observations are noise free*" [18].

1.3. The "fitting everything " argument

The second attack is tightly connected to the previous quotation in [32], and consists in the suspicion that the "high flexibility" of the FLMP would enable it to predict everything. Very interestingly, Cutting et al. [7] attempted to provide a quantitative ground to this idea. They noticed that a model should not only be able to fit experimental data, but that it should also be "selective", and therefore provide a better fit of true data than of false ones. In consequence, they suggested that comparing two models should involve not only fitting experimental data, but also random ones. Using their own results about visual depth perception, they showed that the FLMP performed as well as an "additive model" on true data, but also slightly better on random data, and concluded for the superiority of the additive model, more selective to true data. Massaro & Cohen [17] rather convincingly replied that the difference between both models in fitting random data was quite low, and that the FLMP fit was much poorer on random than on true data (the Root Mean Square Error *rmse* was 8 times greater).

1.4. The topic of the present paper

A careful examination of the "fitting vs. predicting" and "fitting everything" arguments lead us to the view that they capture something very important about FLMP and about models in general, but in an incomplete and partly unsatisfactory way. This paper attempts to determine what exactly might be the "trick" in FLMP, and to propose a statistical framework in which model comparison could be performed in a more convincing way. More precisely, the present author feels that there is a specific experimental configuration (typically, McGurk) where the use of the FLMP might be mathematically unsound, and lead to mistaken interpretations of experimental data. The point is not to say that the FLMP systematically fits random data or solves the fitting problem unsatisfactorily, but that there is a specific methodological modelling problem with *conflicting audio-visual stimuli*, that is McGurk configurations. This will be presented in Section 2, together with an integrated Bayesian framework for comparing models. The objective, if the mathematical arguments are convincing enough, is to enable AV researchers to be aware of the difficulty, to escape possible interpretation mistakes, and to focus on the right questions about AV fusion, about the FLMP and about what Massaro calls the "paradigm". It is important to say at this point that some of these questions have been raised by Massaro himself, and some convincingly solved by the FLMP, as we shall recall in Section 3.

2. A technical analysis of the FLMP in the McGurk paradigm

2.1. FLMP and the 0 / 0 trick

The basic FLMP equation is:

$$p_{AV}(C_i) = p_A(C_i)p_V(C_i) / \sum_j p_A(C_j)p_V(C_j) \quad (1)$$

Ci and Cj being phonetic categories involved in the experiment, and p_A, p_V and p_{AV} the probability of responses respectively in the A, V and AV conditions. The superiority of the FLMP on other models is particularly significant in cases involving conflictual cues [6, 22]: typically, the McGurk situation with audio /b/ plus video /g/. In this situation, the unimodal responses are almost incompatible, and hence all phonetic categories involved in the pattern of responses display at least one very low value, either in the A modality, or in the V modality, or in both. The consequence is that all terms $p_A(C_i)p_V(C_i)$ are likely to be close to zero for all involved categories. Just to take an example, consider what would happen in an extreme situation with two phonetic classes C1 and C2, and a pair of A and V stimuli perfectly conflicting, that is with 100% of C1 responses with the A stimulus, and 100% of C2 responses with the V stimulus (see Table 1). Then, it is easy to show that any response in the AV modality, with a probability of C1 response equal to x and a probability of C2 response equal to (1−x) (x being any value between 0 and 1) can be fitted by the FLMP with a *rmse* exactly equal to 0. Indeed, suppose that the corresponding FLMP parameters for C1 and C2 are respectively set to $(1-\varepsilon_A)$ and ε_A in the A modality, and to ε_V and $(1-\varepsilon_V)$ in the V modality, with ε_A and ε_V two small values. Then the probability of C1 response in the AV condition is given in the FLMP by:

$$\varepsilon_V(1-\varepsilon_A)/[\varepsilon_V(1-\varepsilon_A)+\varepsilon_A(1-\varepsilon_V)] \cong \varepsilon_V / [\varepsilon_V + \varepsilon_A] \quad (2)$$

Hence, x may be perfectly fitted by choosing an $(\varepsilon_A, \varepsilon_V)$ pair such that:

$$\varepsilon_V /[\varepsilon_V + \varepsilon_A] = x \Leftrightarrow \varepsilon_V = \varepsilon_A \, x /(1-x) \quad (3)$$

Then, setting ε_A and ε_V at an arbitrarily low value, provided that they respect Eq. (3), allows a perfect fit (*rmse* equal to 0) to the pattern of experimental data in Table 1, whatever x.

	C1 responses		C2 responses	
	Data	FLMP	Data	FLMP
A cond.	1	$1-\varepsilon_A$	0	ε_A
V cond.	0	ε_V	1	$1-\varepsilon_V$
AV cond.	x	x	1−x	1−x

Table 1

Exactly the same can be done with a 3-classes situation more similar to the McGurk effect (both for fusions and combinations), displayed in Table 2. Indeed, fitting Table 2 data can be done with:

$$\varepsilon_V / [\varepsilon_V + \alpha\varepsilon_A + (1-\alpha)\varepsilon'_A] = x$$
$$\alpha\varepsilon_A / [\varepsilon_V + \alpha\varepsilon_A + (1-\alpha)\varepsilon'_A] = y \quad (4)$$

This set of two equations can once more be satisfied with arbitrary low values of ε_A, ε'_A and ε_V: hence any pattern of (α, x, y) values in Table 2 can be perfectly fitted by the FLMP, with a 0 *rmse*.

	C1 responses		C2 responses		C3 responses	
	Data	FLMP	Data	FLMP	Data	FLMP
A	1	$1-\varepsilon_A-\varepsilon'_A$	0	ε_A	0	ε'_A
V	0	ε_V	α	$\alpha-\varepsilon_V-\varepsilon'_V$	$1-\alpha$	$1-\alpha+\varepsilon'_V$
AV	x	x	y	y	1-x-y	1-x-y

Table 2

This is of course due to a very simple and well-known mathematical fact: 0/0 is an arbitrary value, or, to state this more precisely, lim(x/y) when $x\to0$ and $y\to0$, if it exists, may be any real value.

2.2. Application to real McGurk data

Of course, it could be argued that the previous section did not deal with real data, which seldom produce perfect zero values. However, the McGurk paradigm typically leads to very similar situations, since it deals with conflicting A and V stimuli. In a study of the McGurk effect in French [5], the pattern of responses to $[b_A]$, $[d_V]$, $[g_V]$, $[b_Ad_V]$ and $[b_Ag_V]$ for 126 French subjects, provided in Table 3, was judged surprising by us, since it showed that there were less [d] and more [b] responses to $[b_Ad_V]$ than to $[b_Ag_V]$. This pattern, coherent with most other published data, seems difficult to understand on the classical "$[b_A]$ is similar to $[d_A]$, $[g_V]$ is similar to $[d_V]$" basis. However, to our own surprise, the FLMP performed very well on these data, with an *rmse* of 0.0062[1]. But then, we had the idea to transform the pattern of AV data into arbitrary patterns of response, while keeping the pure A and V data of Table 3. In Table 4, we display the FLMP fits to hypothetical patterns of AV responses with $[b_Ad_V]$ and $[b_Ag_V]$ both perceived as mostly [b] (Test 1), mostly [d] (Test 2), mostly [g] (Test 3), $[b_Ad_V]$ perceived as [b] and $[b_Ag_V]$ as [d] (Test 4) or the inverse, $[b_Ad_V]$ as [d] and $[b_Ag_V]$ as [b] (Test 5). All the fits are equally good, and as good as the fit of true data.

responses	[b]	[d]	[g]	other
$[b_A]$	0.98	0	0	0.02
$[d_V]$	0.005	0.88	0.06	0.055
$[g_V]$	0	0.125	0.845	0.03
$[b_Ad_V]$	0.835	0.095	0	0.07
$[b_Ag_V]$	0.68	0.23	0.02	0.07

Table 3: McGurk data (from [5]) / *rmse* = 0.0062

		ans.[b]	ans.[d]	ans.[g]	*rmse*
Test 1	$[b_Ad_V]$	0.9	0.1	0	0.0049
	$[b_Ag_V]$	0.9	0.1	0	
Test 2	$[b_Ad_V]$	0.1	0.9	0	0.0053
	$[b_Ag_V]$	0.1	0.9	0	
Test 3	$[b_Ad_V]$	0.1	0	0.9	0.0061
	$[b_Ag_V]$	0.1	0	0.9	
Test 4	$[b_Ad_V]$	0.9	0.1	0	0.0082
	$[b_Ag_V]$	0.1	0.9	0	
Test 5	$[b_Ad_V]$	0.1	0.9	0	0.0047
	$[b_Ag_V]$	0.9	0.1	0	

Table 4: FLMP fit to various arbitrary AV patterns

This situation clearly corresponds to the Cutting et al.'s criticism against FLMP: it seems that in this McGurk context, FLMP is able to fit everything, even a random pattern of response. Why wasn't it the case in the random data used in [7]? The reason is that the "fitting noise" argument is not general. Consider the theoretical situation described in Table 5, in which we assume an experiment with two possible responses C1 and C2, and a visual pattern of responses completely ambiguous between C1 and C2. In this kind of configuration, the FLMP provides a good fit only if p_{AV} equals p_A (*rmse* = 0). As soon as p_{AV} differs from p_A by more than 0.02, *rmse* increases above 0.01. In this case, FLMP does not fit random data, but, on the contrary, it provides a strong prediction about what Massaro calls "optimal integration" ([12] p. 749).

	C1	C2
A	p_A	$1-p_A$
V	0.5	0.5
AV	p_{AV}	$1-p_{AV}$

Table 5: Testing "optimal integration" in FLMP

[1] The *rmse* in Cathiard et al. (2001) was slightly larger, because of the use of a threshold on the minimal acceptable probabilities, not used here, apart from the classical constraint that a probability is within [0-1].

So, it seems that the "fitting noise" argument is sometimes true (around "McGurk" configurations of conflicting stimuli) and sometimes not. By the way, let us come back on the "fitting vs. predicting" argument. The "fitting all" methodology recommended by Massaro seems the right one (in spite of our own choice in the past!): see [13], Ch. 10. However, what happens if we test the prediction strategy, using the A and V data in Table 3 to predict the AV data from Eq.(1)? The resulting *rmse* is extremely high: 0.1162! This shows that the FLMP ability to fit any pattern in this region has a severe drawback: the fit is highly unstable, hence the dramatic difference between the "fitting all" and the "prediction" strategies. Therefore, very small variations (+-0.01) to each FLMP parameter around the best fit to the experimental data in Table 3, lead to dramatic changes from the almost perfect value *rmse* = 0.0062 to values as high as 0.25! This is not the case in the stable configuration of Table 5.

2.3. Reconciling fitting and prediction, or fitting and selectivity: the statistical approach

So, where are we now? The "fitting noise" argument is sometimes true, sometimes not. The "fitting all" methodology seems to be the right one, but it may involve a large fit instability in the McGurk region. To deal with this complex pattern, let us recall where does fit come from.

Fit is derived from the logarithm of the maximum likelihood of a model, considering a data set. If D is a set of k data d_i, and M a model with parameters Θ, the estimation of the best θ values is provided by:

$$\theta = \text{argmax}\ p(\Theta|D,M) = \text{argmax}\ p(D|\Theta,M) \qquad (5)$$

and, if the model predicts that the d_i values come from Gaussian models (δ_i, σ_i), we have:

$$\log\,(p(D|\Theta,M)) = ct - 1/2 \sum_i (d_i-\delta_i)^2/\sigma_i^2 \qquad (6)$$

$$\log\,(p(D|\theta,M)) = ct - k/2\ \ rmse/\sigma^2 \qquad (7)$$

$$\text{if } \sigma_i^2=\sigma^2\ \ \forall i$$

Hence the θ parameters maximising the likelihood of M are those providing the best fit measured by *rmse*. But, in the Bayesian theory, the comparison of two models is more complex than the comparison of their best fit (Jaynes, [10]). Indeed:

$$p(D|M)=\int p(D,\Theta|M)d\Theta=\int p(D|\Theta,M)p(\Theta|M)d\Theta \qquad (8)$$

which means that the a priori distribution of data D knowing model M *integrates the distribution for all values Θ of the parameters of the model*. Let us consider two models M_1 and M_2 that have to be compared in relation to a data set D. The best fit θ_1 for model M_1 provides an *a posteriori* likelihood $\Lambda_1=\max p(\Theta_1|D,M_1)$ and the best fit θ_2 for model M_2 provides an *a posteriori* likelihood $\Lambda_2=\max p(\Theta_2|D,M_2)$. From Eq. (8) it follows that the model comparison is not provided by Λ_1/Λ_2 (or by comparing $rmse_1$ and $rmse_2$, as classically done), but by:

$$p(M_1|D) / p(M_2|D) = \Lambda_1 W_1 / \Lambda_2 W_2 \qquad (9)$$

with:

$$W_i = \int [p(D|\Theta_i,M_i)/p(D|\theta_i,M_i)]\ p(\Theta_i|M_i)d\Theta_i \qquad (10)$$

The term $p(D|\Theta_i,M_i)/p(D|\theta_i,M_i)$ evaluates the likelihood of Θ_i values relative to the best set θ_i providing the highest likelihood Λ_i for model M_i. Hence W_i evaluates the volume of Θ_i values providing an "acceptable" fit (not too far from the best one) relative to the whole volume of possible Θ_i values. This relative volume decreases with the increase of the total Θ_i volume: for example with the dimension of the Θ_i space[2]. But it also decreases if the function $p(D|\Theta_i,M_i)/p(D|\theta_i,M_i)$ decreases too quickly: this is what happens if the model is too sensitive, as is the FLMP around its best fit in the McGurk region.

2.4. Implementing the statistical approach

Let us consider a simple case, with two categories C1 and C2, and with the following pattern of data: $p_A(C1)=0.99$, $p_V(C1)=0.01$, and $p_{AV}(C1)=0.95$. Suppose that we try to compare two models on these data, that is the FLMP described by Eq. (1), and an "Audio Model" AM according to which the AV response would be equal to the pure A data, that is:

$$P_{AV}(Ci|AM) = P_A(Ci|AM) \qquad (11)$$

In the following, data are called p or q=1-p, and predictions are called P or Q=1-P. For a binomial law P, the probability to observe a distribution p on n samples is given by:

$$p(D|M) = C_n^{np}\ P^{np}\ Q^{nq} \qquad (12)$$

$$L=\text{Log}(p(D|M)) \cong n(p\text{Log}(P/p)+q\text{Log}(Q/q)) \qquad (13)$$

L is maximum for P=p, and around this value it varies as the opposite of *rmse*, in agreement with Eq. (7). The algorithm for computing the likelihood of each model is the following:

(a) Define a Θ space common to both models, e.g. $\Theta = P_A \otimes P_V$ with $P_A, P_V \in \{0.001 \ldots 0.999\}$

(b) For each $\{P_A,P_V\}$ pair, compute P_{AV} by Eq.(1) for FLMP, or Eq.(11) for the Audio model.

(c) For each $\{P_A,P_V\}$ pair, compute L_A, L_V and L_{AV} from Eq. (13), and the likelihood $\Lambda(P_A,P_V)$ by:

$$\Lambda(P_A,P_V) = \exp(L_A + L_V + L_{AV}) \qquad (14)$$

(d) Add all $\Lambda(P_A,P_V)$ values to obtain the total likelihood of M_1 and M_2 knowing the data D.

On Table 6, we provide both *rmse* and the global likelihood for both models. The conclusion is clear:

---

[2] Massaro proposes to apply a correction factor k/(k-f) to *rmse*, with k the number of data and f the freedom degree of the model ([13], p. 301).

The Audio Model provides a poorer *rmse* (remember that we are in a region where FLMP provides almost perfect fit to any AV value) but it displays a much larger likelihood. The interpretation is straightforward: the Audio Model makes a "stronger" prediction than FLMP on this set of data, which results in a larger volume of acceptable parameters in terms of relative fit, hence a larger likelihood. The lesson of this simple experiment is the following. Comparing models should not involve best fits, but global likelihood. Apart from the difference between a fit and a likelihood (the variances σ_i^2 in Eq. (7) are different from one value to another, since the variance of a binomial law is PQ/n), the less stable the fit is, the lower the likelihood. Jaynes provides an estimation of this contribution. It is outside the scope of the present paper to develop it in the context of AV models comparison, though this will be a major goal for the close future.

	rmse	*Likelihood*
FLMP	0.0122	0.0019
Audio Model	0.0283	0.0093

Table 6: Comparing FLMP and an "Audio Model" on the task of Section 2.4 (n=10)

3. Discussion

The first message in this paper is that fit is not the only truth, and likelihood should be preferred. The debate about fit vs. likelihood is not new ([26] vs. [19]), but the 0/0 trick may raise serious problems for the FLMP in this context. At least, in a comparison of models, the stability of *rmse* with small variations of the parameters around the best fit should be carefully studied, and introduced in some way in the evaluation procedure. Replacing fit by likelihood would probably result in contesting some of the "optimal integration" assumption, and it is likely that AV fusion could be shown to depend on languages [20, 30], attention [31], subjects, etc.

Far from the 0/0 region, the FLMP behaviour is sound. Hence its predictions about optimal integration are valid, and partly backed by experimental data. Of course, since it provides strong predictions, it is also in this region – and only there – that it can be falsified. Thus, the FLMP predicts that two ambiguous stimuli lead to an ambiguous fusion. This is the reason why the data on the "audiovisual VOT" [3] did provide "*an opportunity to illustrate that it is possible to falsify the FLMP prediction when the features are not, in fact, independent of one another*" ([12], pp. 787). Our own data on AV scene analysis as a contribution to perception in noise are another illustration of the non-independence of A and V processing [29].

The argument in this paper is basically *methodological*. It should clarify some technical points in previous debates, signal some bad uses of the FLMP in past studies, and possibly prevent some mistaken analyses in the future, particularly in the McGurk paradigm. However, methodology is just a step towards *understanding*, and a better understanding of AV representations and interactions in speech perception still needs the McGurk paradigm and conflicting stimuli as a key methodological tool [5, 8].

It remains that Massaro and his group have done a large contribution to the field, both experimental and theoretical. In the AV speech perception domain, quantitative models as the FLMP are a crucial piece of work, enabling researchers to think about data and processes: hence, the "Paradigm" stays alive and well, and provides a coherent framework to ask how speech is processed in the human brain.

Ackowledgement – This paper has benefited from many inspiring discussions with my colleagues or former colleagues Rafael Laboissière, Jordi Robert-Ribes & Pierre Bessière.

REFERENCES

[1] Braida, L.D. (1991). Crossmodal integration in the identification of consonant segments. *Quaterly J. Experimental Psychology, 43A*, 647-677.

[2] Braida, L.D., Sekiyama, K., & Dix, K. (1998). Integration of audiovisually compatible and incompatible consonants in identification experiments. *Proc AVSP98*.

[3] Breeuwer, M., & Plomp, R. (1986). Speechreading supplemented with auditorily presented speech parameters. *J. Acoust. Soc. Am., 79*, 481-499.

[4] Campbell, H.W. (1974). *Phoneme recognition by ear and by eye: a distinctive feature analysis*. Doctoral dissertation, Katholieke Universiteit te Nijmegen.

[5] Cathiard, M.A., Schwartz, J.L., & Abry, C. (2001). Asking a naive question to the McGurk effect : why does audio [b] give more [d] percepts with visual [g] than with visual [d] ? Proc. *AVSP'2001,* 138-142.

[6] Cohen, M.M., & Massaro, D.W. (1995). Perceiving visual and auditory information n consonant-vowel and vowel syllabes. In C. Sorin et al. (eds.) *Levels in Speech Communication* (pp. 25-38). Elsevier B.V.

[7] Cutting, J.E. et al. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgements of perceived depth. *J. Experimental Psychology: General, 121*, 364-381.

[8] Girin, L. (2003). Pure audio McGurk effect. *Proc. AVSP03*.
[9] Grant, K.W., Walden, B.E., & Seitz, P.F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition and auditory-visual integration. *J. Acoust. Soc. Am.*, 103, 2677-2690.
[10] Jaynes E.T. ; *Probability theory - The logic of science*. Cambridge University Press (in press). http://bayes.wustl.edu (1995).
[11] Massaro, D.W. (1987). *Speech perception by ear and eye: a paradigm for psychological inquiry*. London: Laurence Erlbaum Associates.
[12] Massaro, D.W. (1989). Multiple book review of speech perception by ear and eye: A paradigm for psychological inquiry. *Behavioral and Brain Sciences*, *12*, 741-794.
[13] Massaro, D.W. (1998). *Perceiving Talking Faces*. Cambridge: MIT Press.
[14] Massaro, D.W., & Cohen, M.M. (1976). The contribution of fundamental frequency and voice onset time to the /zi/-/si/ distinction. *J. Acoust. Soc. Am.*, *60*, 704-717.
[15] Massaro, D.W., & Cohen, M.M. (1990). Perception of synthesized audible and visible speech. *Psychological Science*, *1*, 55-63.
[16] Massaro, D.W., & Cohen, M.M. (1993). Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. *Speech Comm.*, *13*, 127-134.
[17] Massaro, D.W., & Cohen, M.M. (1993). The Paradigm and the Fuzzy Logical Model of Perception are alive and well. *JEP*, *122*, 115-124.
[18] Massaro, D.W., & Cohen, M.M. (1995). Modeling the perception of bimodal speech. *Proc. XIIIth ICPhS*, *3*, 106-113.
[19] Massaro, D.W., Cohen, M. M., Campbell, C.S., & Rodriguez, T. (2001). Bayes factor of model selection validates FLMP. *Psychonomic Bulletin & Review*, *8*, 1-17.
[20] Massaro, D.W., Cohen, M.M., Gesi, A., Heredia, R., & Tsuzaki, M. (1993). Bimodal speech perception: An examination across languages. *J Phon*, *21*, 445-478.
[21] Massaro, D.W., & Egan, P.B. (1996). Perceiving affect from the voice and face. *Psych Bull Rev*, *3*, 215-221.
[22] Massaro, D.W., & Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review*, *97*, 225-252.
[23] Massaro, D.W., & Warner, D.S. (1977). Dividing attention between auditory and visual perception. *Perception and Psychophysics*, *21*, 569-574.
[24] McClelland, J.L., & Elman, J.L. (1986). The TRACE model of speech perception. *Cog. Psych.*, *18*, 1-86.
[25] McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746-748.
[26] Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79-95.
[27] Petajan, E.D. (1984). *Automatic lipreading to enhance speech recognition*. Doc. Thesis, Univ. of Illinois.
[28] Robert-Ribes, J., Schwartz, J.L., & Escudier, P. (1995). A comparison of models for fusion of the auditory and visual sensors in speech perception. *Artificial Intelligence Review*, *9*, 323-346.
[29] Schwartz, J.L., Berthommier, F., & Savariaux, C. (2002). Audio-visual scene analysis: Evidence for a "very-early" integration process in audio-visual speech perception. Proc. *ICSLP'2002*, 1937-1940.
[30] Sekiyama, K., & Tokhura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *J. Phonetics*, *21*, 427-444.
[31] Tiippana, K., Sams, M., & Andersen, T.S. (2001). Visual attention influences audiovisual speech perception. Proc. *AVSP'2001*, 167-171.
[32] Vroomen, J., & de Gelder, B. (2000). Crossmodal integration : a good fit is no criterion. *Trends in Cognitive Sciences*, *4*, 37-38.