A reanalysis of McGurk data suggests that audiovisual fusion in speech perception is subject-dependent

Jean-Luc Schwartz

A reanalysis of McGurk data suggests that audiovisual fusion in speech perception is subject-dependent

Jean-Luc Schwartz

(Jean-Luc.Schwartz@gipsa-lab.inpg.fr)

GIPSA-Lab, Speech and Cognition Department / Institut de la Communication Parlée,

UMR 5216, CNRS – Grenoble University – France

Suggested running title: Audiovisual fusion is subject dependent

ABSTRACT

Audiovisual perception of conflicting stimuli displays a large level of intersubject variability, generally larger than pure auditory or visual data. However, it is not clear whether this actually reflects differences in integration per se, or just the consequence of slight differences in unisensory perception. It is argued that the debate has been blurred by methodological problems in the analysis of experimental data, particularly when using the Fuzzy-Logical Model of Perception (Massaro, 1987) shown to display overfitting abilities with McGurk stimuli (Schwartz, 2006). A large corpus of McGurk data is reanalyzed, using a methodology based on (1) comparison of FLMP and a variant with subject-dependent weights of the auditory and visual inputs in the fusion process, WFLMP, (2) use of a Bayesian Selection Model criterion instead of a Root Mean Square Error fit in model assessment, (3) systematic exploration of the number of useful parameters in the models to compare, attempting to discard poorly explicative parameters. It is shown that WFLMP performs significantly better than FLMP, suggesting that audiovisual fusion is indeed subject-dependent, some subjects being more "auditory" and others more "visual". Inter-subject variability has important consequences for theoretical understanding of the fusion process, and reeducation of hearing impaired people.

Suggested PACS Classification numbers

Main section: 43.71

Detailed classification: 43.71.An, 43.71.Ma

Keywords: audiovisual fusion – fusion models – McGurk effect – interindividual variability

I. INTRODUCTION

When a public demonstration of the McGurk effect (McGurk & MacDonald, 1976) is presented to visitors or students, there appears a large variability in the subjects' audiovisual (AV) responses, some seeming focused on the auditory (A) input, others more sensitive to the visual (V) component and to the McGurk illusion. The existence of possible differences in fusion would have important consequences in both theoretical and practical terms. However, it stays hotly debated, considering that subjects could actually differ in pure auditory and visual performance rather than in fusion per se. In the following, the major elements of discussion and disagreement will be reviewed. Then it will be suggested that the debate has been largely blurred by methodological problems. A way out of these problems will be proposed, which will constitute the core of the present paper. The objective is actually twofold: present a methodological framework for analysis of audiovisual speech perception data, and show that this framework confirms that there are indeed inter-individual differences in the fusion process.

A. Inter-individual differences in audiovisual fusion in the speech perception literature

The possibility that subjects could put more or less "weight" on the auditory or visual inputs is certainly not new. It was, for example, the focus of a paper by Seewald et al. (1985), suggesting that there was a "primary modality for speech perception", either auditory or visual. This hypothesis has received less attention since the work by Massaro and colleagues in the framework of the development of the "Fuzzy-Logical Model of Perception" (FLMP). Indeed, a central assumption of the model is that, apart from possible differences in auditory or visual perception, the fusion mechanism *per se* is exactly the same for all subjects (Massaro, 1987, 1998). The mechanism, actually a multiplicative process applied to fuzzy-

logical levels of confidence provided by audition and vision on all possible answers, is considered as being an optimal process, in the sense that “all sources contribute to a decision but more ambiguous sources are given less of a say in the decision” (Massaro, 1998, p. 115).

A repeated claim by Massaro and colleagues is hence that all subjects are “optimal integrators” and combine auditory and visual evidence for available categories all exactly in the same multiplicative way. Any difference in the output of the audiovisual speech perception process would be only due to differences in auditory and visual processing and unisensory category tuning between subjects.

The hypothesis of a universal and optimal fusion mechanism remains controversial, and was the object of a series of experimental and modeling work by Grant & Seitz (1998), who claimed that, even when unimodal skill levels are taken into account, large differences in individuals’ AV recognition scores persist which “might be attributable to differing efficiency in the operation of a perceptual process that *integrates* auditory and visual speech information” (p. 2438). The debate further continued between Massaro & Cohen (2000) and Grant (2002). Actually, the question of possible inter-subject differences in AV integration, apart from being theoretically challenging, has important potential practical applications. Indeed, if some subjects integrate the audio and visual information less efficiently than others, the focus in a rehabilitation process (in case of hearing impairment for example) should be put on the training of integration, rather than just the training of auditory or visual abilities (Grant & Seitz, 1998). Incidentally, Rouger et al. (2007) claim to have found that cochlear implanted subjects are better at integrating the sound and face of a speaker’s utterances than normal hearing subjects.

In the last 15 years, a number of studies have shown substantial individual variability in AV speech recognition. Sekiyama and Tokhura (1991) showed that McGurk fusion illusions were reduced in Japanese compared with English participants. Since then, several studies

have investigated comparative language effects for audio-visual speech integration: English vs. Japanese (Sekiyama and Tokhura, 1993; Kuhl et al., 1994), English vs. Japanese vs. Spanish (Massaro et al., 1993), Spanish vs. German (Fuster-Duran, 1995), or German vs. Hungarian (Grassegger, 1995). A number of differences have been reported. Some of them come from the nature of the stimuli, differing from language to language. However, there remain differences between linguistic groups perceiving the same stimuli. Sekiyama and Tokhura (1993) claim that they reflect variations in the *weight* different linguistic communities would attribute to the visual input in the integration process. They suggest that the Japanese community could make less use of the visual input because of a cultural difference, namely that “it may be regarded as impolite in Japan to look at someone’s face” (pp. 442). On the other hand, Massaro et al. (1993) and Kuhl et al. (1994) interpret these differences as coming from variations in the inventory of linguistic prototypes rather than from social or cultural variations in the tuning of the audio-visual process. Indeed, Massaro et al. (1993) showed that their own data displaying different audiovisual perception of conflicting AV stimuli by English, Spanish and Japanese subjects, could be perfectly fitted by FLMP. In the FLMP fit, the differences between English, Spanish and Japanese subjects cannot be due to differences in fusion: they are totally due to differences in the unimodal categorization responses.

More recently, Sekiyama et al. (2003) showed that the very early ability to fuse auditory and visual inputs, displayed by a McGurk effect appearing as soon as 4 months in infants' speech perception (Burnham & Dodd, 1996, 2004; Rosenblum et al., 1997), was followed by a developmental evolution of AV fusion after 6 years, and largely between 6 and 8 (Sekiyama & Burnham, 2004) for English children. This increase could be the result of a learning process, and it seems to be blocked in Japanese children, hence resulting in the smaller role of the visual input in AV perception previously described. Once again, however, it could be

argued that this developmental pattern is just indicative of a development of *unimodal* auditory and visual categories rather than of integration *per se*. Basically, children (and particularly English ones) would be progressively more and more accurate in their perception of visual categories, hence the increase in AV performance. In this reasoning, fusion would stay perfectly stable whatever the age, i.e. multiplicative and optimal, in Massaro's sense.

Finally, gender differences in audiovisual fusion have been suggested in various papers, female subjects presenting a higher level of audiovisual performance and a greater level of visual influence on auditory speech (Irwin et al., 2006; Strelnikov et al., 2008), possibly linked to differences in the cortical networks involved, with less left lateralization in females compared with males (Pugh et al., 1996; Jaeger et al., 1998).

B. A methodological caveat: the 0/0 problem in FLMP testing

In a recent paper, Schwartz (2006) displayed a severe technical problem in the comparison of FLMP with other models when using corpora containing McGurk data. Indeed, in the case of conflicting inputs, the audio and visual stimuli provide at least one quasi-null probability in each possible category, and the multiplicative process implied by the FLMP leads to AV predictions equal to 0/0, which is indeterminate. Therefore, any audio-visual response can be fitted by the FLMP. The consequence is double. Firstly, since FLMP may predict any pattern of response in the McGurk case, fitting McGurk data with FLMP cannot help determine if variation in AV perception is actually due to differences in unimodal behavior or in AV fusion. Secondly, the over-fitting ability of the FLMP with discrepant A and V stimuli might well contaminate the global Root Mean Squared Error (RMSE) criterion systematically used when FLMP is compared with other models. For these reasons, it seems more appropriate to use a Bayesian Model Selection (BMS) criterion, which intrinsically

accounts for any over-fitting problem (MacKay, 1992; Pitt & Myung, 2002; Schwartz, 2006).

In this context, the present paper aims at reconsidering the invariant vs. subject-dependent audiovisual fusion problem, in a sound BMS framework. A classical test corpus of audiovisual consonant-vowel stimuli extensively studied by Massaro and colleagues (Massaro, 1998) will provide a basis for assessing possible discrepancies in audiovisual fusion between subjects, independent of any linguistic or developmental effect. For this aim, WFLMP, a variant of FLMP explicitly incorporating subject-dependent weights of the audio and visual inputs in integration, will be compared with FLMP. This will provide the opportunity to use both BMS and RMSE criteria on these two models. This will also lead to a principled methodology for comparing audiovisual speech perception models on a given set of data. This methodology uses a so-called Laplace approximation of BMS, called BMSL, together with a systematic assessment of the number of really useful parameters in the models to compare, relying on the BMS ability to deal with variations in the number of degrees of freedom in these models.

Section II will recall the experimental material and provide a detailed description of the proposed methodology, together with the models to compare, and the assessment criteria. In Section III, the obtained results will be presented, before a discussion in Section IV.

II. METHODOLOGY

A. Experimental material: The UCSC corpus of CV audiovisual discrepant stimuli

The corpus considered here has been extensively used for comparing audiovisual fusion models in speech perception (Massaro, 1998). This corpus crosses a synthetic five-level audio /ba/-/da/ continuum with a synthetic video similar continuum. The 10 unimodal (5A, 5V) and 25 bimodal (AV) stimuli were presented for /ba/ vs. /da/ identification to 82

subjects, with 24 observations per subject. The responses are kindly made available by Massaro and colleagues on their web site (http://mambo.ucsc.edu/ps1/8236/).

B. Model comparison

1. RMSE and corrected RMSE

Let us consider a given speech perception experiment consisting in the categorization of speech stimuli involving n_E experimental conditions Ej, and in each condition, n_C possible responses corresponding to different phonetic categories Ci. In most papers comparing models in the field of speech perception, the tool used to compare models is the "fit" estimated by the "root mean square error" *RMSE*, computed by taking the squared distances between observed and predicted probabilities of responses, averaging them over all categories Ci and all experimental conditions Ej, and taking the square root of the result:

$$RMSE = \left[\left(\sum_{Ej,\,Ci} \left(P_{Ej}\left(C_i\right) - p_{Ej}\left(C_i\right)\right)^2\right) / \left(n_E\ n_C\right)\right]^{1/2} \qquad (1)$$

(observed probabilities are in lower case and predicted probabilities in upper case throughout this paper).

Considering that two models M_A and M_B might differ in their number of degrees of freedom, Massaro (1998) proposes to apply a correction factor k/(k-f) to *RMSE*, with k the number of data and f the number of degrees of freedom of the model (p. 301). This provides a second criterion:

$$RMSEcor = k/(k\text{-}f)\left[\left(\sum_{Ej,\,Ci} \left(P_{Ej}\left(C_i\right) - p_{Ej}\left(C_i\right)\right)^2\right) / \left(n_E\ n_C\right)\right]^{1/2} \qquad (2)$$

2. BMSL

If **D** is a set of k data d_i, and M a model with parameters Θ, the fit may be derived from the logarithm of the *maximum likelihood of the model* considering the data set, that is the value

of Θ maximizing $L(\Theta|M) = p(\mathbf{D}|\Theta,M)$. However, comparing two models by comparing their best fits means that there is a first step of estimation of these best fits, and it must be acknowledged that the estimation process is not error-free. Therefore, the comparison must account for this error-prone process, which is done in Bayesian Model Selection by computing the total likelihood of the model knowing the data. This results in integrating likelihood over all model parameter values. Taking the opposite of the logarithm of total likelihood leads to the so-called "Bayesian Model Selection" (BMS) criterion that should be minimized for model evaluation (MacKay, 1992, Pitt & Myung, 2002) [(1)]:

$$BMS = -\log \int L(\Theta|M)\, p(\Theta|M)\, d\Theta \qquad (3)$$

The computation of BMS through Eq. (3) is complex. It involves the estimation of an integral, which generally requires use of numerical integration techniques, typically Monte-Carlo methods (e.g. Gilks et al., 1996). However, Jaynes (1995, ch. 24) proposes an approximation of the total likelihood in Eq. (9), based on an expansion of log(L) around the maximum likelihood point θ.

$$\mathrm{Log}(L(\Theta)) \cong \mathrm{Log}(L(\theta)) + 1/2\, (\Theta-\theta)' \left[\partial^2 \log(L) / \partial \Theta^2\right]_\theta (\Theta-\theta) \qquad (4)$$

where $[\partial^2 \log(L) / \partial \Theta^2]_\theta$ is the Hessian matrix of the function log(L) computed at the position of the parameter set θ providing the maximal likelihood L_{max} of the considered model. This leads to the so-called Laplace approximation of the BMS criterion (Kass & Raftery, 1995):

$$BMSL = -\log(L_{max}) - m/2 \log(2\pi) + \log(V) - 1/2 \log(\det(\Sigma)) \qquad (5)$$

where V is the total volume of the space occupied by parameters Θ, m is its dimension, that is the number of free parameters in the considered model, and Σ is defined by:

$$\Sigma^{-1} = -\left[\partial^2 \log(L) / \partial \Theta^2\right]_\theta \qquad (6)$$

The preferred model considering the data **D** should *minimize* the *BMSL* criterion. There are in fact three kinds of terms in Eq. (5). Firstly, the term $-\log(L_{max})$ is directly linked to the maximum likelihood of the model, more or less accurately estimated by *RMSE* in Eq. (1): the larger the maximum likelihood, the smaller the *BMSL* criterion. Then, the two following terms are linked to the dimensionality and volume of the considered model. Altogether, they result in handicapping models that are too "large" (that is, models with a too high number of free parameters) by increasing $BMSL^{(2)}$. Finally, the fourth term provides a term favoring models with a large value of $\det(\boldsymbol{\Sigma})$. Indeed, if $\det(\boldsymbol{\Sigma})$ is large, the determinant of the Hessian matrix of log(L) is small, which expresses the fact that the likelihood L does not vary too quickly around its maximum value L_{max}. This means that the fit provided by the model around its maximum likelihood point is stable: exactly the contrary of FLMP with McGurk data, since its overfitting abilities result in very rapid modifications of the prediction even for very small changes in the unimodal values, making the integration process quite unstable and over-sensitive to the tuning of free parameters in the model. Derivation of the exact formula in (5), together with a practical implementation of *BMSL*, can be found in http://www.icp.inpg.fr/~schwartz/fichiers_pdf/BMSL_tutorial.pdf.

Bayesian Model Selection has already been applied to the comparison of AV speech perception models, including FLMP (see Myung & Pitt, 1997; Massaro et al., 2001; Pitt et al., 2003). However, this involved heavy computations of integrals in Eq. (3) through Monte Carlo techniques, which would be difficult to apply systematically in model comparisons. *BMSL* has the advantage of being easy to compute, and interpret in terms of fit and stability. Furthermore, if the amount of available data is much higher than the number of parameters involved in the models to compare (that is, the dimension m of the $\boldsymbol{\Theta}$ space) the probability distributions become highly peaked around their maxima, and the central limit theorem shows that the approximation in Eqs. (4-5) becomes quite reasonable (Walker, 1967). Kass

& Raftery (1995) suggest that the approximation should work well for a sample size greater than 20 times the parameter size m (see Slate, 1999, for further discussions about assessing non-normality).

3. Estimating the "true" number of degrees of freedom in a model

The number of model parameters in most model comparison studies in AV speech perception is generally kept fixed to the "natural number of degrees of freedom" of the model, that is the number of free parameters necessary to implement the model in its most extensive definition. Care is generally taken to check that the models have basically the same number of degrees of freedom, otherwise the *RMSE* correction described previously could be applied. Notice that this correction looses some sense if a parameter is introduced with no effect on the model likelihood (a "useless parameter") while *BMSL* naturally discards useless parameters through the integration in formula (3).

Of course, completely useless parameters generally do not exist, since this would correspond to some kind of misconception of the model. However, it is important to assess the possibility that some parameters are not really useful in the model behavior. For example, while all model comparisons generally involve a subject-by-subject assessment – and it will also be the case here – it could be interesting to test if some parameters could not in fact be similar from one subject to the other. The same could be done from one experimental condition to the other. Therefore, various implementations of the models to compare will be systematically tested, with a progressively increasing number of fixed parameters and thus a decreasing number of free parameters, in order to attempt to determine the *true* number of degrees of freedom of the model, that is the number of free parameters really useful, and providing the highest global likelihood of the model knowing the data. Our

basic assumption is that it is under the condition of true number of degree of freedom that models can be really assessed and compared in sound conditions.

Decreasing the number of free parameters raises two problems. Firstly, the parameters to fix must be adequately selected. This may be done on a statistical objective basis, for example through Principal Component Analysis techniques, but this results in combinations of parameters difficult to interpret. A heuristic approach was preferred in which the observation of experimental data guided the selection of possible parameters to be kept fixed from one subject to another. The second problem is to estimate the value of the parameters being kept fixed. This was done through a Round Robin technique, in which a given parameter for one subject is estimated from the mean value taken by the parameter in the whole corpus excluding the current subject from the computation. This technique, classical and computationally simple, prevents from any artefactual introduction of the current data to model inside the "fixed" parameter used to model the data in a circular approach, which would be inappropriate.

C. Models

Two models were compared, FLMP and a variant with weighted contribution of the auditory and visual inputs in the integration, WFLMP. For each corpus, each model (including the variants associated with the decrease in the number of degrees of freedom) was fitted to the data separately for each subject. This enabled us to compute both mean values of the selected criteria, averaged over all subjects, and to assess differences between models by applying Wilcoxon signed-rank tests over the compared criteria for each subject.

1. FLMP

In a speech perception task consisting in the categorization of auditory, visual and audio-visual stimuli, the FLMP may be defined as a Bayesian fusion model with independence between modalities, and the basic FLMP equation is:

$$P_{AV}(C_i) = P_A(C_i)P_V(C_i) / \sum_j P_A(C_j)P_V(C_j) \quad (7)$$

Ci and Cj being phonetic categories involved in the experiment, and P_A, P_V and P_{AV} the model probability of responses respectively in the A, V and AV conditions.

2. WFLMP

The weighted FLMP model, called WFLMP, is defined by Eq. (8):

$$P_{AV}(C_i) = {P_A}^{\lambda_A}(C_i)\ {P_V}^{\lambda_V}(C_i) / \sum_j {P_A}^{\lambda_A}(C_j)\ {P_V}^{\lambda_V}(C_j) \quad (8)$$

where λ_A and λ_V are subject-dependent factors used to weight the A and V inputs in the computation of the audiovisual responses estimated by $P_{AV}(C_i)$ (see other introductions of weights inside FLMP in Schwarzer & Massaro, 2001; or, for a similar kind of weighted fusion model applied to speech recognition, in various implementations since Adjoudani & Benoît, 1996: see a review in Teissier et al., 1999). For each subject, a lambda value is defined between 0 and 1, and λ_A and λ_V are computed from lambda by: λ_A = lambda / (1 – lambda) and λ_V = (1 – lambda) / lambda, with thresholds maintaining λ_A and λ_V between 0 and 1. Figure 1 shows how lambda controls the weights λ_A and λ_V and how this results in varying P_{AV} from a value close to P_A when lambda is close to 0, to a value close to P_V when lambda is close to 1, passing by a value identical to the FLMP prediction when lambda is set at 0.5, with λ_A and λ_V both equal to 1.

FIG. 1

III. RESULTS

A. Analysis of individual experimental data

The UCSC corpus has been extensively used in AV speech perception model assessment, generally with a good fit using the FLMP and RMSE criterion (Massaro, 1998; see also Massaro et al., 2001, for an assessment of FLMP with a *BMS* criterion on this corpus). However, looking at the data, there seems to appear an effect not predicted by the FLMP, that is inter-individual differences in AV interaction. This is displayed in Fig. 2, showing two subjects with very close auditory and visual performances, though with quite different audiovisual responses. It seems that the weight of the visual modality is respectively high for the first one (Fig. 2a) and low for the second one (Fig. 2b). Though the FLMP does not incorporate A and V weights, the fit is however quite acceptable (with *RMSE* values respectively 0.04 and 0.02 for these two subjects). This good fit is actually obtained because of the 0/0 instability: indeed, the FLMP simulation of unimodal data for the first subject is drawn towards slightly more ambiguous values for A responses and less ambiguous values for V responses (see Fig. 2a), while the inverse is done for the second subject (see Fig. 2b). This is the indirect way the FLMP may decrease the importance of a modality in fusion, by slightly but consistently misfitting the unimodal data without introducing subject-specific weights, and while keeping a very low *RMSE* value (a very good fit) because of the 0/0 problem (Schwartz, 2006). Such consistent misfits of unimodal data, if they happen in a significant number of cases, would indicate a problem in modeling. They should be taken into account in a BMS criterion, though they are almost undetectable in a RMSE criterion.

FIG. 2

B. Selected degrees of freedom for FLMP and WFLMP

The first implementation of FLMP needs 10 parameters for each subject, that is 5 values $Ai=P_{Ai}$(/da/) and 5 values $Vj=P_{Vj}$ (/da/) for the 5 stimuli of each continuum. Since the WFLMP model needs one more parameter per subject, one parameter was removed by fixing the value of the parameter A5 (audio response for the fifth audio stimuli, higher than 0.99 in average) at a value equal to the mean of the value it takes for the other subjects.

To explain how the number of parameters was decreased, on Fig. 3 a sample of auditory and visual identification curves is displayed for 10 of the 82 subjects. In the audio results (Fig. 3a), the curves are all S-shaped from a value close to 0 to a value close to 1, with less variation on the sides (for A5, A1 and to a lesser extent A4 and A2). Therefore, it was attempted to fix these parameters, in this order, with the Round Robin procedure. In the visual curves (Fig. 3b) the configuration is different, and suggests that it should be possible to describe these curves by estimating some values by a linear regression prediction on logit values of Vi, that is log(Vi/(1-Vi)). For this aim, two linear regression predictions on logit values were defined, one predicting V2 and V4 from the parameters V1, V3 and V5, and the other predicting V2, V4 and V5 from the parameters V1 and V3. Altogether, this lead to five variants of the FLMP and WFLMP models respectively with 10, 6, 5, 4 and 3 free parameters per subject (Table 1) [(3)].

FIG. 3

Table 1

C. Modeling results

Figure 4 shows the results for the two models with their 5 free-parameter variants. For each case, means and standard deviations computed on the modeling results for the 82 subjects are presented.

FIG. 4

With 10 parameters per subject, the FLMP fit is good, with an average *RMSE* value at 0.051 (see Massaro, 1998, p. 64). Interestingly, the fit is significantly better for the WFLMP with the same number of free parameters, with an average *RMSE* value at 0.0445 (since N=82 is higher than 20, z-ratios are used following a unit normal distribution, $u=4.92$, $p<0.001$). *RMSE* then logically increases when the number of parameters decreases (Fig. 4a). The portrait for *RMSEcor* is the same (Fig. 4b). However, *BMSL* reaches a minimum for 6 parameters, both for FLMP and WFLMP (Fig. 4c). In this variant of both models, A1 and A5 are fixed, together with A4 for WFLMP. V2 and V4 are estimated from V1, V3 and V5 by logit linear regression. For this optimal six-parameter implementation, there is a significant gain of WFLMP over FLMP ($u=4.77$, $p<0.001$).

In Figure 5, the histogram of logarithms of estimated lambda values are plotted for all subjects for WFLMP with 6 parameters. It appears that the range is indeed large, with auditory subjects on the right of the 1-value, and visual subjects on the left. Under a criterion of λ_A / λ_V values respectively higher than 1.5 or lower than 0.67 (1/1.5), there are 33 «audio» and 14 «visual» subjects, the remaining 35 being intermediary.

FIG. 5

Figure 6 shows how WFLMP models the two subjects compared in Fig. 2. Typically, the auditory and visual fits are similar between subjects – as in the experimental data themselves – while the good fit of the differences between subjects in audiovisual values is due to large differences in the lambda values, as shown on Fig. 5. This confirms that subject (a) is rather "visual" (with a λ_A / λ_V ratio at 0.23) and subject (b) is rather "auditory" (with a λ_A / λ_V ratio at 3.89), as suggested by the data themselves.

FIG. 6

IV. GENERAL DISCUSSION

Two topics are addressed in the present work. One concerns methodology for model comparison, which is of particular importance in audiovisual fusion, as evidenced by the very large number of controversies regularly arising in the domain. This also has implications for designing models for audiovisual fusion in speech perception. The second one concerns the invariant vs. subject-dependent nature of audiovisual fusion and more generally the parameters able to intervene in fusion. These topics will be addressed one after the other.

A. An adequate methodology for comparing models

There are two important claims in our methodological approach. Firstly, a local criterion such as RMSE, or its quasi-equivalent Maximum Likelihood, can be inappropriate, particularly in cases involving models which have a tendency to overfit the data, e.g. with FLMP and

McGurk data. A global BMS criterion is theoretically sounder, as has been discussed in a large number of papers, unfortunately not much acknowledged in the speech perception community. The local approximation provided by BMSL is simple to compute, easy to interpret, and efficient assuming that the number of experimental data points are sufficient. In the present paper it appears that RMSE and BMSL converge on showing the superiority of WFLMP over FLMP for the McGurk data. But this is not systematically the case (see e.g. Schwartz & Cathiard, 2004; Schwartz, 2006). Therefore, we suggest that any model comparison involving FLMP based on an RMSE criterion should be taken cautiously, and its conclusions should be considered as probably arguable unless a new evaluation based on Bayesian Model Selection is undertaken.

Secondly, varying the number of parameters in model assessment is very important. Of course, the difficulty is that there is much freedom in the strategies that can be proposed for this principled reduction. This should involve both a simple and intuitive approach, and a substantial number of variants to be able to assess the approximate number of really meaningful parameters for comparing models.

In fact, these two claims are related. Decreasing the number of parameters *forces one* to use a criterion able to take the size of the model parameter set into account. This is the case of BMS and its BMSL approximation, and *not* the case of RMSE, in which any correction is largely arbitrary. Variation in the number of parameters showed that there was indeed some redundancy in the 10 free parameters per subject involved in FLMP for the present corpus, 6 appearing as a more plausible number of degrees of freedom (3 for the visual input, 3 for the audio input). It is interesting to note that a reduced number of degrees of freedom provided a much larger difference in favor of WFLMP, compared with the complete set of 10 parameters.

B. Audio-visual fusion models

The present paper is focussed on FLMP for methodological reasons associated to its very frequent use in publications, and its excessive adaptability to McGurk data leading in several cases to inappropriate or mistaken analyses of experimental results. However, FLMP is actually neither weakened nor strengthened by the present paper.

It is not weakened since we proposed both an adequate method for testing it in safer conditions – through the BMS framework – and a possible variant with subject-specific modulation – WFLMP – which could provide the route for new developments more in line with evidence that fusion is subject dependent. Massaro and colleagues already introduced both ingredients in some of their work, but the present paper shows that they are actually *required* in any further use of FLMP in speech perception, particularly (but not exclusively) in experiments involving incongruent stimuli, as in the McGurk effect.

It is not strengthened either, since the present analysis could have been applied to other fusion models, such as Braida's Pre- and Post-Labeling models (Braida, 1991; see also Grant & Seitz, 1998; Grant, 2002), with quite probably the same conclusions. Actually, a number of models have been recently developed explicitly taking into account the possibility that one modality could be favored in the fusion process. This is the point addressed by Ernst and colleagues with their Maximum Likelihood framework according to which integration would be "optimal" in leading to the largest possible reduction in the variance of the multisensory output. This is achieved by adaptively weighting modalities in relation to their reliability (or variance) for the considered task (Ernst & Banks, 2002; Ernst & Bülthoff, 2004; Deneve & Pouget, 2004; Körding et al., 2007; and a precursor use of this concept in audiovisual speech perception models in Robert-Ribes et al., 1995). This "optimal integration" view is different from the optimal Bayesian fusion of decisions implemented in FLMP, since it occurs at a pre-categorical level.

C. What drives audiovisual fusion in speech perception?

The major theoretical output of the present paper is that it clearly shows that fusion is subject-dependent. There are indeed large inter-subject differences in audiovisual fusion for speech perception, with various groups of subjects, some being more “auditory”, others more “visual”. Many papers mention such a large variability in audiovisual performance. However, it was always unclear whether this was due to differences in unisensory performance, or multisensory integration. The present analysis strongly reinforces the second view.

This opens the route to a number of questions about the fusion mechanism itself. Differences between subjects in the McGurk paradigm could result from a general “orientation” of a given subject towards one or the other modality for individual reasons (specific or related to e.g. culture, language, sex or age). They could also be the consequence of properties of the task or the experimental situation, which could have driven the subject towards one rather than the other stimulus input in a bimodal task.

Inter-individual factors

It could well be the case that some subjects rely more on audition and others more on vision, and that they weight audiovisual fusion accordingly (see Giard & Peronnet, 1999). Hence, it could be assumed that there is for a given subject a general trend to favor one modality over another one, whatever the task. This should result in future studies comparing audiovisual fusion in various speech and non-speech tasks, searching for individual portraits stable from one task to the other. These different behaviors could also be associated to differences in neuroimagery experiments, in terms of the involved cortical networks, and of the quantitative role of each part in the global portrait.

We have already discussed in the Introduction section possible factors likely to play a role in sensor fusion: some languages could use the visual input more than others (e.g.

English more than Japanese), female subjects could use it more than males, adults more than children. Notice that the methodology employed here could be used to re-analyze all data relevant for these claims, in order to carefully disentangle the role of unimodal and multimodal factors in the corresponding studies.

This opens an important question, which is to know to what extent the weighting process can be dynamically modified during the subject's life. We have already mentioned the developmental evolution leading to an increase in the role of the visual input (see e.g. Sekiyama et al., 2003; Sekiyama & Burnham, 2004). Recent data by Schorr et al. (2005) suggest that there is a critical period for the development of audiovisual fusion in speech perception, before 2.5 years. In the case of a perturbation of one or the other modality, related to age or handicap, the question becomes to know if a subject can, voluntarily or through any reeducation means, selectively reinforce the weight of the most efficient modality.

In a recent study, Rouger et al. (2007) claim that this could indeed be the case, hearing impaired subjects equipped with cochlear implants displaying, in their terms, "a greater capacity to integrate visual input with the distorted speech signal" (p. 7295). Actually, these data should be considered with caution, being a possible case of unimodal effects interpreted as bimodal. Indeed, Rouger et al. compare three populations of subjects: hearing impaired subjects equipped with cochlear implants (CI), normal hearing subjects with audition degraded by noise (NHN) and normal hearing subjects presented with noise-band vocoder speech degrading audition in a way supposed to mimic the cochlear implant (NHV). They show that for a similar level of audio performance, the audiovisual recognition is larger for CI than for NHV, NHN being in the middle. Two factors could be, in their view, responsible for this pattern: differences in the visual performance, and in integration per se. A modeling approach leads them to claim that while the global visual scores are actually better in CI compared with NHN and NHV, there would be an additional gain in CH compared with

NHV, hence the claim about a "greater capacity to integrate visual input with the distorted speech signal". Notice that integration efficiency would be as high in NHN as in CH according to their analysis. However, careful inspection of auditory confusion matrices for NHV and NHN, available in Rouger (2007), shows that the structure of these matrices is quite different. Importantly, the transmission of the voicing mode was poorer in speech degraded with noise-band vocoder (NHV) than with white noise (NHN), suggesting that there could be a poorer complementarity in the audio and visual inputs in NHV, logically resulting in lower audiovisual scores. Differences in audiovisual performance would hence result from the *structure of the unimodal inputs* (being less complementary for normal hearing subjects presented with noise-band vocoder speech) rather than from integration per se. Actually, in this case, a study based on WFLMP and BMS would probably *not* reveal any discrepancy in integration between impaired subjects equipped with cochlear implants (CI) and normal hearing subjects (NHV and NHN).

Intra-individual factors

Finally, it is quite possible that the weight of one modality depends on the experimental situation per se. Firstly, stimuli themselves could possibly drive the weighting factor. In a review of intersensory interactions, Welch & Warren (1986) proposed "Modality Precision" or "Modality Appropriateness" as a key factor in explaining which modality should dominate intersensory judgements. Evidence for the role of reliability in audiovisual fusion for speech perception can be found in the study by Lisker & Rossi (1992) on the auditory, visual and audiovisual identification of vocalic rounding. Careful inspection of their data shows that though auditory identification seems in some cases quite accurate, there is a systematic trend for putting more weight in the visual modality within audiovisual data, as if the subjects "knew" that their eye was better than their ear at this particular task. Conversely, Robert-

Ribes et al. (1998) on their study of audiovisual vowel perception in noise report that with a very high level of noise, some subjects consistently select a given response (e.g. [ɔ]) for all vowels in noise in the auditory modality, which could lead in a model as FLMP to a large probability of response of this category in the audiovisual modality. However, this is not the case, showing that subjects know that the auditory modality is not reliable at high levels of noise, and hence discard it almost completely from the fusion process.

Secondly, the experimental conditions could lead to enhance or decrease the role of one modality in the fusion process. Attentional mechanisms should play a role at this level. Actually, while it had been initially claimed since McGurk & MacDonald (1976) that the McGurk effect was automatic and not under the control of attention, it appeared later that the instruction to attend more to audition or to vision might bias the response (Massaro, 1998). A recent set of experiments by Tiippana et al. (2004) showed that if the attention is distracted from the visual flow, the role of the visual input seems to decrease in fusion, with less McGurk effect. Notice that the authors themselves attempted to simulate their data with FLMP, and argued that the good fit of their data by a model claiming that fusion is automatic appeared as “a paradox” (p. 458). Actually, reanalysis of their data in a BMS framework with a weighted FLMP suggests that there are indeed attentional factors intervening in fusion itself, independently of unimodal effects (Schwartz & Tiippana, in preparation). This is confirmed by a number of recent experiments showing the possibility to modulate the McGurk effect based on manipulations of attention (e.g. Alsius et al., 2005, 2007), though here again, a precise analysis of experimental results in a BMS framework could provide an adequate control for disentangling unimodal from multimodal factors.

V. CONCLUSION

The present work proposed a new methodology for comparatively assessing models of audiovisual speech perception. This methodology is based on both the use of a Bayesian Model Selection criterion approximated in a computationally simple way (BMSL), and on a systematic variation in the number of degrees of freedom of the models to assess, in order to reveal the "true" number of parameters in a given model for a given task. The comparison of FLMP with a variant with auditory and visual weights varying from one subject to another (WFLMP) lead to the conclusion that weights are indeed variable, and hence that audiovisual integration seems subject-dependent.

This could have important consequences in future studies about audiovisual speech perception. Firstly, from a methodological point of view, it suggests that studies on audiovisual speech perception should consider these differences and possibly separate experimental groups into "auditory" or "visual" sub-groups on the basis of such criteria as McGurk performance. Secondly, from an audiological point of view, this indicates that subjects should be assessed on the basis of their audiovisual fusion abilities, and considered differently – in terms of reeducation and practice – depending on whether they are more "auditory" or more "visual" in their behavior.

Footnotes

1. In the following, bold symbols deal with vectors or matrices, and all optimizations are computed on the model parameter set Θ.

2. The interpretation of the term log(V) is straightforward, and results in handicapping large models by increasing BMSL. The term – m/2 log(2π) comes more indirectly from the analysis, and could seem to *favor large models*. In fact, it can only *decrease the trend to favor small models* over large ones.

3. It could seem paradoxical to maintain the number of free parameters similar for each subject, while attempting to show inter-individual differences. This is not the case actually. The principle is to freeze as much as possible the structure of the model for all subjects, in order to let the differences appear in an objective way.

References

Adjoudani, A., & Benoît, C. (**1996**). "On the integration of auditory and visual parameters in an HMM-based ASR," in *Speechreading by humans and machines*, edited by D.G. Stork & M.E. Hennecke (Berlin: Springer-Verlag), pp. 461-472.

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (**2005**). "Audiovisual integration of speech falters under high attention demands," Curr Biol. **15**, 839-43.

Alsius, A., Navarra, J., & Soto-Faraco, S. (**2007**). "Attention to touch weakens audiovisual speech integration," Exp Brain Res. **183**, 399-404.

Braida, L. D. (**1991**). "Crossmodal integration in the identification of consonant segments," Q. J. Exp. Psychol. **43**, 647– 677.

Burnham, D., & Dodd, B. (**1996**). "Auditory-visual speech perception as a direct process: The McGurk effect in human infants and across languages," in *Speechreading by humans and machines*, edited by D.G. Stork & M.E. Hennecke (Berlin: Springer-Verlag), pp. 103-114.

Burnham, D. & Dodd, B. (**2004**). "Auditory-visual speech integration by pre-linguistic infants: Perception of an emergent consonant in the McGurk effect," Developmental Psychobiology **44**, 209-220.

Denève, S., & Pouget, A. (**2004**). "Bayesian multisensory integration and cross-modal spatial links," Journal of Neurophysiology (Paris) **98**, 249-258.

Ernst, M.O., & Banks, M.S. (**2002**). "Humans integrate visual and haptic information in a statistically optimal fashion," Nature **415**, 429–433.

Ernst, M. O., & Bulthoff, H. H. (**2004**). "Merging the senses into a robust percept," Trends Cogn. Sci **8**, 162-169.

Fuster-Duran, A. (**1995**). "McGurk effect in Spanish and German listeners. Influences of visual cues in the perception of Spanish and German conflicting audio-visual stimuli," *Proceedings of Eurospeech 95*, 295-298.

Giard, M.H, & Peronnet, F. (**1999**). "Auditory-visual integration during multi-modal object recognition in humans: a behavioral and electrophysiological study," J Cogn Neurosci **11**, 473--490.

Gilks, W.R., Richardson, S., & Spiegelhalter, D.J. (**1996**). *Markov Chain Monte Carlo in Practice*. New-York: Chapman & Hall.

Grant, K.W. (**2002**). "Measures of auditory-visual integration for speech understanding: A theoretical perspective (L) ," J. Acoust. Soc. Am. **112**, 30-33.

Grant, K.W. & Seitz, P.F. (**1998**). "Measures of auditory-visual integration in nonsense syllables and sentences," J. Acoust. Soc. Am. **104**, 2438-2449.

Grassegger, H. (**1995**). "McGurk effect in German and Hungarian listeners," *Proceedings of the XIIIth Int. Cong. of Phon. Sciences*, 210-113.

Irwin, J.R., Whalen, D.H., & Fowler, C.A. (**2006**). "A sex difference in visual influence on heard speech," Perception & Psychophysics **68**, 582-592.

Jaeger, J., Lockwood, A., Van Valin, R. D. Jr,, Kemmerer, D. L., Murphy, B. W., & Wack, D. S. (**1998**). "Sex differences in brain regions activated by grammatical and reading tasks," **Neuroreport** 9, 2803-2807.

Jaynes E.T. (**1995**). *Probability theory - The logic of science*. Cambridge University Press. http://bayes.wustl.edu.

Kass, R.E., & Raftery, A.E. (**1995**). "Bayes factor," Journal of the American Statistical Association **90**, 773-795.

Körding, K.P., Beierholm, U., Ma, W., Quartz, S., Tenenbaum, J., Shams, L., (**2007**). "Causal Inference in Cue Combination," PLOSOne **2**, e943. doi:10.1371/journal.pone.0000943.

Kuhl, P.K., Tsuzaki, M. Tohkura, Y., & Meltzoff, A.N. (**1994**). "Human processing of auditory-visual information in speech perception: Potential for multimodal human-machine interfaces," *Proc. Int. Conf. on Spoken Language Processing* (Yokohama, Japan), 539-542.

Lisker, L. & Rossi, M. (**1992**). "Auditory and visual cueing of the [± rounded] feature of vowels," Language and Speech **35**, 391-417.

MacKay, D.J.C. (**1992**). "Bayesian interpolation," Neural Computation **4**, 415-447.

Massaro, D.W. (**1987**). *Speech perception by ear and eye: a paradigm for psychological inquiry*. London: Laurence Erlbaum Associates.

Massaro, D.W. (**1998**). *Perceiving Talking Faces*. Cambridge: MIT Press.

Massaro, D.W., & Cohen, M.M. (**2000**). "Tests of auditory-visual integration efficiency within the framework of the fuzzy-logical model of perception," J. Acoust. Soc. Am. **108**, 784-789.

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.

Massaro, D.W., Cohen, M.M., Gesi, A., Heredia, R., & Tsuzaki, M. (**1993**). "Bimodal speech perception: An examination across languages," J. Phonetics **21**, 445-478.

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

Myung, I. J., & Pitt, M. A. (**1997**). "Applying Occam's razor in modeling cognition: A Bayesian approach, " Psychonomic Bulletin & Review **4**, 79-95.

Pitt, M.A., Kim, W., & Myung, I.J. (**2003**). "Flexibility versus generalizablity in model selection.," Psychonomic Bulletin & Review **10**, 29-44.

Pitt, M.A., & Myung, I.J. (**2002**). "When a good fit can be bad," Trends in Cognitive Science **6**, 421-425.

Pugh, K.R., Shaywitz, B.A., Shaiwitz, S.E., Fulbright, R.K., Byrd, D., Skudlarski, P., Shankweiler, D.P., Katz, L., Constable, R. T., Fletcher, J., Lacadie, C., Marchione, K., & Gore, J. C. (**1996**). "Auditory selective attention: An fMRI investigation," NeuroImage **4**, 159-173.

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 Journal **9**, 323-346.

Robert-Ribes, J., Schwartz, J.-L., Lallouache, T. & Escudier, P. (**1998**). "Complementarity and synergy in bimodal speech: Auditory, visual and audio-visual identification of French oral vowels in noise," J. Acoust. Soc. Am. **103**, 3677-3689.

Rosenblum, L.D., Schmuckler, M.A., & Johnson, J.A. (**1997**). "The McGurk effect in infants," Perception & Psychophysics **59**, 347-357.

Rouger, J. (**2007**). "Perception audiovisuelle de la parole chez le sourd postlingual implanté cochléaire et le sujet normo-entendant: étude longitudinale psychophysique et neurofonctionnelle," PhD Université Paul Sabatier, Toulouse.

Rouger, J ., Lagleyre, S., Fraysse, B., Deneve, S., Deguine, O., & Barone, P. (**2007**). "Evidence that cochlear-implanted deaf patients are better multisensory integrators," Proc Natl Acad Sci USA **104**, 7295-7300.

Schorr, E.A., Fox, N.A., van Wassenhove, V., & Knudsen, E.I. (**2005**). "Auditory-visual fusion in speech perception in children with cochlear implants," Proc Natl Acad Sci USA **102**, 18748–18750.

Schwartz, J.L. (**2006**). "Bayesian model selection: The 0/0 problem in the Fuzzy-Logical Model of Perception," J. Acoust. Soc. Am., **120**, 1795-1798.

Schwartz, J.L., Berthommier, F., & Savariaux, C. (**2004**). "Seeing to hear better: Evidence for early audio-visual interactions in speech identification," Cognition **93**, B69–B78.

Schwartz, J.L., & Cathiard, M.A. (**2004**). "Modeling audio-visual speech perception. Back on fusion architectures and fusion control," Proc. *ICSLP'2004*, Jeju, Korea, pp. 2017-2020.

Schwartz, J.L., Robert-Ribes, J., & Escudier, P. (**1998**). "Ten years after Summerfield ... a taxonomy of models for audiovisual fusion in speech perception," in R. Campbell, B. Dodd & D. Burnham (eds.) Hearing by eye, II. Perspectives and directions in research on audiovisual aspects of language processing (pp. 85-108). Hove (UK) : Psychology Press.

Schwarzer, G., & Massaro, D.W. (**2001**). "Modeling face identificaton processing in children and adults," Journal of Experimental Child Psychology **79**, 139-161.

Seewald, R., Ross, M., Giolas, T., & Yonovitz, A. (**1985**). "Primary modality for speech perception in children with normal and impaired hearing," Journal of Speech and Hearing Research **28**, 36–46.

Sekiyama, K., Burnham, D., Tam, H., & Erdener, D. (**2003**). "Auditory-Visual Speech Perception Development in Japanese and English Speakers," in *Proceedings of the International Conference Audio–Visual Speech Processing 2003*, edited by J.-L. Schwartz, F. Berthommier, M.-A. Cathiard & D. Sodoyer (St.Jorioz, France), pp. 43-47.

Sekiyama. K. & Burnham, D. (**2004**). "Issues in the development of auditory-visual speech perception: Adults, infants and children," in *Proceedings of the 8th International Conference on Spoken Language Processing*, edited by Soon Hyob Kim & Dae Hee Yuon (Seoul, Sunjin Printing, Korea), pp 1137-40

Sekiyama, K., & Tokhura, Y. (**1991**). "McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility," J. Acoust. Soc. Am. **90**, 1797-1825.

Sekiyama, K., & Tokhura, Y. (**1993**). "Inter-language differences in the influence of visual cues in speech perception," J Phon **21**, 427-444.

Slate, E.H. (**1999**). "Assessing multivariate nonnormality using univariate distributions," Biometrika **86**, 191-202.

Strelnikov, K, Rouger, .J., Lagleyre, S., Fraysse, B., Deguine, O., & Barone, P. (**2008**). "Improvement in speech-reading ability by auditory training: evidence from gender differences in normally-hearing, deaf, cochlear implanted subjects," to appear in Neuropsychologia.

Teissier, P., Robert-Ribes, J., Schwartz, J.L., & Guérin-Dugué, A. (1999). Comparing models for audiovisual fusion in a noisy-vowel recognition task. IEEE Trans. Speech and Audio Processing 7, 629-642.

Tiippana, K., Andersen, T.S., & Sams, M., (2004) Visual attention modulates audiovisual speech perception, European Journal of Cognitive Psychology, 16 457-472.

Tables

Number of parameters per subject	Parameters for FLMP	Parameters for WFLMP
10	V1..5 A1..5	+ lambda A5 fixed
6	V1, V3, V5 A2, A3, A4 V2, V4 estimated by linear regression A1, A5 fixed	+ lambda A4 fixed
5	V1, V3 A2, A3, A4 V2, V4, V5 estimated by linear regression A1, A5 fixed	+ lambda A4 fixed
4	V1, V3 A2, A3 V2, V4, V5 estimated by linear regression A1, A4, A5 fixed	+ lambda A2 fixed
3	V1, V3 A3 V2, V4, V5 estimated by linear regression A1, A2, A4, A5 fixed	+ lambda V1 fixed

Table 1 – ***The five variants of FLMP and WFLMP.***
All fixed parameters are estimated by the Round Robin technique

Figure captions

Figure 1 – Variations of weighting coefficients λ_A and λ_V (left) and predicted p_{AV} (right) as a function of the lambda parameter tuning fusion in WFLMP. When lambda is close to 0, the audio weight decreases towards zero, the video weight increases towards one and the modeled p_{AV} reaches a value close to p_V. Conversely, when lambda is close to 1, the audio weight increases towards one, the video weight decreases towards zero and the modeled p_{AV} reaches a value close to p_A. Notice that for a lambda value at 0.5, both audio and video weights are set to one, which provides exactly the FLMP predictions. In this example, p_V is set to 0.2 and p_A to 0.8.

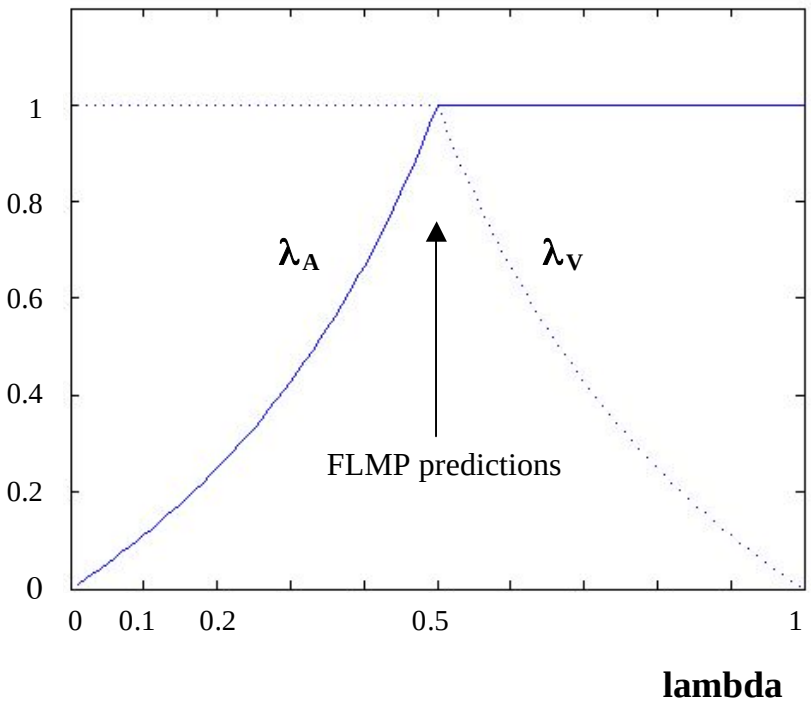
Figure 2 – (a) Audio (top left), visual (top right) and audiovisual (bottom) data for subject 3 in UCSC corpus: data in solid lines, FLMP predictions in dotted lines; (b) Same as Fig. 2a for subject 18 in UCSC corpus.

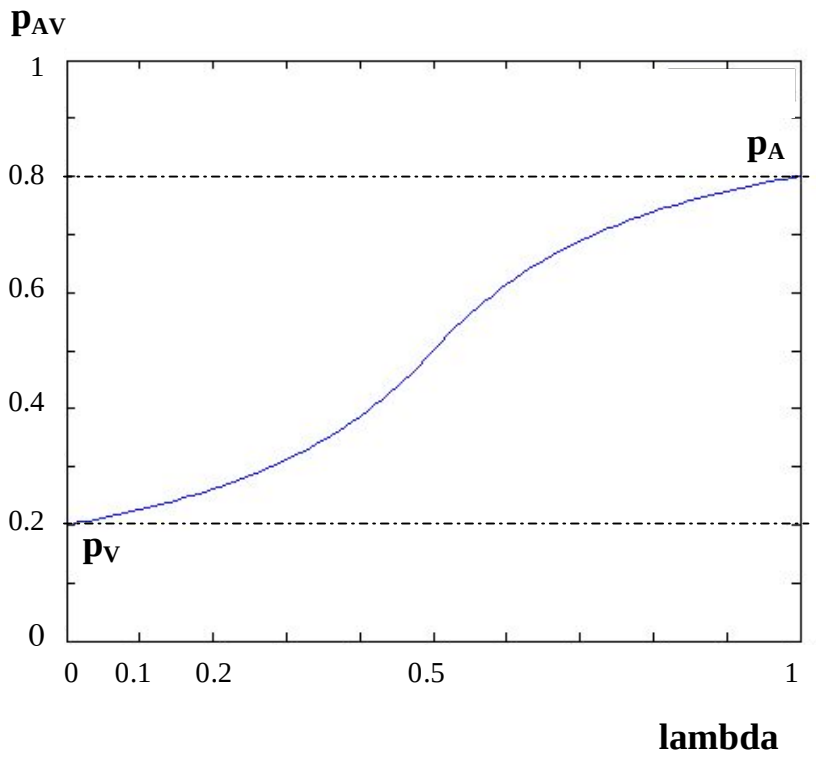
Figure 3 – left: audio identification for 10 subjects / right: visual identification for 10 subjects.

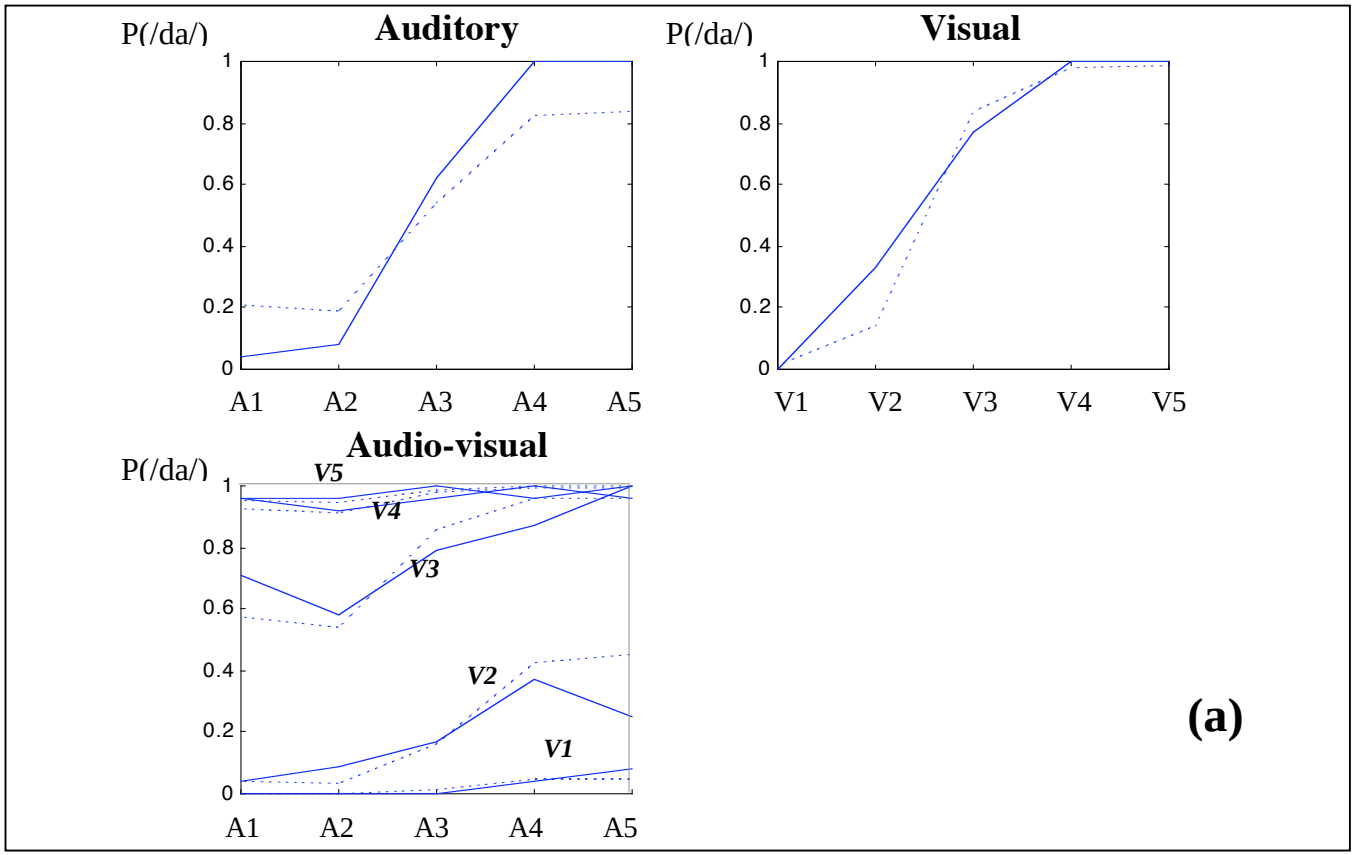
Figure 4 – (a) Compared RMSE values for FLMP vs. WFLMP simulations with 3 to 10 free parameters per subject; (b) Compared corrected RMSE values for FLMP vs. WFLMP simulations with 3 to 10 free parameters per subject; (c) Compared BMSL values for FLMP vs. WFLMP simulations with 3 to 10 free parameters per subject.

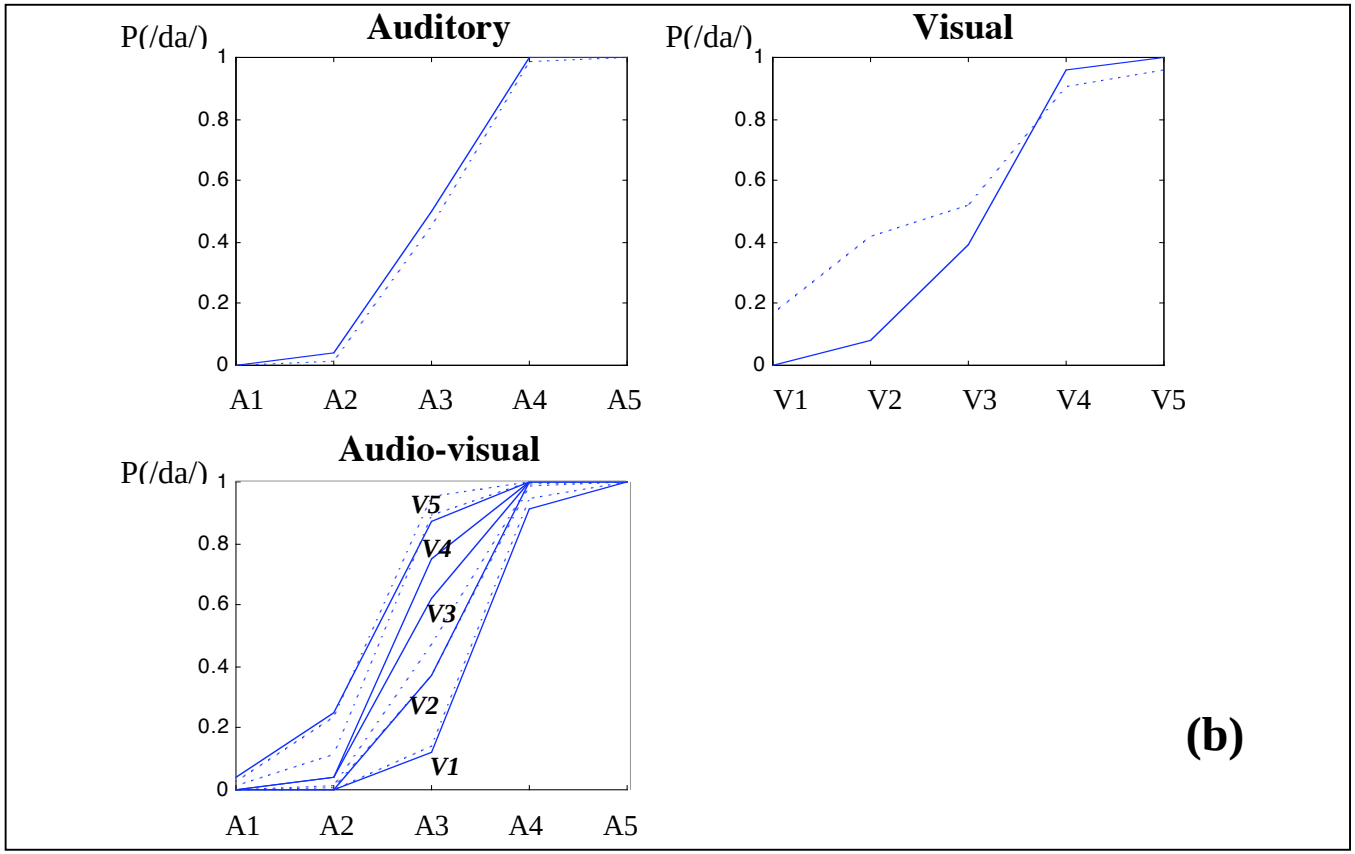
Figure 5 – Histogram of $\lambda_A/\ \lambda_V$ values controlling the fusion process for the 82 subjects in the WFLMP model with 6 parameters. Values for subject 3 (data displayed in Fig. 1a) and subject 18 (data displayed in Fig. 1b) are superimposed on the figure.

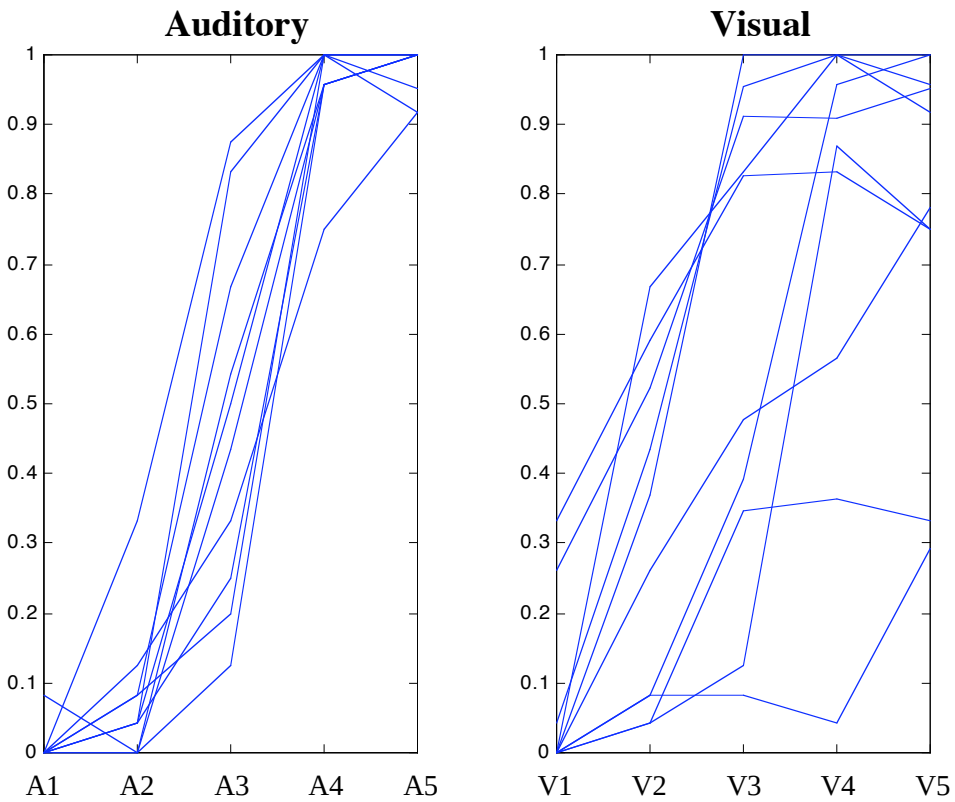
Figure 6 – (a) Audio (top left), visual (top right) and audiovisual (bottom) data for subject 3 in UCSC corpus: data in solid lines, predictions with WFLMP with 6 parameters in dotted lines; (b) Same as Fig. 6a for subject 18 in UCSC corpus.

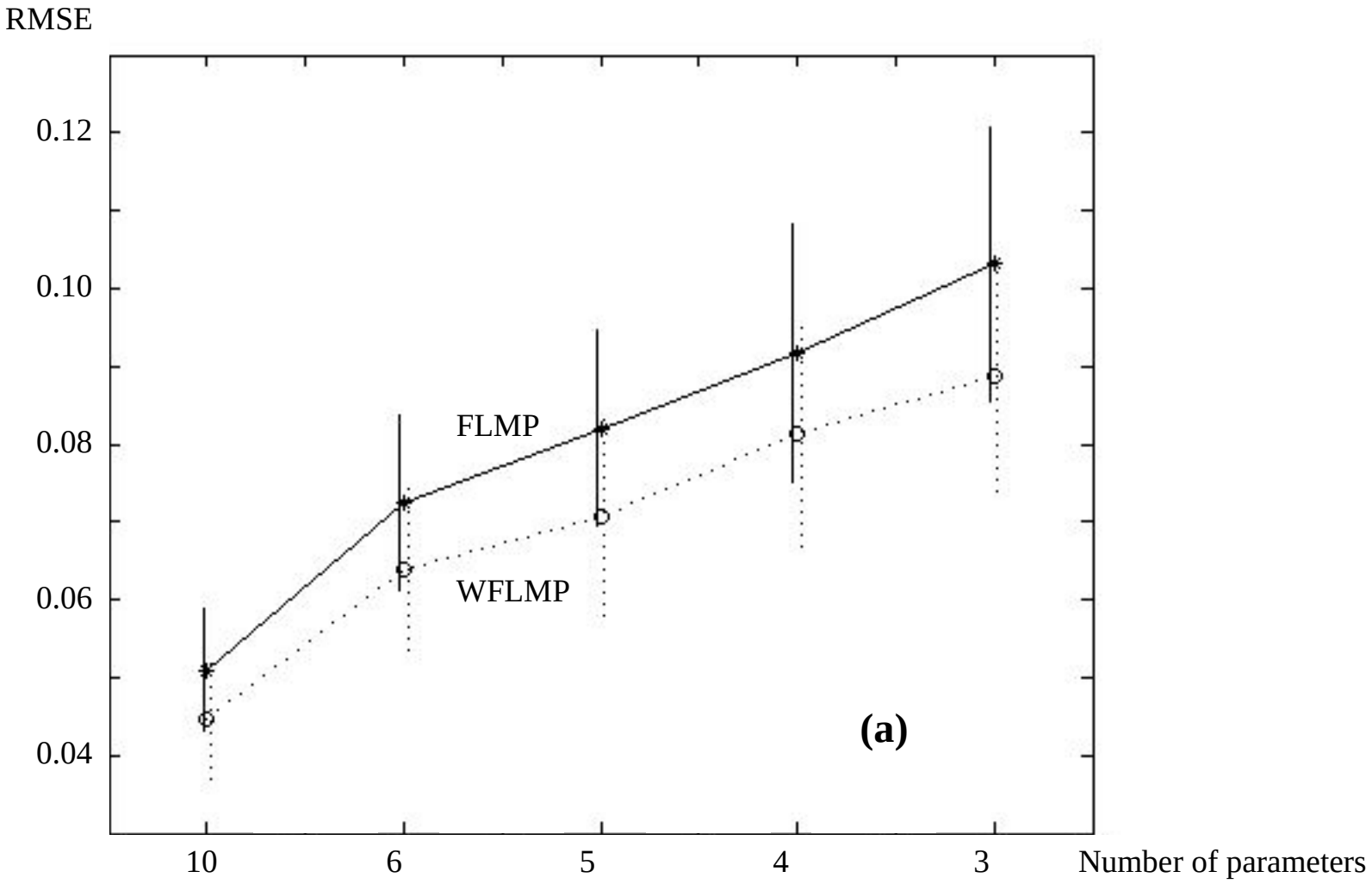

1
0.8
0.6
0.4
0.2
0
0
0.1
0.2
0.5
1
λA
λV
FLMP predictions
lambda

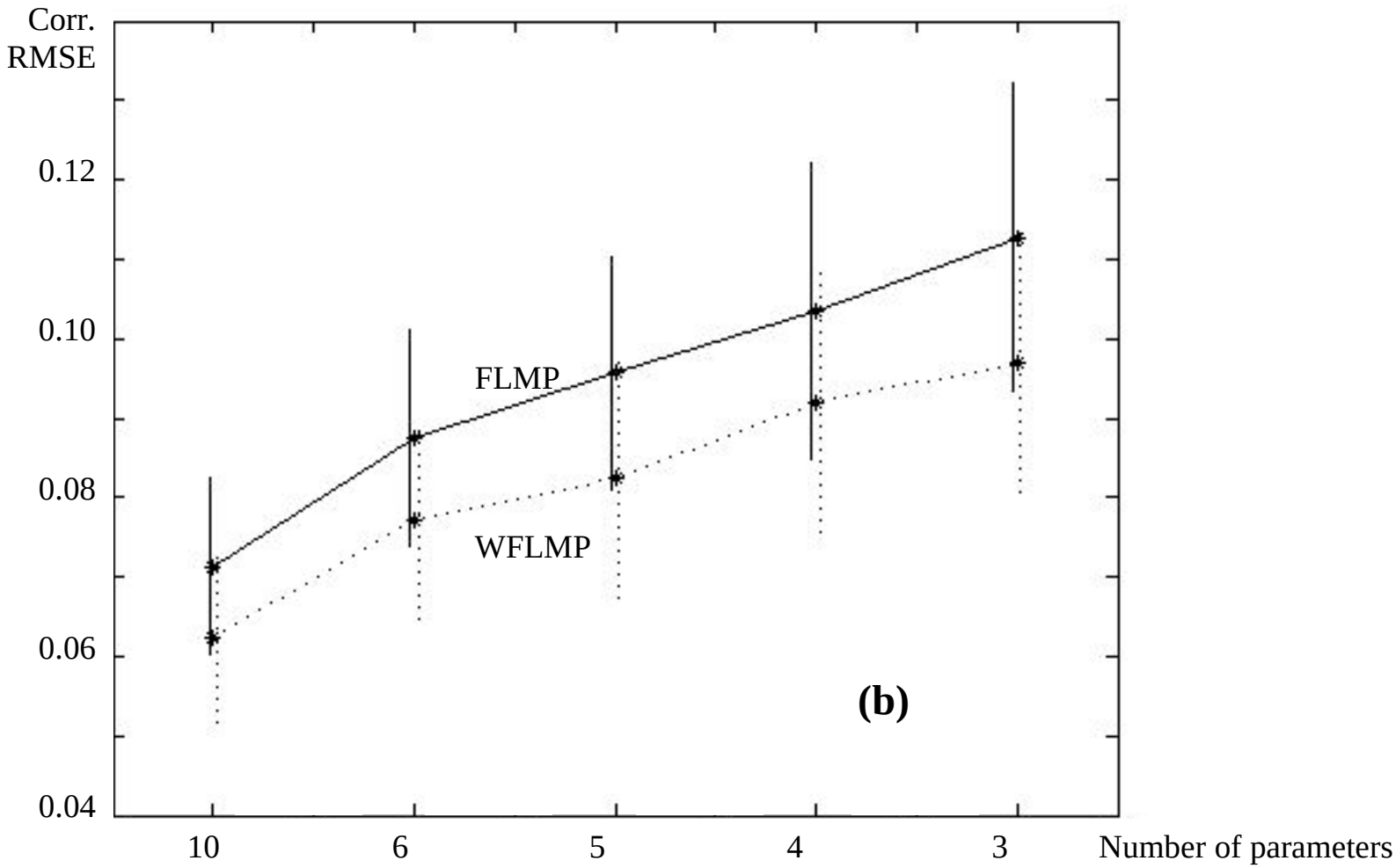

pAV
1
0.8
0.6
0.4
0.2
0
0
0.1
0.2
0.5
1
pA
pV
lambda

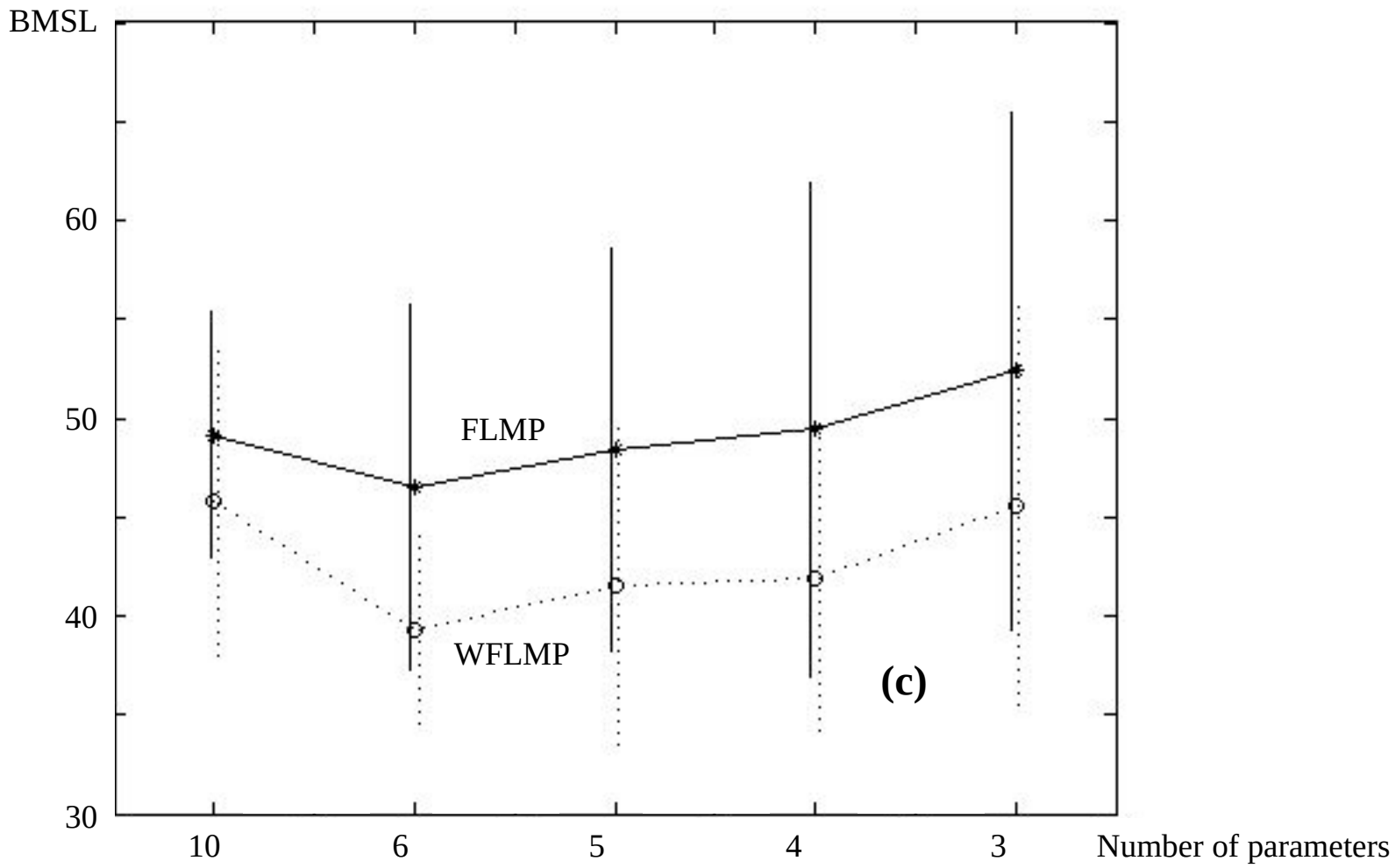

Auditory
P(/da/)
1
0.8
0.6
0.4
0.2
0
A1
A2
A3
A4
A5
Visual
P(/da/)
1
0.8
0.6
0.4
0.2
0
V1
V2
V3
V4
V5
Audio-visual
P(/da/)
V5
V4
V3
V2
V1
1
0.8
0.6
0.4
0.2
0
A1
A2
A3
A4
A5
(a)

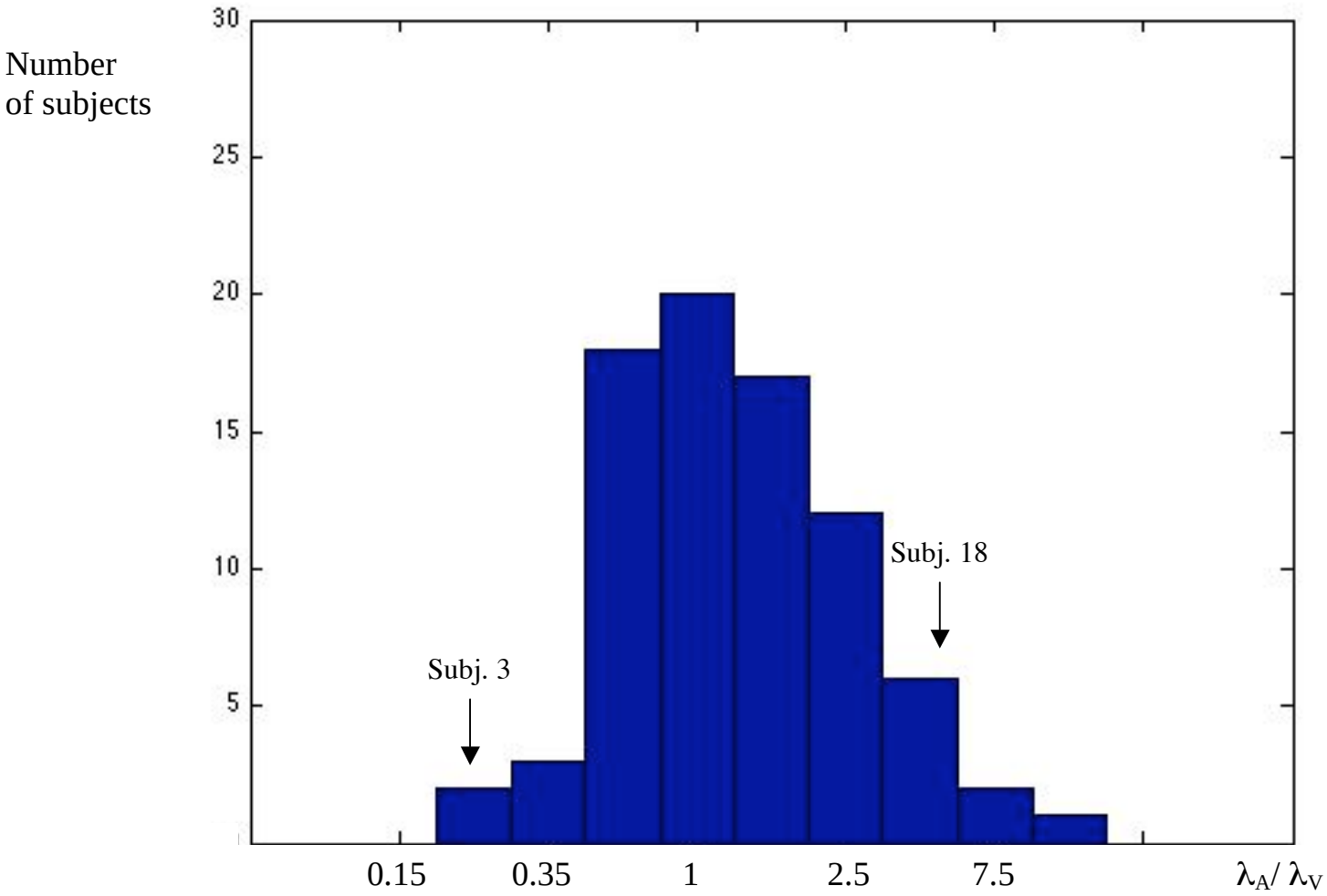

Auditory
P(/da/)
1
0.8
0.6
0.4
0.2
0
A1
A2
A3
A4
A5
Visual
P(/da/)
1
0.8
0.6
0.4
0.2
0
V1
V2
V3
V4
V5
Audio-visual
P(/da/)
1
0.8
0.6
0.4
0.2
0
V5
V4
V3
V2
V1
A1
A2
A3
A4
A5
(b)

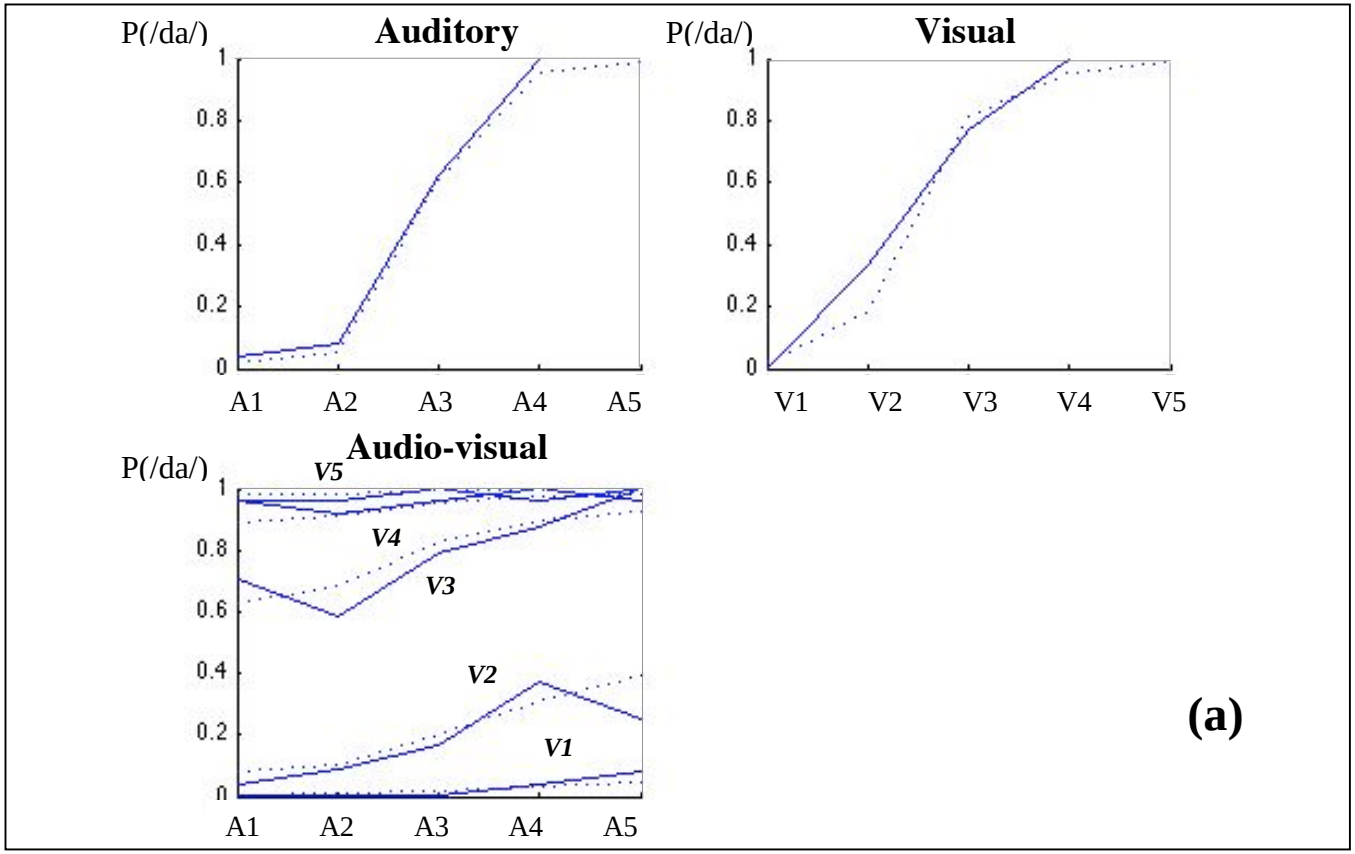

Auditory
Visual
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
A1
A2
A3
A4
A5
V1
V2
V3
V4
V5

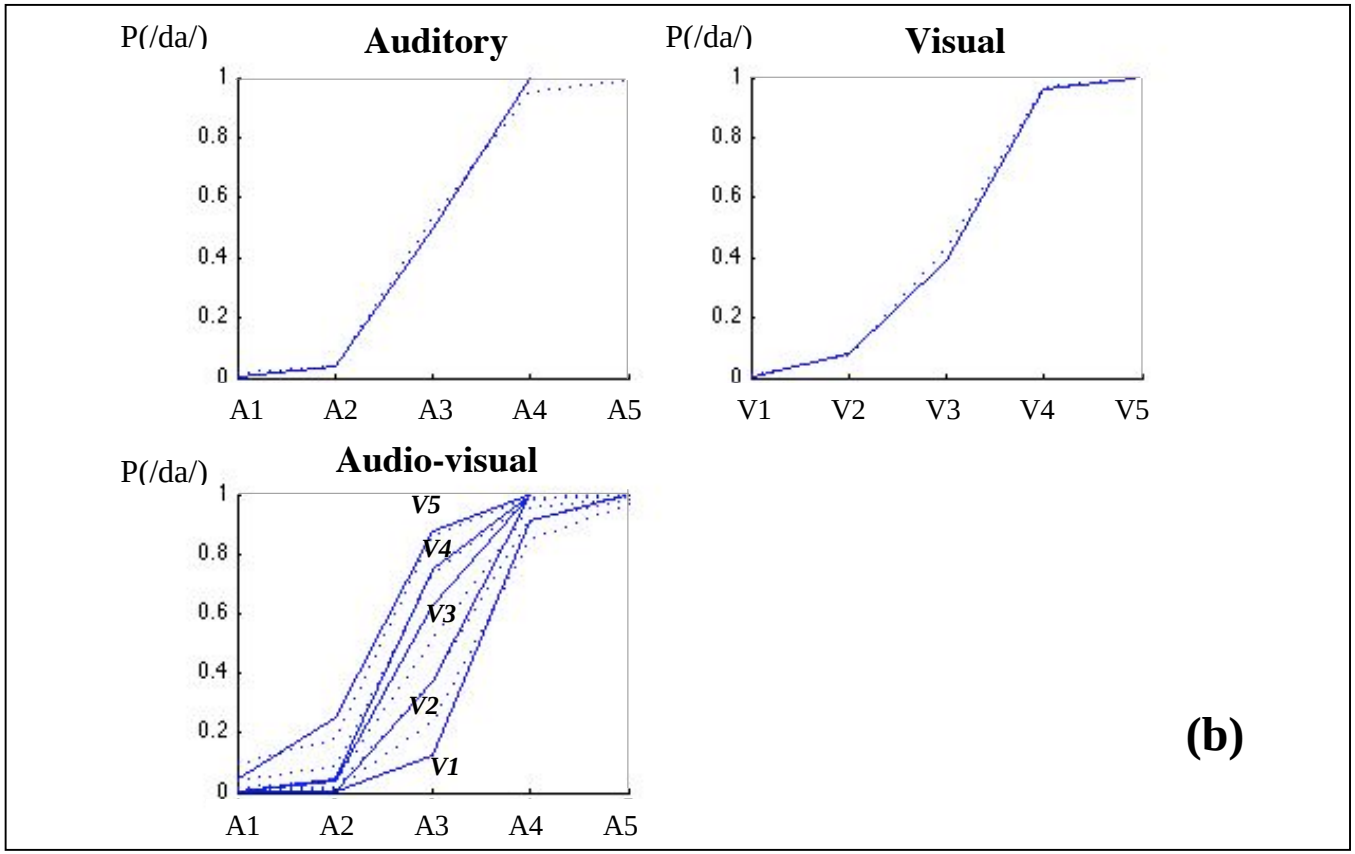

RMSE
0.12
0.10
0.08
0.06
0.04
FLMP
WFLMP
(a)
10
6
5
4
3
Number of parameters


Corr.
RMSE
0.12
0.10
0.08
0.06
0.04
FLMP
WFLMP
(b)
10
6
5
4
3
Number of parameters


BMSL
60
50
40
30
FLMP
WFLMP
(c)
10
6
5
4
3
Number of parameters


Number
of subjects
30
25
20
15
10
5
Subj. 18
Subj. 3
0.15
0.35
1
2.5
7.5
λA/ λV


P(/da/)
Auditory
1
0.8
0.6
0.4
0.2
0
A1
A2
A3
A4
A5
P(/da/)
Visual
1
0.8
0.6
0.4
0.2
0
V1
V2
V3
V4
V5
P(/da/)
Audio-visual
V5
V4
V3
V2
V1
1
0.8
0.6
0.4
0.2
0
A1
A2
A3
A4
A5
(a)


P(/da/)
Auditory
1
0.8
0.6
0.4
0.2
0
A1
A2
A3
A4
A5
P(/da/)
Visual
1
0.8
0.6
0.4
0.2
0
V1
V2
V3
V4
V5
P(/da/)
Audio-visual
V5
V4
V3
V2
V1
1
0.8
0.6
0.4
0.2
0
A1
A2
A3
A4
A5
(b)