Modeling audio-visual speech perception

Back on fusion architectures and fusion control

Jean-Luc Schwartz & Marie Cathiard

Institut de la Communication Parlé (ICP) (Speech Communication Institute)
CNRS UMR 5009, INPG / Université Stendhal, Grenoble France
schwartz, cathiard@icp.inpg.fr

ABSTRACT

In a review paper about audio-visual (AV) fusion models in speech perception, we (Schwartz et al., 1998) proposed a taxonomy of models around two basic questions: architecture and control. Six years after, it appears that the proposals we made still seem rather convenient for discussing major questions about AV fusion. Moreover — and more importantly — recent experimental and theoretical progress seem to provide some elements of answer in both aspects. The aim of this paper is to review these elements, and to incorporate them into the general architecture-and-control framework.

1. FUSION ARCHITECTURES

1.1. The four architectures for audio-visual fusion

In his well-known presentation of audio-visual models of speech perception, Summerfield (1987) introduced the concept of “metrics for audio-visual integration”, focusing on “the representations of the auditory and visual streams of information at their conflux”. The general literature on sensory interactions in cognitive psychology, and on sensor fusion in information processing, lead us conclude that there are four basic architectures (Schwartz et al., 1998). Their common point is that they should connect two separate inputs to one common output. The conception of a single output “loosing” in some sense the monosensorial nature of each input may be discussed (see the “convergence vs. association” debate raised by Bernstein et al., in press). However, even in a conception of two separate routes in interaction from the input to the output, the questions addressed in this section remain valid, provided that they are rephrased in terms of: under what format are the A and V representations represented in their sensory pathway when they interact in the route towards phonology or lexicon? In the Separate Identification (SI) model, the A and V representations are phonetic, that is mediated by the knowledge the subject has of his/her own language. In the Dominant Recoding (DR) model, the visual input is recoded into an equivalent sound, or of his/her own language. In the Direct Identification (DI) model, none of these process occur before phonetic identification which operates directly on a set of joined A and V parameters.

In our view, the static vs. dynamic issue (or the shape vs. movement debate) is independent of the architecture. In consequence, the preference for static or dynamic parameters, if any, should not lead to select one or the other architecture. Notice that this position is itself controversial, and it has been often argued that movement and motor representation were linked topics (see e.g. Rosenblum & Saldana, 1998; Whalen et al., in press). However, we have several times advocated that recovering the vocal tract shape could be done without necessarily calling for dynamic features (Cathiard et al., 1996), and proposed a “shape from shading from movement” approach in line with recent neurophysiological computational models (Cathiard et al., 2003).

1.2. The “very early” route

The four architectures share a common assumption of independence of the primitive monosensory processing. That is, information would be first extracted separately in each sensorial channel before interaction and fusion. However, a number of recent studies on the detection of speech in noise have raised serious doubts about this assumption (since Grant and Seitz, 2000). Our own contribution was to determine if this gain in detection could contribute to a gain in identification. In a recent study (Schwartz et al., 2004), we showed, thanks to an original paradigm, that seeing the speaker’s lips does enable to better hear and hence better understand. The stimuli used in this set of experiments could not lead to lipreading per se since they corresponded to exactly the same lip gesture. However, intelligibility of these stimuli merged in noise was improved just because the acoustic cues were better extracted thanks to vision. The experimental trick consisted in dubbing the same lip gesture on a number of visually similar but auditorily different configurations, e.g. [y u ty ku du gu yu gu] in French. The visual stimulus did not enable to identify the syllable, but it provided a temporal cue improving the audio identification of these stimuli embedded in a large level of cocktail-party noise, and particularly the identification of plosive voicing. Replacing the visual speech cue (the lip rounding gesture) by a non-speech one with the same temporal pattern (a red bar on a black background, increasing and decreasing in synchrony with the lips) removed the benefit. Therefore, cross-modal interactions can occur early to enhance speech in noise and improve intelligibility. This indicates that there is, whatever the architecture, a preliminary set of interactions, that we called “very early” to make clear that they correspond to a contact point that should be distinguished from early interactions in the classical sense. This is likely to provide a number of interesting technological counterparts in terms of speech enhancement, source separation and audiovisual scene analysis (e.g. Girin et al., 2001; Sodoyer et al., 2002; Berthommier, 2003).
2. FUSION CONTROL

2.1. FLMP and contextual factors: a critical view and a methodological proposal

Once the architecture is defined, the problem is to determine the nature of AV interactions at each contact point. The question we address here concerns the existence of possible contextual factors modifying the function of the fusion operator. A number of potential contextual factors have been identified in the literature, such as inter-individual variability (e.g. Seewald et al., 1985), linguistic variability (since Sekiyama & Tohkura, 1991), and attentional mechanisms (Tiippana et al., 2001).

Experimental data should be related to computational models, and used as benchmark sets for model evaluation and comparison. In this field, Massaro has always played a leading role with a very large set of comparisons between the classical “FLMP” (Fuzzy-Logical Model of Perception) and other competing models, always shown to be poorer (Massaro, 1998). It is puzzling that various experiments displaying contextual effects (e.g. linguistic specificity or visual attention) have been modelled by the FLMP with an excellent fit, while the FLMP is obviously a model defined in our terms by an SI architecture (fusion after phonetic evaluation) with a context-independent operator (multiplicative fusion). This problem has been extensively addressed in a recent paper (Schwartz, 2003), with two major points. Firstly, it appears that the FLMP is able to fit random data in the “McGurk region”, that is in all conditions including conflicting stimuli, thanks to what we call the “0/0 trick”. Remembering that the basic FLMP equation is:

\[ p_{AV}(C_i) = p_A(C_i)p_V(C_i) / \sum p_A(C_i)p_V(C_i) \]  

where \( C_i \) and \( C_j \) are 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 unimodal responses in the McGurk paradigm 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, and \( p_{AV}(C_i) \) is undetermined, and can be fitted to any value: this is the 0/0 trick.

This lead us to recall that fitting models is nothing else that assessing the “likelihood” of a given model considering a set of experimental data. 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 \( \Theta \), the estimation of the best \( \Theta \) values is provided by:

\[ \Theta = \text{argmax} p(\Theta|D,M) = \text{argmax} p(D|\Theta,M) \]  

and, in the simple case of a Gaussian model \( M \), the 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, in press). Indeed:

\[ p(D|M) = \int p(D|\Theta,M) d\Theta = \int p(D|\Theta,M)p(\Theta|M)d\Theta \]  

which means that the a priori distribution of data \( D \) knowing model \( M \) integrates the distribution for all values \( \Theta \) of the parameters of the model. This integral probability depends not only on the maximal value of \( p(D|\Theta,M) \) (that is, on the best fit) but also on the volume of \( \Theta \) values providing an “acceptable” fit (not too far from the best one) relative to the whole volume of possible \( \Theta \) values. This relative volume decreases if the model is too sensitive, as is the FLMP around its best fit in the McGurk region. The lesson of this is very simple. Considering that the data set \( D \) for an AV speech perception experiment generally consists of A, V and AV data, if a model makes a “strong prediction” for a given set of A and V data, then it will fit only a restricted set of AV responses: fit will be good if these responses are indeed displayed, and poor if they are not. The consequence is that the global fit (measured by the Bayesian term \( p(D|M) \)) will be large if the AV response is coherent with the A and V responses in relation with the predictions of the model. On the other side, since the FLMP in this special case (for conflictual A and V inputs) may predict any AV response, fit will always be good, but the available region of the parameter space able to fit the true data will be quite narrow (since if you want to predict everything, you must narrowly divide your prediction space), and the \( p(D|FLMP) \) value will be quite low. Therefore, FLMP will be a good model in terms of local best fit, and a poor model in terms of global likelihood!

2.2. A test case displaying inter-individual variability

The test case we selected to apply the Bayesian framework of model comparison was provided by an experiment by Cathiard (1994) studying A, V and AV perception of [i#a] vs. [i#i] transitions in French. The experimental procedure was based on gating. In the auditory test, the excerpt to identify included the beginning of the transition (that is, an attack “t” as “...–you said ...”– followed by the first vowel [i] plus the silent pause) and a short portion of the second vowel ([i] or [a]), with three durations of this short portion: 2, 6 or 10 ms. In the visual test, a 1-s excerpt was extracted from various initial points (10 points varying with 40 ms steps), thus including a variable extent of the trajectory from the first vowel [i] to the second vowel [i] or [a]. In the audio-visual condition, the two excerpts were combined, with a varying degree of desynchronization from 0 to 360 ms by 40 ms steps (10 steps, with the sound in synchrony or in advance). The stimuli were produced by a French male speaker, and all tests were passed 10 times by 10 French naive subjects with no deficit in vision or audition.

Auditory and visual results are displayed in Fig. 1, for sequences [i#a]. The auditory identification of [a], averaged over the 10 subjects, is weak for 2-ms stimuli, with a large inter-subject variability, and large and less variable for 6- and 10-ms stimuli (Fig. 1a). The visual identification (Fig. 1b) increases from 0 (where no portion of the transition from the lip-spread [i] to the lip-open [a] is visible) to 1 (when the whole transition is visible). The inter-subject variability is not very large, except for the sequence finishing in the middle of the transition (all articulatory analyses of lip trajectories are provided in the original reference). The interesting point is the large inter-individual variability for AV identification, with three groups (Fig. 2). In group G1 (Fig. 2a), the AV responses closely mirror the V ones, in spite of the fact that at the left of the curve, the image is just a stable [i], while the sound is perfectly identified as an [a], at least for the two longest durations. On the contrary, for G3, audition seems to dominate the AV answer, since all sequences are perceived as [a] (at last for the longest audio durations) though the visual...
2.3. Applying the Bayesian framework to the test case

To apply the Bayesian procedure introduced previously, we attempted to compare two models on the data presented in the previous section, that is the FLMP, defined by Eq. 1, and a variant of the FLMP, that we call WFLMP, in which the audio and video inputs are weighted by subject-dependent factors:

\[ p_{AV}(C) = p_A^{\lambda_A}(C) p_V^{\lambda_V}(C) / \sum p_A^{\lambda_A}(C) p_V^{\lambda_V}(C) \]  

(4)

\( \lambda_A \) and \( \lambda_V \) allow to increase or decrease the weight of the A and V contributions in the computation of the model outputs, that is the \( p_{AV}(C) \) terms. For each subject, we define a lambda value between 0 and 1, and we compute \( \lambda_A \) and \( \lambda_V \) from lambda by:

\[ \lambda_A = \text{lambda} / (1 - \text{lambda}) \quad \text{and} \quad \lambda_V = (1 - \text{lambda}) / \text{lambda} \]

with thresholds maintaining \( \lambda_A \) and \( \lambda_V \) between 0 and 1.

The first implementation of FLMP needs 13 parameters for each subject, that is 10 values of \( p_A[a] \) for each of the 10 initial positions of the gating window (V1..10) and 3 values of \( p_V[a] \) for the three durations of the initial portion of the stimulus (A1..3). There are 43 data per subject (10 V, 3 A and 30 AV responses, each response being a probability of identifying [a] on 10 trials). This provides 130 parameters altogether, for fitting a set of 430 data. Since the WFLMP model needs one more parameter per subject, we removed one parameter by fixing the value of the parameter A3 (response at a 10-ms duration) at a value equal to the mean of the value it takes for the other subjects. Then we determined for FLMP and WFLMP the best fit to the data, and computed the root mean square error for each model. We also computed the global likelihood of each model based on an approximation of the integrals in Eq. (3) provided in Jaynes (in press, p. 2404, Eq. 24-11).

Then we attempted to decrease the number of parameters, considering (from the data in Fig. 1) that some parameters could be fixed between subjects. We considered five variants of the FLMP model respectively with 13, 9, 5, 4 and 3 free parameters per subject (Table 1). On Fig. 3, we display the corresponding results. It appears that for 13 parameters per subject, the fit is good (low rmse) in both models, with no significant difference between FLMP and WFLMP. rmse increases when the number of parameters decreases. However, the log-likelihood values increase when the number of parameters decreases, until a maximum for 4 parameters. For this optimal four-parameter implementation, there is a significant gain of WFLMP over FLMP (t(9)=1.7745, p<0.06). On Fig. 4, we plot the estimated lambda values for all subjects for WFLMP with 4 parameters: the three groups clearly emerge from the picture: the visual subjects in Group G1 have a high \( \lambda_V \), the auditory ones in G3 have a high \( \lambda_A \), while \( \lambda_A \) and \( \lambda_V \) are similar in group G2.

This shows two important properties of the Bayesian framework. First, it enables to select the appropriate number of free parameters for a given experiment. Indeed, it seems reasonable to focus on 1 or 2 parameters per subject for describing the auditory curve in Fig. 1a, 1 or 2 parameters for the visual curves on Fig. 1b, plus the lambda factor. Secondly, it enables to demonstrate for the first time, in a sound statistical framework, that audio-visual fusion is not a context-free multiplicative process, but on the contrary that it depends on individual factor effects, some subjects being more “auditory” and others “visual”, independently of the A or V performances.

3. CONCLUSION

A lesson of these last years seems to be that the contact points between the A and V pathways in their route towards speech understanding are many. They include a very early level of interaction for detection and feature extraction. They should also incorporate an ability to compare a sound and an image and detect conflicts (Summerfield & McGrath, 1984), implying at some level a common format, probably pre-linguistic and motor, considering the motor activity displayed by infants in AV association tasks (Kuhl & Melzoff, 1982). Moreover, a number of cortical imagery experiments about the perception and production of actions, including speech, converge towards the existence of a cortical “Perceptual Action Understanding” system, linking a temporal region containing the multisensory description of the characteristics of action (around the Superior Temporal Sulcus, STS, with a convergence of the auditory and visual pathways); a posterior parietal area coding the motor specification of the perceived action; and an inferior frontal region (including the Broca area) coding and interpreting the goals of the action, hence the partner’s intentions (see e.g. Iacoboni, 2004). This cortical circuit is a good candidate for implementing a Motor Recoding model in which audio-visual interactions would be mediated by action understanding. Finally, the weight of the auditory and visual pathway at the various contact points probably depends on various stimulus-dependent and stimulus-independent factors, which should also play an important role in the neurophysiological pattern of interactions in the brain (e.g. Giard & Peronnet, 1999).

References
Hove: Psychology Press.
Summerfield, Q. (1987). In B. Dodd and R. Campbell (Eds.), Hearing by eye (pp. 3-51). Lawrence Erlbaum Associates.
Whalen et al. (in press). A sex difference in audio-visual integration of speech.
Table 1 – The five variants of FLMP and WFLMP

<table>
<thead>
<tr>
<th>Npar (param. per subject)</th>
<th>Parameters for FLMP</th>
<th>Par. for WFLMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>V1..10, A1..3</td>
<td>+ lambda A3 fixed</td>
</tr>
<tr>
<td>9</td>
<td>V3..8, A1..3</td>
<td>+ lambda A3 fixed</td>
</tr>
<tr>
<td></td>
<td>V1..2 and V9..10 fixed</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>V5..6, A1..3</td>
<td>+ lambda A3 fixed</td>
</tr>
<tr>
<td></td>
<td>V1..4 and V7..10 fixed</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>V5, A1..3</td>
<td>+ lambda A3 fixed</td>
</tr>
<tr>
<td></td>
<td>V1..4 and V6..10 fixed</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>V5, A1..2</td>
<td>+ lambda A2 fixed</td>
</tr>
<tr>
<td></td>
<td>V1..4, V6..10 and A3 fixed</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1 – Auditory (a) and Visual (b) identification scores (% of [a] responses)

Figure 2 – Audio-visual identification scores (% of [a] responses)
Columns: G1 (left, a), G2 (middle, b), G3 (right, c)
Lines: Audio duration= 2 ms (top), 6 ms (middle), 10 ms (bottom)

Figure 3 – Variations of rmse (left) and log-likelihood (right) for FLMP and WFLMP, in function of the number of parameters per subject (13, 9, 5, 4 or 3 parameters)

Figure 4 – $\lambda_\alpha/\lambda_\nu$ for the 10 subjects in the 4-param. model
Notice the different values for subjects of groups G1, G2, G3