AUDITORY, VISUAL AND AUDIOVISUAL VOWEL REPRESENTATIONS: EXPERIMENTS AND MODELLING

Jordi Robert-Ribes, Jean-Luc Schwartz, Pierre Escudier
Institut de la Communication Parlée, Grenoble, France
{jordi,schwartz}@icp.grenet.fr

ABSTRACT

Audiovisual (AV) speech perception exploits the inherent complementarity of the auditory (A) and visual (V) sensors. We provide new data on the expansion of the vowel triangle in the auditory and visual domain, and on the optimal use of the A-V complementarity for fusion. Then we propose a taxinomy of models for AV integration, and we show that the data in the literature are rather in favor of the so-called MR model, recoding the A and V inputs into an intermediary motor space where integration occurs. Finally, we show that the MR model is not only *plausible* but also *functional*, since it efficiently models AV identification for vowels in noise.

INTRODUCTION

It is now largely accepted that speech is a multimodal means of communication that is conveyed by the auditory and visual system ([1], [2], [3], [4], [5]). How do humans fuse the auditory and visual information? How can recognition systems integrate audio and visual information? Answering the first question we will deal with plausibility, while answering the second we will have to do with functionality.

Studies on audio-visual integration generally have one (*and only one*) of these two approaches: (1) engineering approach, dealing with functionality constraints or (2) experimental psychology approach, dealing with plausibility constraints (and most of the time bypassing the conversion process from inputs into internal representations). This paper studies models of audio-visual speech integration taking into account both plausibility constraints *and* functionality constraints, with an application to vowel perception.

Apart from general questions about sensory interactions and cognitive processes, we have been defending for years the idea that perceptual processing cannot be understood without a deep knowledge of the *structure of the stimuli* [6]. Our section 1 will here be concerned with a set of new data about the perceptual expansion of the vowel triangle in the A and V domain. This will clearly show the *complementarity* of the A and V sensors for vowel place of articulation. In section 2 we will propose a taxinomy of models for AV integration in speech perception, with three successive binary questions leading to four categories of models. We will show how experimental data constrain the choice towards one category, the so-called Motor Recoding Model. Finally, we will show that this model is not only plausible, but also functional, since it efficiently models AV identification of vowels in noise.

1. AUDIO-VISUAL VOWELS

1.1. Physical characteristics

We recorded the vowels [i e ɛ y ø œ u o ɔ a] from a French speaker with the ICP "Video-Speech Workstation" [7]. We obtained images of the speaker's face with the corresponding synchronised sounds. We recorded, for each vowel, of 100 realisations of 64 ms of sound with the corresponding image.

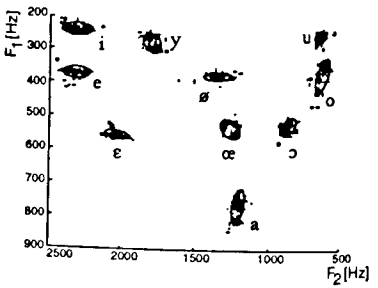

Figure 1. F_1/F_2 representation of the acoustic stimuli used

Their two first formants are presented in Figure 1. We observe the vocalic triangle with vowel [y] close to vowel [i] but far apart from vowel [u]. Notice that even considering higher formants by a global spectral shape analysis, the [i-y] distance remains smaller than the [y-u] one [8].

We extracted from the image the following geometrical parameters of the lip shape: inner-lip horizontal width (A), inner-lip vertical height (B) and inner-lip area (S). Figure 2 presents the stimuli in the A/B plane. We observe a clear separation of rounded vowels ([y ø u o]), semi-rounded vowels ([ɔ œ]) and unrounded vowels ([i e ɛ a]). Notice that vowel [y] is close to vowel [u] but very far apart from vowel [i].

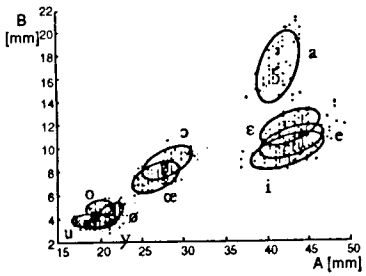

Figure 2. Width (A) / height (B) representation of the optic stimuli used

We summarize the [i y u] acoustical and optical contrasts in Figure 3.

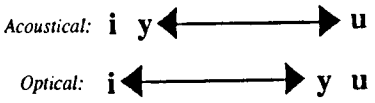

Figure 3. Acoustical and optical contrasts

We see that our stimuli are clearly *complementary*. A number of further analyses of the same stimuli [8] confirm that height contrasts (e.g. [i] vs. [e] vs. [ɛ] vs. [a]) are best represented within the acoustic stimuli. On the contrary, optic stimuli differentiate mainly the stimuli by their rounding (e.g. [i] vs. [y]).

1.2. Audio, visual and audio-visual perception

We carried out a perceptual test on these vowels presented with acoustic noise.

Method

A group of 21 French subjects was presented with 10 realisations of the 7 French isolated (without context) vowels [i e y ø u o a] in audio-visual, visual and audio conditions, with 7 signal-to-noise ratios (SNR). We tested these 7 vowels instead of the 10 available because we didn't want to test the mid-low/mid-high contrast (e.g. [ɛ] vs. [e]) which may be lost in isolation.

Results

Figure 4 shows the correct identification results in percentage corrected to the random level (zero percent means that scores were at random level). This figure shows better audio-visual scores (AV) than audio alone scores (A) and visual scores (V). More detailed analyses based on transmitted information show that this pattern is true for individual phonetic dimensions, namely rounding, height and front-back contrast [8]: we call this the "complementarity rule".

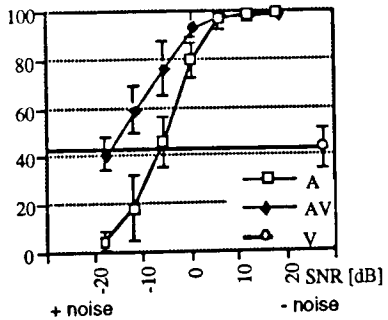

Figure 4. Correct identification scores for the perception test

We also studied the perceptual structure of the auditory and visual perception (for full details, see [8]). We show in Figure 5 a schematic display of the structure we found. We can see that the auditory geometry is stretched in the height dimension ([i] vs. [a]) while the visual geometry is stretched in the rounding dimension ([i] vs. [y]).

Hence our data reveal some audio-visual complementarity: the best information about place of articulation perceived by audition is the worst perceived by vision and vice-versa. Notice that up to now audio-visual complementarity had been rather conceived as an Audition-Mode Vision-

Place complementarity [5]. The complementarity we found in our test is a *complementarity within place of articulation.* The complementarity found in the stimuli (section 1.1) is enhanced by the perceptual system (section 1.2).

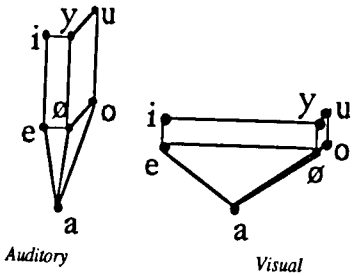

Figure 5. Perceptual structures

2. MODELS FOR AUDIO-VISUAL INTEGRATION

The main theoretical options for modelling audio-visual integration in speech perception have been presented by Summerfield [5]. In the following sections we present these options with slight modifications from our own.

2.1. Four models

Direct Identification (DI Model)

This model assumes a direct identification of the inputs without any transformation. The audio (A) and visual (V) inputs are compiled in a bimodal vector and then classified. Some components of the bimodal vector are A and some components are V (Fig. 6).

Key characteristic of the DI model:

There is *no representation level common to both modalities* between the signal and the percept.

Separate Identification (SI Model)

This model assumes that both inputs have been compared to prototypical forms before the fusion stage. The comparison to prototypes can even lead to a classification of each modality. The two codes (one from the A input and one from the V input) are then combined by means of rules or logical criteria (Fig. 7).

The information at the fusion level can also be continuous (and not discrete).

The key point here, however, is whether it is the result of a comparison to prototypes or not. Therefore, the FLMP (Fuzzy Logical Model of Perception [4], [9]) is one of the SI models because the information at the fusion level is information "indicating the degree of support for one alternative" ([9], p. 743).

Key characteristic of the SI model:

Inputs are compared to prototypes (or even classified) *before fusion.*

Dominant modality Recoding (DR Model)

One of the possibilities that Cognitive Psychology presents for fusing two modalities is the recoding of one modality into the other —supposed to be the dominant modality— [10]. The DR model assumes that the auditory modality is dominant in speech perception. Thus the visual input is recoded into an auditory space where both sources of information are fused [11].

In this model the visual input is used to estimate the vocal tract filter. This estimation is then in some sense averaged with the one derived from auditory processing, while the source characteristics are estimated only from the auditory path. The combined source and filtering characteristics thus estimated are then provided to a phonetic classifier (Fig. 8).

Key characteristic of the DR model:

The *visual modality is recoded into an auditory representation space* where it is fused with the auditory information.

Motor space Recoding (MR Model)

This model assumes that both inputs are projected into a common amodal space where they are fused. This space is amodal because it is homogeneous to neither of the modalities (auditory or visual): it is a motor representation space. This supposes that, in order to perceive speech, we recover the common cause of both the auditory and the visual signals, namely: the motor representation [12] (Fig. 9).

Key characteristic of the MR model:

Both inputs are projected into a motor representation space where fusion occurs.

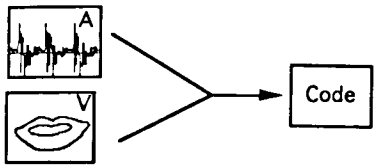

Figure 6. DI Model

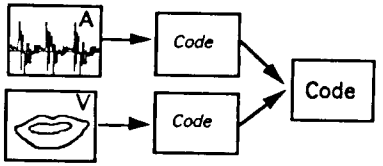

Figure 7. SI Model

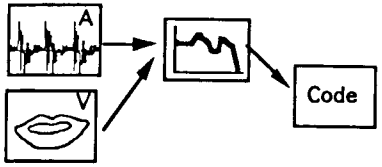

Figure 8. DR Model

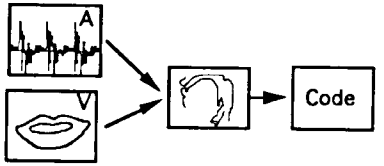

Figure 9. MR Model

2.2. Three questions for a taxonomy

We can attempt to classify the above presented architectures under a synthetic form by comparing them to the general cognitive psychology models. To do so, we can ask three questions.

(1) Does the interaction between modalities imply a *common intermediate representation* or not? The answer allows a first opposition between Model DI (for which there is NO common representation) and the other architectures.

(2) If there is an intermediate representation, does it rely on the existence of prototypes or not? In other words, is it "late" or "early" (see [13], p. 25)?

Integration is considered "late" when it occurs after the decoding processes or the comparison to prototypes (Model SI), even if these processes give continuous data (as with the FLMP). Otherwise, integration is "early" when it applies to continuous representations, common to both modalities, and which are obtained through low-level mechanisms which do not rely on any decoding process nor any comparison to prototypes: It is the case with models DR and MR.

(3) Is there at last any dominant modality which can give a common intermediate representation in an early integration model (DR)? Or is this common representation amodal (such as in the MR model)?

These questions lead to the taxonomy presented on Figure 10.

2.3. Three "plausible" responses to the three questions

About the need for a common representation

In a study on the audio-visual perception of vowels, Summerfield and McGrath [14] showed that subjects detect the incompatibility between the auditory and visual inputs, while they cannot however avoid fusing both inputs.

Even young babies are sensitive to the correspondence between auditory and visual information of a speaking face [15]. When they are presented with two faces and only one sound, babies prefer to look at the face that is articulating the

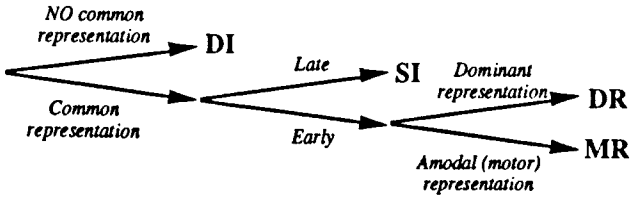

Figure 10. Taxonomy of models

sound they are hearing.

Therefore, the need of a common representation of auditory and visual stimuli is inescapable. Stimuli can be compared in that space before being fused: *stricto sensu*, model DI has to be rejected.

About the need for early integration

Subjects are able to estimate temporal co-ordinations between the auditory signal and the visual signal: they can estimate the VOT audiovisually ([16], [3]). This is clearly incompatible with a late integration model where neither of the two inputs provides information enough to identify the voicing feature, which is hence recovered from the co-ordinations between the auditory and visual signals. It seems that the evaluation of this co-ordination should be done on signals that are not the output of a prototype comparison (in which case the information leading to the co-ordinations would be lost).

In another experiment, the speaking rate perceived audiovisually is the mean of the speaking rate perceived auditorily and the speaking rate perceived visually [17]. In addition, the audio-visual speaking rate can change the phonemic frontier between voiced and unvoiced phonemes [18]. This is hardly compatible with a late integration model because rate is a quantitative information which would be lost after comparison to prototypes.

These facts indicate that subjects can make decisions from audio-visual information, decisions that are impossible to make for each modality independently. The fusion has to be done at an early stage of processing. Late models (SI Model) have to be rejected.

About dominance and complementarity

We have seen in section 1.2 that the best information perceived by audition is the worst perceived by vision and vice versa. The DR Model cannot exploit this complementarity because all the inputs are transformed into an acoustic representation. Thus, audiovisual confusions will be similar to acoustic confusions. Let us develop this idea.

Model DR recodes the visual input into an auditory representation where fusion takes place. From our data in section 1.1, two facts have to be pointed out: (1) Model DR will try to recode visual stimuli distant in the visual space (as [i] and [y]) into close points in the auditory space, and (2) stimuli close in the visual space (as [y] and [u]) will need to be recoded into distant points in the auditory space. This is represented in Figure 11.

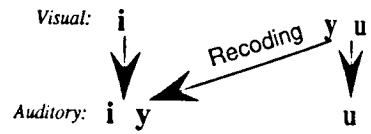

Figure 11. Bad recoding of Model DR

Point (2) results in a lack of stability of the visual-to-auditory transform, which makes the problem of building this association not trivial [19]. Point (1) results in a lack of optimality, which makes the DR Model not very efficient for dealing with vowels in noise (see section 3). But, more seriously, the DR Model predicts that a given V input can modify an A percept only if the V input is conflictual or if the A input is noisy (see an application of the DR Model structure for noisy speech enhancement in [20]). However, some data in a study by Lisker and Rossi [21] show that a V input can bias a congruent and clear auditory stimulus. Their subjects (French-speaking speech researchers) were asked to decide on the rounding category of each vowel. Let us concentrate on the case of [ɯ]. This vowel when presented visually was considered a rounded vowel 1% of the time. When presented auditorily, it was rather considered rounded (60% vs. 40% unrounded responses). Finally, when presented audio-visually the percentage of rounding judgements dropped to 25% (vs. 75% unrounded).

Hence, it is clear that some subjects perceived the vowel [ɯ] as rounded when presented auditorily but they judged it unrounded when presented visually and audio-visually. This fact is hardly conceivable within a DR Model which recodes visual input into an auditory space. Consequently, the DR Model has to be rejected.

Conclusion: plausibility constraints

We have seen that (1) a plausible model of audio-visual integration has to use a common representation between audio and visual inputs, (2) this representation must be placed at an early stage of processing, and (3) this representation cannot be the auditory representation.

The only model that we have not been able to eliminate (Model MR) is in agreement with these three points. We will see in section 3.5 that it can simulate the results found by Lisker and Rossi [21].

In conclusion, the MR Model is "plausible": we will see in the next section that it is also "functional".

3. MR MODEL FUNCTIONALITY

We will present in this section an implementation of the MR Model for the recognition of French vowels. This is the first implementation of this model in the literature. We proved elsewhere ([8], [22]) that the DR Model is less functional than the MR one, namely that it has worse results in a recognition task.

The MR Model implementation is based on simple but controllable tools, which were chosen to allow a good comparison with the DR model (see [8]).

3.1. Motor representation

A crucial choice in the MR Model concerns the definition of the "motor space" in which integration should occur. Since we deal with static vowels, we have chosen articulatory representations based on three parameters, namely X, Y (which are respectively the horizontal and vertical co-ordinates of the highest point of the tongue) and S (the inner-lip area). Of course, X, Y and S respectively provide articulatory correlates of the front-back, open-close and rounding dimensions (see Fig. 12).

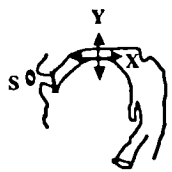

Figure 12. Motor representation

3.2. From the inputs to the motor representation

This stage was implemented thanks to linear associations.

The auditory inputs were 20-dimensional dB/Bark auditory spectra. The outputs were typical values for X and Y for French vowels, and the S value extracted on the corresponding image in the corpus.

The visual inputs were (A, B, S) triplets. X (the tongue front-back position) was supposed to be impossible to estimate from the visual path, and S was directly transmitted from the visual input, hence only the association between (A, B, S) and Y had to be learned.

Notice that all the corpus (e.g. 10 vowel classes) was used in this study.

3.3. Fusion of motor representations

The integration consisted in a weighted sum of the representations obtained by each path (audio and visual). An audio-visual estimate of the (X, Y, S) set was finally derived. The parameter X was estimated only from the acoustic path, while the other two parameters were determined from the corresponding ones provided by both paths. This was performed using the following formulas:

$$Y_{AV} = \alpha_Y Y_A + (1-\alpha_Y) Y_V$$

and $$S_{AV} = \alpha_S S_A + (1-\alpha_S) S_V$$

where index A means auditory, V visual and AV audio-visual. Parameters α_Y and α_S are sigmoidal functions of SNR. The parameter α_Y varied between a value close to 0 for low SNR values (too much noise; almost no available information in the acoustic signal) and a value lower than 1 for high SNR values (no noise; the audio-visual percept is influenced by both the visual and the auditory inputs). On the other hand, the parameter α_S was never higher than 0.3 (indicating that the estimation of S is mainly done from the information of the visual path). The parameters of the sigmoids were learned under a criterion of minimal global error for all learning realisations at all SNR values.

3.4. Vowel identification

Classification was achieved by a Gaussian classifier. We used a Gaussian classifier in the (X, Y, S) space, with a

choice of one between ten classes. The learning corpus for estimating the mean and covariance matrix for each vowel class was based on (X, Y, S) triplets delivered by the auditory path alone on realisations presented at 4 different levels of noise covering a large range between no noise and largely degraded but still partly recognisable stimuli (SNR = 99, 24, 12 or 0 dB).

The whole schema is displayed in Figure 13.

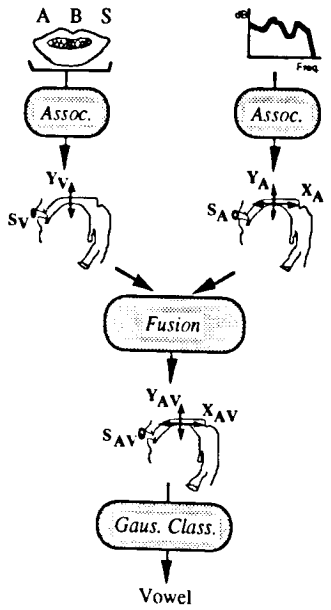

Figure 13. Implementation of the MR Model

3.5. MR Model and complementarity

How could this implementation simulate Lisker and Rossi's data [21] presented in section 2.3? It is known that rounding is highly correlated with parameter S (inner-lip area). We saw that the audio-visual estimate of S by the MR Model mainly depends on the value of S provided by the visual input. However, one value of S is also estimated from the auditory input. Then, a decision about rounding can be taken from the auditory input alone or from the visual input alone. When both inputs (audio and visual) are present, the rounding decision mainly depends on the decision taken from the visual input. This was exactly the case found in the experience of Lisker and Rossi [21] described earlier. Our implementation of the MR Model can simulate this result.

3.6. Results

We present in Figure 14 the identification scores when the audio or the audio-visual stimuli were presented at the input of the implementation. Since dimension X (front-back) cannot be estimated from the visual input, we estimated a visual score by considering (arbitrarily) all rounded vowels as being back-vowels.

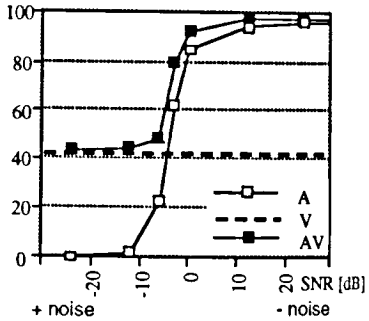

Figure 14. Correct identification for the MR Model

Recognition scores in Fig. 14 are lower than perceptual scores in Fig. 4. This is due to both the higher number of classes in the second case (10 vs. 7) and the simplicity of tools in MR implementation. However, the interesting point is that the visual gain (difference between the A and AV conditions) is high in the model, and it respects the "complementarity rule": transmitted information on each of the three phonetic dimensions is greater for the AV condition than for both the A and V conditions [8].

CONCLUSION

We have shown that a plausible model of audio-visual integration for speech perception requires three characteristics: (1) use a common representation between audio and visual inputs, (2) fuse the modalities at an early stage of processing, and (3) do not use the auditory space as the fusion space.

A model that fuses both informations in the motor space (Model MR) has these three characteristics. We have shown that this model is also functional in a vowel recognition task. We are currently attempting to adapt this model to dynamic stimuli, with more complex processing tools and architectures. We hope that this model should be able to produce some McGurk effect (see [8]).

Acknowledgement

This work has been supported by ESPRIT-BR (6975) funding (Speech Maps Project).

REFERENCES

[1] Erber, N.P. (1975). "*Auditory-visual perception of speech*", J. Speech and Hearing Disorders 40, 481-492.

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

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

[4] Massaro, D.W. (1987). *Speech perception by ear and eye : a paradigm for psychological inquiry*. London: LEA.

[5] Summerfield, Q. (1987). "Some preliminaries to a comprehensive account of audio-visual speech perception", in Dodd, B. and Campbell, R. (Eds.) *Hearing by eye : the psychology of lipreading*. (pp. 3-51). London: LEA.

[6] Schwartz, J.L. and Escudier, P. (1991). "Integration for extraction: What speech perception researchers can learn from Gibson and Marr," in *XII Congrès International des Sciences Phonétiques*. (Vol. 1, pp. 68-72).

[7] Lallouache, M.T. (1990). "Un poste 'visage-parole'. Acquisition et traitement de contours labiaux," in *XVIII Journées d'Études sur la Parole*. pp. 282-286).

[8] Robert-Ribes, J. (1995). *Models of audiovisual integration*. PhD Thesis Institut National Polytechnique de Grenoble.

[9] 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.

[10] Hatwell, Y. (1986). *Toucher l'espace. La main et la perception tactile de l'espace*. Lille: Presses Universitaires de Lille.

[11] Yuhas, B.P., Goldstein, M.H., and Sejnowski, T.J. (1989). "*Integration of acoustic and visual speech signals using neural networks*", IEEE Communications Magazine Nov. 89, 65-71.

[12] Fowler, C.A. and Rosenblum, L.D. (1991). "The perception of phonetic gestures", in Mattingly, I.G. and Studdert-Kennedy, M. (Eds.) *Modularity and the motor theory of speech perception*. (pp. 33-59). Hillsdale (NJ): Erlbaum.

[13] Vroomen, J.H.M. (1992). *Hearing voices and seeing lips: Investigations in the psychology of lipreading*. PhD Thesis Katolieke Univ. Brabant.

[14] Summerfield, Q. and McGrath, M. (1984). "*Detection and resolution of audio-visual incompatibility in the perception of vowels*", Quarterly J. Experimental Psychology : Human Experimental Psychology 36A, 51-74.

[15] Kuhl, P.K. and Meltzoff, A.N. (1982). "*The bimodal perception of speech in infancy*", Science 218, 1138-1141.

[16] Rosen, S., Fourcin, A.J., and Moore, B. (1981). "*Voice pitch as an aid to lipreading*", Nature 291, 150-152.

[17] Green, K.P. and Miller, J.L. (1985). "*On the role of visual rate information in phonetic perception*", Percep. and Psychophysics 38, 269-276.

[18] Green, K.P. and Kuhl, P.K. (1989). "*The role of visual information in the processing of place and manner features in speech perception*", Perception and Psychophysics 45, 34-42.

[19] Robert-Ribes, J., Lallouache, T., Escudier, P., and Schwartz, J.L. (1993). "Integrating auditory and visual representations for audiovisual vowel recognition," in *Proc. 3rd Eurospeech*. (pp. 1753-1756).

[20] Girin, L., Feng, G., and Schwartz, J.L. (1995). "Noisy speech enhancement with filters estimated from the speacker's lips," in *Proc. 4th Eurospeech-95*.

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

[22] Robert-Ribes, J., Schwartz, J.L., and Escudier, P. (in press). "*A comparison of models for fusion of the auditory and visual sensors in speech perception*", Artificial Intelligence Review.