AUDIO-VISUAL RECOGNITION OF SPEECH UNITS: A TENTATIVE FUNCTIONAL MODEL COMPATIBLE WITH PSYCHOLOGICAL DATA

Jordi Robert-Ribes, Jean-Luc Schwartz, Pierre Escudier

Institut de la Communication Parlée
Institut National Polytechnique de Grenoble (France)

ABSTRACT - We compare four models (issued from psychology) of Audio-Visual (AV) integration for speech perception. First, we show that two of them are incompatible with psychological data. Then we present a perception test on AV identification with evidence for a complementarity of A and V in what concerns *place of articulation*. Finally, we present an implementation of the two remaining plausible models, and we show that one of them makes a better use of the complementarity, and hence achieves higher recognition scores.

INTRODUCTION

Speech is a multimodal means of communication. It is conveyed by acoustic and optic stimuli, and processed both by the auditory and visual system. Our work deals with the way of combining this two streams of information.
In section I, we will review some of the models proposed in the literature for audio-visual fusion in speech perception. Section II will try to expose the facts that enable us to choose between these models. In section III we will present psychological data on the audio-visual perception of French vowels. And section IV will deal with our implementation of two of the models described before. This comparison will lead us to conclude that the best model is the one that transcodes both inputs into a motor space in order to fuse the two informations in that space.

I. MODELS OF SENSORY FUSION IN SPEECH PERCEPTION

The problem of intermodal transfer and integration is well-known and intensively studied in psychology (see review in Stein & Meredith, 1993), with three main categories of proposals considering intersensory fusion, namely (1) language or more generally symbolic attributes as the support for fusion, (2) recoding of one modality into another, generally supposed to be "dominant" for a given perceptual task, or (3) amodal continuous representations characterizing the physical properties of the distal source independently of the proximal sensations (direct realist theory of perception, see Gibson, 1966). In the field of speech perception, the main theoretical options have been most clearly presented by Summerfield (1987), with four main options:
(DI) *Direct Identification from audio-visual configurations*, or an extended version of Klatt's "Lexical access from spectra" (Klatt, 1979) into a "Lexical access from spectra and face parameters" model;
(SI) *Separate Identification from auditory and visual inputs*, followed by a fusion of the phonetic features identified by one or the other modality; the fusion can then operate either on logical values, as in the "Vision Place Auditory Mode" model in which each modality in the audio-visual mode is in charge for a specific set of phonetic features (McGurk & MacDonald, 1976; Summerfield, 1987), or on fuzzy logical values (as in Massaro's Fuzzy Logical Model of Perception, FLMP, see Massaro, 1987) or some equivalent as in the attempts to use neural network models (see Robert-Ribes et al., in press).
(RD) *Recoding in the Dominant modality* (acoustic),visual speech as a means of estimating the vocal tract filter, the estimation being 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.
(RM) *Recoding in the Motor space*, transcoding of both sensory modalities into estimates of gestural representations at any level (vocal tract geometry or deep motor commands), followed by a fusion in the common motor representation space and further identification in that space, as suggested by tenants of the Motor Theory (Liberman & Mattingly, 1985) or the Direct Realist Theory (Fowler, 1986) of speech perception.
We observe that Models (DI) and (SI) enter in the first category of models in which fusion occurs at a symbolic level, while Model (RD) which is based on a recoding of the visual modality into the dominant auditory one enters in the second category, and Model (RM) is a clear member of the third category of neo-Gibsonian models.

The basic structure of Models DI, SI, RD and RM is described in Fig. 1. In Model DI, there is a direct link between a set of auditory and visual features and a given lexical category. In Model SI, an intermediary step before lexical identification consists in a separate identification from either the auditory or the visual stimulus, and the fusion occurs later ("late integration", see Vroomen, 1992). In Models RD and RM, there is also an intermediary step into a common representation space. However, this common space where fusion occurs *precedes* the identification level, contrary to Model SI : Models RD and RM hence assume "early integration".

II. PSYCHOLOGICAL DATA IN FAVOUR OF A COMMON CONTINUOUS REPRESENTATION BEFORE FUSION AND IDENTIFICATION

II.1 The case of Model DI

The difference between Model DI and Models SI, RD and RM is that the first one involves *no common intermediary representation* before integration. Hence the main characteristic of Model DI is that it *does not allow any "comparison" or "matching" between the auditory and the visual inputs*. However, a number of experimental data show that the human brain performs such a comparison, in order to detect incongruities between both inputs (Kuhl and Meltzoff, 1982; Summerfield and MacGrath, 1984; Sekiyama, 1994).Therefore, there must exist some intermediary common stage where the auditory and visual inputs may be compared before fusion, and MODEL DI HAS TO BE REJECTED.

II.2 Early vs. late integration (Models RD and RM *vs.* SI)

Data by Green & Miller (1985) showing that the lip-read duration of a syllable influences the auditory decision whether the initial consonant of an audio syllable is voiced or voiceless are difficult to understand with a late-integration model, since the visually perceived syllable duration is a quantitative data that would be lost in a phonetic trace of the visual input, and hence it would not intervene in the late fusion process. Moreover, a number of data show that subjects are able to extract temporal coordinations between the auditory and the visual inputs in experiments on the audio-visual recovery of voicing (Breeuwer & Plomp, 1986). This is obviously incompatible with a late-integration model, in which neither the auditory nor the visual phonetic decoding modules would have *any* cue enabling to take a decision in respect to the voicing feature. Therefore the balance is rather in favour of Models RD or RM, and MODEL SI MAY BE REJECTED.

II.3 Optimal exploitation of the complementarity between audition and vision

A formal study (Robert-Ribes et al., in press) shows that if there is a system of axes exhibiting a natural complementarity between the two inputs, one being more precise on the first axis and the other on the second axis, the fusion should be realized in this system. We shall see in the next sections that this leads to an advantage for Model RM.

III. BIMODAL PERCEPTION OF VOWEL IN NOISE

We shall see in this section that there is a natural complementarity between audition and vision in what concerns vowel identification.

III.1 Method

A group of 21 French subjects was presented with 10 realizations of the 7 French vowels [a, e, i, ø, y, o, u] in audio, visual and audio-visual conditions, with 7 signal to noise ratios, namely: no noise, 12 dB, 6 dB, 0 dB, -6 dB, -12 dB and -18 dB.
Subjects were asked which vowel they thought the speaker had pronounced.

III.2 Results

The results are presented on Fig. 4. for the audio-alone stimuli and on Fig. 5 for the audio-visual and visual-alone stimuli. On Fig. 6 we present the gain due to the visual information expressed as the difference between the audio-visual scores and the audio-alone scores.
On Fig. 7, we present the probability of correct identification of three phonetic features, namely tongue height (/i, y, u/ vs /e, ø, o/ vs /a/), tongue front-back position (/i, e, ø, y/ vs /o, u/) and lip rounding (/i, e/ vs /ø, y, o, u/). We can see that for visual stimuli, lip rounding is perfectly identified, tongue height is moderately perceived and front-back is poorly identified. On the other hand, for the audio stimuli, tongue height is the identification most robust to noise, and front-back and rounding are less robust, with a little advantage for the former.
We see that the best information about place of articulation perceived by vision is the worst perceived by audition and vice-versa. Notice that up to now audio-visual complementarity has been rather conceived as an Audition-Mode Vision-Place complementarity : it is the first time that complementarity WITHIN PLACE FEATURES is demonstrated.

In the next section we shall investigate whether one of the models (RD or RM) takes better advantage of this complementarity than the other.

IV. MODELLING THE BIMODAL IDENTIFICATION OF VOWELS IN NOISE

This section will provide a comparison of models RD and RM in respect to the bimodal identification of French stationary vowels in acoustic noise. We shall see that Models RM performs a better exploitation of the natural complementarity described in section III.2.

IV.1 Corpus

We have selected 100 "realizations" of each of the 10 French oral vowels [a, ɛ, e, i, œ, ø, y, ɔ, o, u]. The auditory input for each occurrence was defined as the representation of the vowel spectrum in a classical (Bark, dB) space with 20 1-Bark wide channels between 0 and 5 kHz. Noisy acoustic signals were obtained by adding various amounts of Gaussian noise on the temporal stimuli, before further spectral analysis. In what concerns the visual input, we selected three main characteristics of lip gestures (Abry & Boë, 1986), namely horizontal width (A) and vertical height (B) of the internal lip opening, and lip area (S). (A, B, S) triplets provided the visual input for each occurrence.

IV.2 Models

IV.2.1 Implementation of Model RM

The whole schema is displayed in Fig. 2. A crucial choice in this model concerns the definition of the "motor space" in which integration should occur. 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.

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

From visual inputs to articulatory representations – The visual inputs were (A, B, S) triplets. X (the tongue front-back position) was supposed to be impossible to estimate on 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.

Fusion in the articulatory space – 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:

$$Y_{AV} = \alpha_Y Y_A + (1-\alpha_Y) Y_V \qquad \text{and} \qquad S_{AV} = \alpha_S S_A + (1-\alpha_S) S_V$$

where index A means auditory, V visual and AV audio-visual. α_Y and α_S are sigmoidal functions of S/N, varying between a value close to 0 for low S/N values (too much noise; almost no available information in the acoustic signal) and a value lower than 1 for high S/N values (no noise; the audio-visual percept is influenced by both the visual and the auditory inputs). The parameters of the sigmoids were learned under a criterion of minimal global error for all learning realizations at all S/N values.

Vowel identification – We used a Gaussian classifier in the (X, Y, S) space, with a choice to perform in 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 realizations presented at 4 different levels of noise covering a large range between no noise and largely degraded but still partly recognizable stimuli (S/N = 99, 24, 12 or 0 dB).

IV.2.2 Implementation of Model RD

This model was first implemented for the English vowels by Yuhas et al. (1989). The whole schema is displayed in Fig. 3.

From visual inputs to "auditory equivalent" representations – Once more, images were represented by (A, B, S) triplets, and sounds by 20-dimensional dB/Bark auditory spectra. However, since the final Gaussian classifier in Model RM was fed with 3-dimensional inputs (namely audio-visual estimates of X, Y and S), we found it fair to compare its classification results with a 3-dimensional version of Model RD. A second version (in which the auditory representation consisted in the first three principal components of the 20-dimensional auditory spectra, previously determined through a Principal Component Analysis) can be found in Robert-Ribes et al. (1993). Results of this second version were worse than results of the version presented here.

There is for French vowels a problem linked to the existence of both front and back rounded vowels, with almost the same lip pattern, but quite different spectra, and Model RD has to associate one spectrum to each lip pattern! Our approach to solve this problem has been treated elsewhere (Robert-Ribes et al., in press).

Fusion in the auditory space – The integration of the auditory spectrum $spect_A$ with the "visual" one $spect_V$ estimated from the (A, B, S) triplet was done point by point according to the formula :

$$spect_{n,AV} = \alpha\ spect_{n,A} + (1-\alpha)\ spect_{n,V}$$

with $\alpha \in [0,1]$, $n \in \{1,...,20\}$

α is a sigmoidal function of S/N whose parameters are learned using the same criterion as for model RM. A unique α function was used for Model RD so as to keep the number of parameters to tune at a reasonable level in relation to the size of our corpus.

Vowel identification – The parameters of the final Gaussian classifier were estimated on the auditory path alone, through the same learning corpus as for Model RM (50 realizations, 4 levels of noise).

IV.3 Results

Correct identification is presented as a function of S/N for Models RM and RD on Fig. 4 for the auditory-alone path and on Fig. 5 for the audio-visual path. On Fig. 4, it appears that Model RM performs surprisingly well, with results almost as good as Model RD for S/N higher than -6dB: the results are poorer only for the very high or very low S/N values for which the high dimensionality of Model RD allows the capture of fine discriminating details for classification. This result, quite unexpected, shows that the linear rearrangement of auditory spectra estimating the "articulatory" representation (X, Y, S) is in fact quite efficient for vowel classification in silence or reasonable levels of noise.

There is for both models an improvement due to the visual input on the recognition of noisy acoustic stimuli. However, the improvement is higher for Model RM than for Model RD (see Fig. 6), which is not due to a ceiling effect, since this happens even in cases where Model RM is already better than Model RD in the auditory-alone condition. Altogether, Model RM is clearly the best for audio-visual inputs, which shows that it takes a better advantage of the visual input than models integrating in an auditory space. This is probably due to the fact that there is in the (X, Y, S) space a natural decomposition into "visible" and "invisible" components, which makes the integration process much more efficient by taking advantage of the complementarity.

Notice finally that the perceptual data are better than all modelling experiments, which is mainly due to (1) the very simple (linear) structure of our classifiers, and (2) the fact that the /e-ɛ/, /ø, œ/ and /o, ɔ/ distinctions were ignored in the psychological experiment.

CONCLUSION

Audio-visual fusion in speech perception seems to be characterized by four main principles:

1. There must exist a *common representation space* where the auditory and visual inputs are projected and may be matched and compared before a final lexical decision is taken;
2. This representation must be *continuous* and it is likely to *precede phonetic processing* (early integration);
3. In order to exploit optimally the *complementarity* between audition and vision, this common representation space is unlikely to be the "dominant" auditory space. Indeed, the complementarity means that audition is not entirely dominant, but rather that *each modality has its own preferred tasks*, though audition is globally the speech modality. Automatic audio-visual speech recognition systems can take advantage of this fact;
4. The fusion must be tuned by some principle related to the *precision* of a given modality for a given decoding task, the precision being probably related to the inverse of the *variance* of a given perceptual estimate. All these facts can provide some guidelines to research on synthesis of speaking faces (Benoît, 1992), by telling us which are the most important parameters of a face that have to be synthesized accurately. Hearing-aid systems should also benefit of a better knowledge of what are the most important acoustic features that need to be presented to the hard of hearing in order to complement speech reading.

The problem of intersensory integration in speech perception remains a challenge for future research, for which closer relationship between specialists of audition, vision and natural language will be particularly welcome.

REFERENCES

Abry, C. & Boë, L.-J. (1986) *Laws for lips*, Speech Communication, 5, 97-104.

Benoît, C. (1992) *The intrinsic bimodality of speech communication and the synthesis of talking faces* J. on Communications of Scientific Soc. of Telecommunications Hungary, 43, 32-40.

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

Fowler, C.A. (1986) *An event approach to the study of speech perception from a direct-realist perspective*, J. of Phonetics, 14, 3-28.

Gibson, J.J. (1966) *The senses considered as perceptual systems*, (Houghton Mifflin C°: Boston).

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

Klatt, D.H. (1979) *Speech perception: A model of acoustic-phonetic analysis and lexical access*, J. of Phonetics, 7, 279-312.

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

Liberman, A. & Mattingly, I. (1985) *The motor theory of speech perception revised*, Cognition, 21, 1-33.

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

(Laurence Erlbaum Associates: London).

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

Robert-Ribes, J., Schwartz, J.L. & Escudier, P. (in press) *Fusion of the auditory and visual sensors in speech perception – A comparison of models on the basis of psychological data and functional arguments*, in press in AI Review – Special Vol: Integration of Natural Language and Vision.

Robert-Ribes, J., Lallouache, T., Escudier, P. & Schwartz, J.L. (1993) *Integrating auditory and visual representations for audiovisual vowel recognition*, Proc. of 3rd European Conference on Speech Communication and Technology (by ESCA), Berlin, Germany, 1753-1756.

Sekiyama, K. (1994) Differences in auditory-visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility, J. Acoust. Soc. Jpn. (E), 15(3), 143-158.

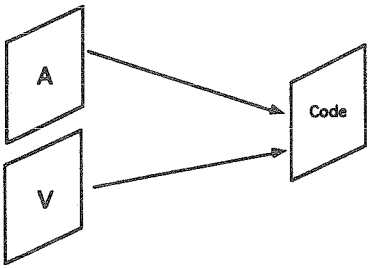
Stein, B.E. & Meredith, M.A. (1993) *The merging of the senses*, (Cambridge: MIT Press).

Summerfield, Q. & MacGrath, M. (1984) *Detection and resolution of audio-visual incompatibility in the perception of vowels*,Quart. J. of Exp. Psycho. : Human Experimental Psychology, 36, 51-74.

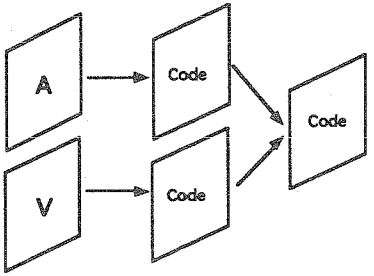
Summerfield, Q. (1987) *Some preliminaries to a comprehensive account of audio-visual speech perception*, in B. Dodd and R. Campbell (Eds.), Hearing by eye (pp 3-51), (Lawrence Erlbaum Associates: London).

Vroomen, J.H.M. (1992) *Hearing voices and seeing lips: Investigations in the psychology of lipreading*, Doctoral dissertation, Katolieke Univ. Brabant.

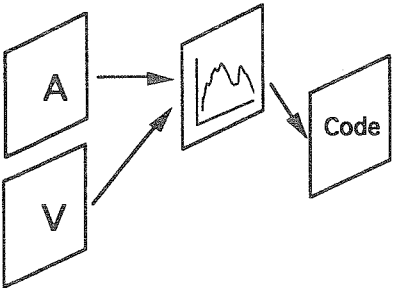
Yuhas, B.P., Goldstein, M.H. & Sejnowski, T.J. (1989) *Integration of Acoustic and Visual Speech Signals Using Neural Networks*, IEEE Communications Magazine, 65-71.



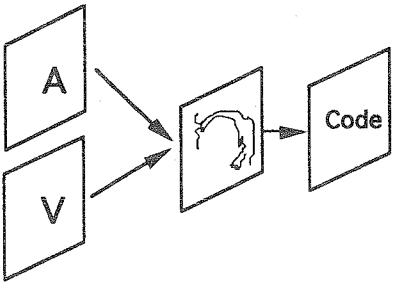
Model DI: direct integration, without common intermediary representation



Model SI: late integration, the common representation is phonetic



Model RD: recoding in the dominant modality



Model RM: recoding int motor space

Fig. 1 Basic structure of models DI, SI, RD and RM

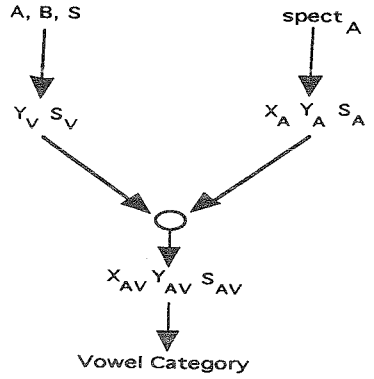


Fig. 2 Schema of model RM (Integration in a "motor space")

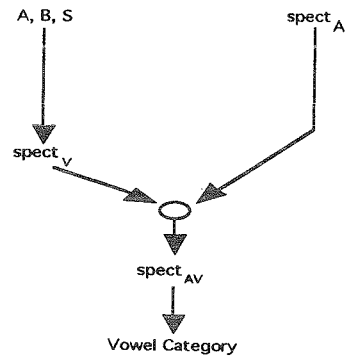


Fig. 3 Schema of model RD (Integration in an "auditory space")

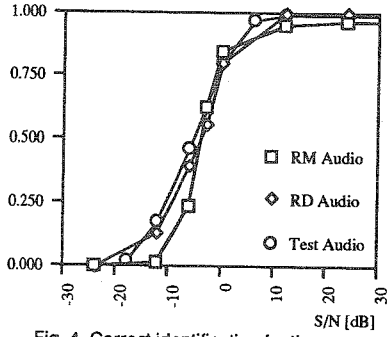


Fig. 4 Correct identification for the auditory-alone path (perception Test and Models RM and RD)

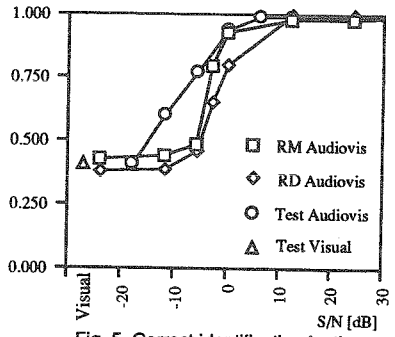


Fig. 5 Correct identification for the audiovisual path (perception Test and Models RM and RD)

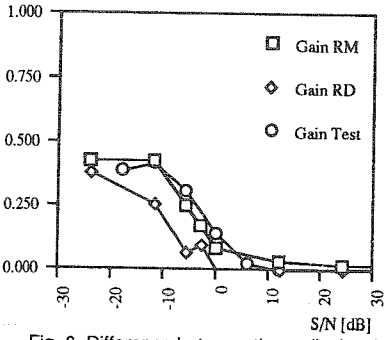


Fig. 6 Difference between the audiovisual and the auditory-alone identifications (perception Test and Models RM and RD)

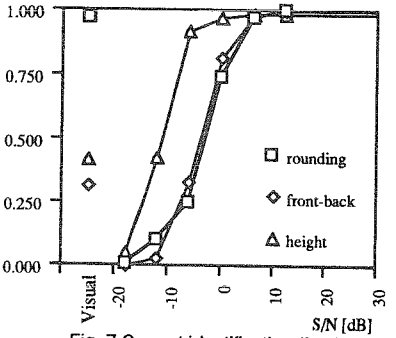


Fig. 7 Correct identification (for the perception Test) for three features with audio stimuli and visual stimuli