NON-LINEAR REPRESENTATIONS, SENSOR RELIABILITY ESTIMATION AND CONTEXT-DEPENDENT FUSION IN THE AUDIOVISUAL RECOGNITION OF SPEECH IN NOISE

Pascal Teissier (1) (2), Jean-Luc Schwartz (1) and Anne Guérin-Dugué (2)

(1) Institut de la Communication Parlée CNRS UPRESA 5009 / INPG - U. Stendhal
ICP, INPG, 46 Av. Félix-Viallet, 38031 Grenoble Cedex 1 / (teissier, schwartz)@icp.grenet.fr
(2) Laboratoire de Traitement d'Images et de Reconnaissance des Formes
LTIRF, INPG, 46 Av. Félix-Viallet, 38031 Grenoble Cedex 1 / guerin@tirf.inpg.fr

ABSTRACT

The paper involves the recognition of French audiovisual vowels at various signal-to-noise ratios (SNRs). It deals with a new non-linear preprocessing of the audio data which enables an estimation of the reliability of the audio sensor in relation to SNR, and a significant increase in the recognition performances at the output of the fusion process.

1. INTRODUCTION

Since several years, a number of automatic speech recognition systems insert a visual input in their system in order to enhance speech identification in acoustical noise [8]. Then the challenge is to obtain the best synergy: the audio-visual recognition rate must be higher than both audio and visual scores separately as it is for human subjects. In order to realize this challenge one must make three important choices. First, various *architectures for decision fusion* have been proposed [7] and their performances must be assessed and compared. Second, the *fusion process* may incorporate an external variable (a "context"), possibly linked to the *reliability* of each sensor [1]. Third, the *representation* of the input data and the nature of the monosensory classifiers may considerably modify the behaviour of the whole system. We discuss these three points in relation to the audio-visual (AV) recognition of vowels embedded in acoustical noise.

2. EXPERIMENTAL CONDITIONS

2.1 Audiovisual data

The corpus consists in 100 repetitions of each of the 10 French oral vowels pronounced in isolation by a single speaker. The corpus was recorded with a Video-Speech Workstation [5] which allows us to automatically extract three basic parameters of the labial contours, namely inner-lip width (A), height (B) and area (S). Noisy acoustical signals were obtained by adding various amounts of white gaussian noise, with 8 signal-to-noise ratios: no noise, 24 dB, 12 dB, 6 dB, 0 dB, -6 dB, -12 dB and -24 dB. The audio coefficients are 20-dimensional spectral values in a perceptual bark-scale (see [6,9] for more details).

2.2 Recognition paradigms

The efficiency of the audiovisual system was assessed in various situations depending on two main factors.

2.2.1 Learning and test corpus

We consider here an "extrapolation" paradigm in which the learning corpus contains acoustical samples at high SNR values (no noise, 24 dB, 12 dB, 0 dB) and the system is then tested with non-learned samples at all SNR values including the lowest ones. We used ten different partitions to increase the reliability of recognition scores estimation.

2.2.2 Introduction of a contextual input

The reliability of each sensor plays a crucial role in sensor fusion. We consider in our work the possibility to directly introduce in the fusion process a control of the acoustical input linked to the SNR value: this is called "extrapolation-with-context" paradigm. In the "extrapolation-without-context" paradigm, we shall discuss the possibility to *infer* the reliability of the acoustical input directly from the data: this is called "extrapolation-with-estimated-context" paradigm.

3. AUDIO-VISUAL FUSION ARCHITECTURES

In other works, we have defined four possible architectures for sensor fusion [7], and studied their performances without preprocessing of the acoustical input [9]. In the present work, where we test the interest of a non-linear preprocessing, we focus on the most classical architecture, the Separate Identification (SI) model. In this model, the acoustical and optical data are separately classified, and then a decision-to-decision fusion process is applied to estimate the audio-visual score.
For each monomodal classifier, we use the quadratic discriminant analysis (gaussian classifier), in which for a

given partition we first estimate means m_i and the covariance matrices V_i for each class ω_i (with 10 classes), and then compute the *posterior* probability $P(\omega_i/x)$ for a given input vector x. Let us call P_A and P_V the probabilities at the output of the Audio and Visual classifier respectively, and P_{AV} the probabilities at the output of the fusion process. When no context is introduced, P_{AV} is computed thanks to a classical multiplicative process. However in the "context" paradigm, we introduce weighting power factors α_i and $(1-\alpha_i)$ selectively reinforcing the weight of the audio or visual decisions in the multiplicative fusion process :

$$P_{AV}(\omega_i/x) = \frac{\left[P_A(\omega_i/x)\right]^{\alpha_i} \cdot \left[P_V(\omega_i/x)\right]^{1-\alpha_i}}{\sum_{j=1}^{10}\left[P_A(\omega_j/x)\right]^{\alpha_j} \cdot \left[P_V(\omega_j/x)\right]^{1-\alpha_j}} \qquad (1)$$

Finally, the classification is based on the choice of *argmax* $[P_{AV}(\omega_i/x)]$.

4. NON-LINEAR PREPROCESSING OF THE AUDIO DATA

A difficulty for audio classification is that in the auditory space, the configurations of the vowel clusters (including all SNRs) are rather folded. The objective of the non-linear preprocessing is to simplify the trajectories of the stimuli when noise increases (trajectory "unfolding") in order to facilitate the classification and the SNR estimation. A 3-D Principal Component Analysis (PCA; linear preprocessing) performed on the corpus (all SNRs) showed us the complexity of the trajectories produced by the deformation of a vowel spectrum with increasing noise (Fig. 1a: each symbol represents the cluster center at different SNR). In the extrapolation condition, these folded trajectories lead to a poor recognition rate in audio and audiovisual conditions. Hence we used an original non-linear projection algorithm, called Curvilinear Component Analysis (CCA), to attempt to "unfold" these trajectories. The CCA principle relies on the idea that two points in the input space *xi* and *xj* with a distance *Xij* between them must be located in the low-dimensional output space with a distance *Yij* as close as possible to *Xij*. However, due to the reduction dimension process, this is not possible for all the range of distances. Then the cost function to minimize for the matching between *Xij* and *Yij* includes a monotonously decreasing weighting function *F(Yij)* such that short-range distances are favored relative to longer-range ones. So, it is possible to re-shape a data structure by unfolding it into an output space of lower dimension. The basic principle is to constraint the transformation to reveal the vowels trajectories due to the noise. For this, the trajectories unfolding is organised in a supervised manner: acoustic data for each level of noise (called a "layer of noise") are sequentially organized from the highest SNR to the lowest SNR so that acoustic data at

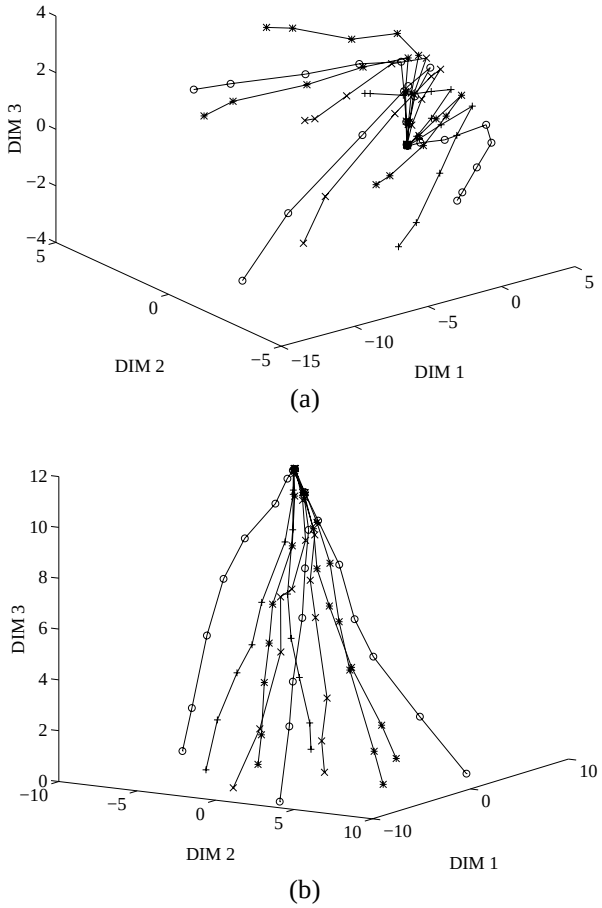
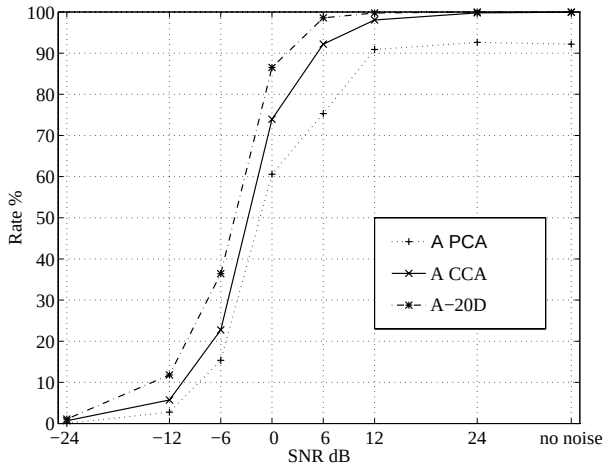
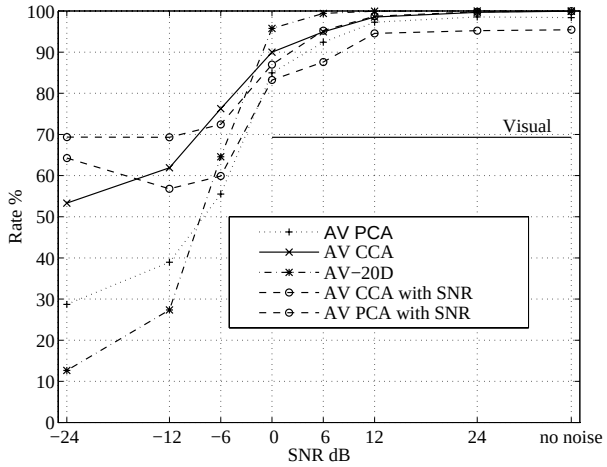


Figure 1 - Representation of the trajectories of the vowel cluster centers for eight SNRs after a 3-D CCA (a) and a 3-D PCA (b) projections

the level SNR = N_i are organized from themselves and from the organization already obtained with the level SNR = N_{i-1}. From the point of view of the output dimension, we know that each layer can be reasonably well unfolded into a 2D (vocalic triangular shape) space. Then if we consider simultaneously two layers, the dimension of the output space must be increased towards (2+1) dimensionnal space. In order to constraint this supplementary dimension to capture the basic shape (triangle) displacement with noise, this dimension is only added at the beginning of the process when the second layer (in our case 24 dB) is organized from the first layer (in our case "no noise"). For the following layers, the dimension of the output space is not further increased and remains at (2+1) (see [2, 3] for more details). Fig. 1b displays the 3D CCA projection of the audio corpus. We clearly see the unfolding of the trajectories and the vowel space shrinkage (from bottom to top) due to noise. We shall see in the next section the potential interest of data unfolding.



(a)



(b)

Figure 2 – Extrapolation-without-context recognition for various frontends (20D, 3D-PCA, 3D-CCA), (a) A classifier (b) AV classifier; "with SNR" means weighted fusion with the SNR estimated from the acoustic input

5. RECOGNITION EXPERIMENTS

5.1 Extrapolation-without-context paradigm

We present on Fig. 2a the recognition at the output of the acoustical classifier: we notice that the 3-D CCA preprocessing leads to performances slightly lower than with the complete 20-D inputs, but better than with a linear 3-D PCA frontend. On Fig.2b, we notice that the AV scores with CCA are much better than with both 3-D PCA and 20-D inputs for small SNRs. The reason is that with CCA, the trajectories are quite straight, and the audio classifier provides *posterior* probabilities not very contrasted in large amount of noise: hence, in the fusion process, the audio decisions play a small part. On the contrary, with PCA or no frontend, the trajectories are more folded, hence the audio classifier makes errors with important *posterior* probabilities, which play a large role in the fusion with the video classifier. In consequence, the audiovisual rate at SNR = -24 dB

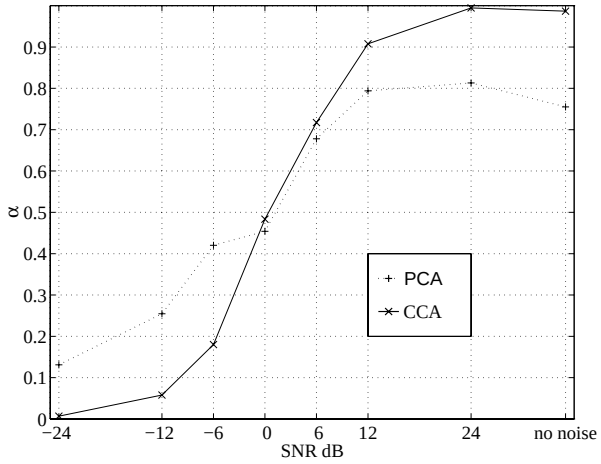


Figure 3 –Evaluation of the average estimation of α (for the 10 vocalic classes) during the test phases vs the SNR values

increases from 13% without data processing, to 29% after PCA up to 53% after CCA, not very far from the video score at 69% (see figure 2.b): some true "extrapolation" occurs thanks to data unfolding.

5.2 Extrapolation-with-estimated-context paradigm

In order to enhance the system performance we attempted to estimate the audio reliability in the extrapolation paradigm (the α_i parameter in Eq.1). We have studied two methods to evaluate the audio reliability [4]: derive it from the ambiguity at the output of the acoustical classifier or from an estimation of SNR by the stimulus position on the vocalic trajectory. "Ambiguity scores" are easy to compute but it may happen that ambiguity is small in a however large level of noise, which makes it a poor candidate for estimating the audio reliability. Hence we prefered the second method. With CCA preprocessing, the SNR estimation can be computed from the last dimension of the CCA (or 3D dimensions of the PCA) representation which is highly correlated with noise thanks to unfolding (see figure 1a). Then, we will set α_i from an estimation of a ratio ζ_i , called "noise factor", normalized between 0 and 1 (see Eq. 2). This factor is linked with the SNR value by equation 3.

$$\zeta_i(x) = \frac{Power(Signal)}{Power(Signal) + Power(Noise)} \qquad (2)$$

$$\zeta_i(x) = \frac{10^{\frac{SNR(x)}{10}}}{1 + 10^{\frac{SNR(x)}{10}}}, x \in \omega_i \qquad (3)$$

The evalutation of the noise factor is realized by a polynomial regression method (2^{nd} order). To account for differences in the vowel trajectories with noise, we consider ten regression modules. This parabolic regression is sufficient to make an estimation of the noise

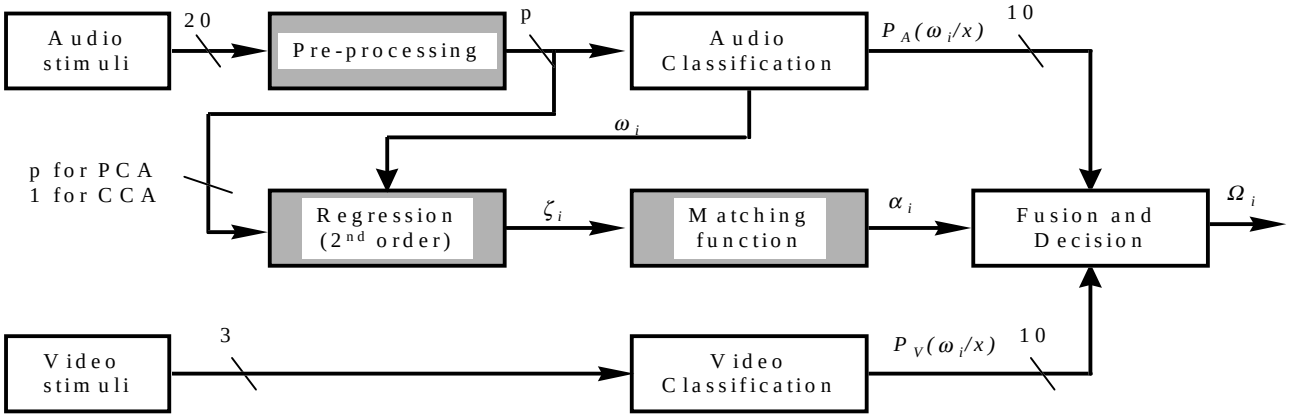


figure 4 - Global architecture with preprocessing of the audio data, context estimation and control of the fusion process

factor $\bar{\zeta}_i$ and derive the weighting factor α_i by a simple matching function with threshold at 0 and saturation at 1. The evaluation of the weighting factor is displayed in Fig. 3. With PCA we clearly see that the estimation (at all the SNRs) is imprecise due to the folded trajectories. Finally, in Fig. 2b, we see that the introduction of a reliability factor in the SI model, derived from SNR estimation in CCA, is very efficient for all the SNRs contrary to PCA because of the poor context estimation. In this recognition paradigm, we have only used the location of the cluster along the trajectories to derive a weighting factor $(\bar{\zeta}_i)$ for the fusion process. Notice that this factor directly linked with the SNR can be also used as an estimate of the SNR for a more general recognition application.

6. CONCLUSION

The global architecture progressively elaborated in this work is displayed in Fig 4. Altogether, CCA preprocessing, SNR estimation and control of the sensor fusion process by the reliability of the audio sensor enable us to realize our challenge : the audiovisual recognition score converges towards the visual recognition score as the SNR decreases and remains always superior or equal to both the visual and the audio scores. Our future work will aim to estimate the context for more complex corpus (dynamic) and noise (coktail party).

7. ACKNOWLEDGMENTS

This work has been realized with the support of the French CNRS-INPG “Fédération de Laboratoires” ELESA

8. REFERENCES

[1] Bloch, I. (1996). Information Combination Operators for Data Fusion: A Comparative review with Classification. *IEEE Trans on SMC*, A, vol 26, 1, 52-67.

[2] Demartines, P. & Hérault, J. (1997). Curvilinear Component Analysis : A Self-Organizing Neural Network for Non Linear Mapping of Data Sets. *IEEE Trans on Neural Networks*, Vol 8, 1, 148-154.

[3] Guérin-Dugué, A., Teissier, P., Schwartz, J.L. & Hérault, J. (1997). Non linear representation for audio-visual fusion in a noisy-vowel recognition task,. *NEURAP'97*,Marseille,31-40.

[4] Guérin-Dugué, A., Teissier, P., Schwartz, J.L. & Hérault, J. (1997). Constrained neural network for estimating sensor reliability in sensors fusion, *IWANN'97*, to appear.

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

[6] Robert-Ribes, J., Schwartz, J.L. & Escudier, P. (1995). A comparison of models for fusion of auditory and visual sensors in speech perception. *Artificial Intelligence Review Journal*, 9, 323-346.

[7] Schwartz, J.L., Robert-Ribes, J., & Escudier, P. (In press). Ten years after Summerfield... a taxanomy 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.* Erlbaum / Psychology Press.

[8] Stork, D.G. & Hennecke, M.E. (1996). (Eds.) *Speechreading by Man and Machine: Models, Systems and Applications*. NATO ASI Series, Springer.

[9] Teissier, P., Robert-Ribes, J., Schwartz, J.L. & Guérin-Dugué, A. (1997). Comparing models for audiovisual fusion in a noisy-vowel recognition task. Submitted to *IEEE Trans. Speech and Audio Processing.*