# Projections non linéaires, estimation de la fiabilité d'un capteur et fusion dépendante du contexte pour la reconnaissance audiovisuelle de parole dans le bruit

**Pascal Teissier** (1) (2)**, Jean-Luc Schwartz** (1) **et 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

RÉSUMÉ

Si on veut exploiter pleinement la complémentarité des capteurs auditif et visuel et réaliser un système de reconnaissance audiovisuelle de la parole robuste en milieu bruité, il faut introduire au cœur du processus de fusion de capteurs une information contextuelle portant sur la fiabilité des capteurs. Nous montrons ici comment estimer efficacement la fiabilité du capteur acoustique par une "Analyse en Composantes Curvilinéaires" (ACC), et comment introduire cette estimation de contexte dans le processus de fusion dans deux architectures de reconnaissance.

ABSTRACT

If we want to exploit fully the auditory and visual complementarity, and realize an audiovisual speech recognition system robust in noise, we have to insert in the sensor fusion process a contextual information about the sensors reliability. Here, we show how efficiently estimate the acoustic sensor reliability thanks to a une "Curvilinear Component Analysis" (CCA), and the way of insert this context estimated into the fusion process in two recognition architectures.

## 1 Introduction

Depuis plusieurs années, un certain nombre de systèmes de reconnaissance automatique de la parole intègrent une entrée visuelle de façon à améliorer l'identification de la parole dans le bruit acoustique [9]. Le challenge consiste alors à obtenir la meilleure synergie : le taux de reconnaissance audiovisuelle doit être plus élevé que les scores auditif et visuel séparément comme cela est constaté pour des sujets humains. Pour réaliser ce challenge, trop rarement atteint, il est nécessaire d'introduire une information (contexte) afin de guider le système vers la modalité la plus efficace. Le problème est alors le suivant : comment estimer la fiabilité d'un capteur (contexte) et comment introduire cette information dans le système de fusion ?

Le présent travail porte sur l'estimation et l'introduction de contexte sur deux modèles de fusion audiovisuelle pour une application à la reconnaissance de voyelles audiovisuelles du Français à différents rapports signal sur bruit (RSB).

## 2 Conditions expérimentales

### 2.1 Données audiovisuelles

Le corpus comporte 100 répétitions de chacune des 10 voyelles orales du français [i, e, a, u, y, o, œ, ε, χ, ⏐ ] prononcées isolément par un seul locuteur. Le corpus a été enregistré avec un poste "visage parole" [5] qui nous permet d'extraire trois paramètres caractéristiques du contour labial, l'étirement (A), la hauteur (B) et la surface (S) du contour interne des lèvres qui fournissent les trois entrées vidéo. Les signaux acoustiques bruités sont obtenus en additionnant un taux variable de bruit blanc gaussien avec 8 rapports signal sur bruit : sans bruit, 24 dB, 12 dB, 6 dB, 0 dB, -6dB, -12dB et –24 dB. Les entrées audio sont 20 coefficients spectraux en sortie d'une analyse spectrale utilisant une échelle perceptuelle en bark pour les fréquences (voir [10] pour plus de détails).

### 2.2 Corpus d'apprentissage et de test

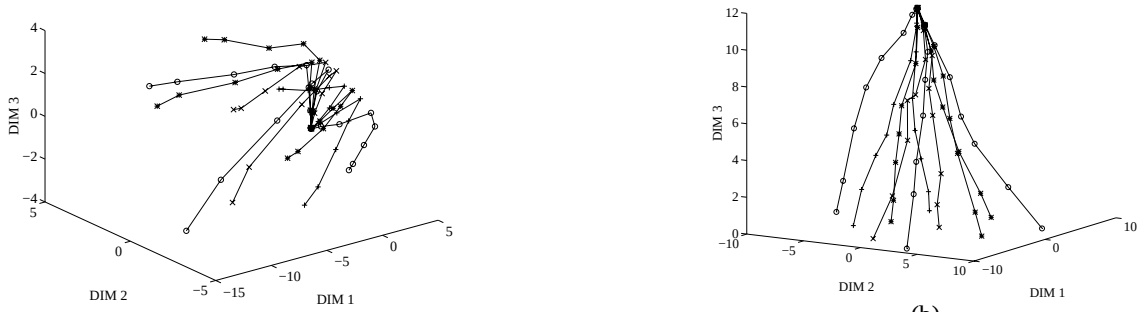
On considère ici le paradigme "d'extrapolation" pour lequel le corpus d'apprentissage contient uniquement les échantillons à fort RSB (sans bruit, 24 dB, 12 dB, 0 dB) et le système est ensuite testé avec les échantillons non appris sur tous les RSB. Nous avons utilisé dix partitions différentes de façon à améliorer la qualité des estimations des scores du système.

## 3 Architectures de fusion audiovisuelle

Dans des travaux précédents, nous avons défini quatre architectures possibles pour la fusion de capteurs auditif et visuel [8] et montré [10] comment introduire dans chacune de ces architectures des principes de fusion de divers types, selon une taxinomie proposée par Bloch [2]. Nous nous intéressons ici aux deux architectures les plus classiques, le modèle à Identification Séparée (IS) et le modèle à Identification Directe (ID), avec dans chaque cas une fusion dépendante du contexte.

### 3.1 Modèle à Identification Séparée (IS)

Dans ce modèle, les données acoustiques et visuelles sont identifiées séparément, puis on applique un processus de fusion de décisions pour estimer le score audiovisuel. Pour chaque classifieur monomodal, nous utilisons l'analyse




(a) (b)

**Figure 1 –** Représentation des trajectoires des centres de classe pour les 8 RSB à 3 dimensions après une projection par ACP (a) et ACC (b)

discriminante quadratique (classifieur gaussien) : pour une partition donnée nous estimons la moyenne $m_i$ et la matrice de covariance $V_i$ pour chaque classe $\omega_i$ (avec 10 classes), puis nous calculons la probabilité *a-posteriori* $P(\omega_i/x)$ pour un vecteur d'entrée $x$ donné. On appelle respectivement $P_A$ et $P_V$ les probabilités à la sortie du classifieur auditif et visuel, et $P_{AV}$ la probabilité à la sortie du processus de fusion. Lorsque le contexte n'est pas introduit $P_{AV}$ est calculé par un processus classique de multiplication. On introduit le contexte par le biais de facteurs de pondération exponentiels $\alpha$ et $(1-\alpha)$ qui renforcent sélectivement le poids des décisions auditive et visuelle dans le processus de fusion multiplicative :

$$P_{AV}(\omega_i / x) = \frac{[P_A(\omega_i / x)]^{\alpha} \cdot [P_V(\omega_i / x)]^{1-\alpha}}{\sum_{i=1}^{10} [P_A(\omega_i / x)]^{\alpha} \cdot [P_V(\omega_i / x)]^{1-\alpha}} \qquad (3.1)$$

La décision est fournie par *argmax* $P_{AV}(\omega_i/x)$. Le principe d'estimation du contexte $\alpha$ sera précisé dans la section 5.

### 3.2 Modèle à Identification Directe (ID)

Dans ce modèle on concatène les données acoustiques et visuelles : on a donc un classifieur bimodal. Pour ce classifieur, nous utilisons également l'analyse discriminante quadratique. On introduit le contexte par le biais d'un facteur de pondération $\alpha$ (respectivement $1-\alpha$) qui contrôle le poids de la composante acoustique (respectivement visuelle) en fonction du RSB, selon la formule :

$$P_{AV}(\omega_i / x) \propto \exp\left(-0.5[W(x-m_i)]^T V_i^{-1}[W(x-m_i)]\right) \Big/ \sqrt{|V_i|}$$
$$\text{avec } W = \begin{bmatrix} \alpha \mathbf{I}_{nA} & 0 \\ 0 & (1-\alpha)\mathbf{I}_{nV} \end{bmatrix} \qquad (3.2)$$

où $I_n$ est la matrice identité n x n, $n_A$ ($n_V$) correspond au nombre de dimension auditive (visuelle). En augmentant $\alpha$, on augmente la contribution acoustique et diminue la contribution visuelle à la distance entre l'entrée et la moyenne de chaque classe, tandis qu'en diminuant $\alpha$ on a l'effet inverse. Des valeurs de $\alpha$ respectivement 0 et 1 conduisent à des performances proches de celles d'un classifieur gaussien appliqué aux données respectivement auditives ou visuelles[1].

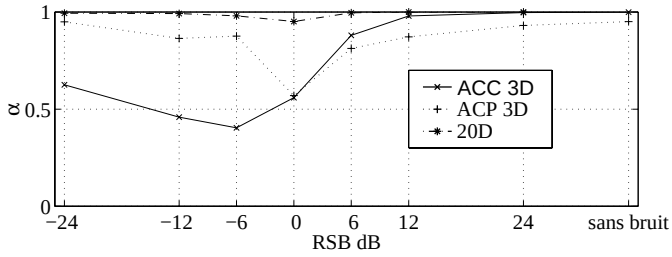
## 4. Prétraitement non linéaire des données auditives

Une difficulté pour le classifieur audio et l'estimation de la fiabilité de ce classifieur est que dans l'espace auditif, les trajectoires produites par la variation des spectres vocaliques dans une catégorie donnée lorsque le bruit augmente sont très complexes. Une Analyse en Composantes Principales (ACP, prétraitement linéaire) réalisée sur tout le corpus confirme la complexité des trajectoires produites par la déformation du spectre de voyelle quand le bruit augmente (voir figure 1.a, chaque symbole représente le centre d'une classe à différents RSB). Ceci rend difficile l'estimation d'un contexte et amène à des scores de reconnaissance auditive et audiovisuelle très faibles [4]. L'objectif du prétraitement non-linéaire est de simplifier les trajectoires des stimuli quand le bruit augmente (trajectoire "dépliée") de façon à faciliter la classification et l'estimation du contexte. Pour cela, nous utilisons une transformation non linéaire, réalisée par un réseau de neurones artificiels appelée "Analyse en Composantes Curvilinéaires" (ACC, voir [3][4] pour plus de détails) pour essayer de déplier ces trajectoires. La figure 1.b nous montre la projection en 3D du corpus audio en sortie de l'ACC. Nous voyons clairement le dépliage des trajectoires et le rétrécissement de l'espace des voyelles (de bas en haut) dû au bruit. Au cours de travaux précédents, nous avons montré l'intérêt du dépliage des données pour la classification [4] ; nous allons voir dans la prochaine section l'intérêt potentiel du dépliage pour l'estimation d'un contexte.

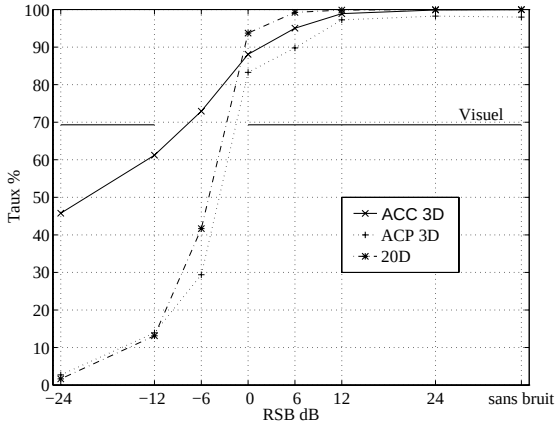
## 5 Estimation du contexte

L'idée de l'introduction de paramètres de contrôle du processus de fusion reliés à la fiabilité des capteurs auditif et visuel (contexte) a fait son apparition récemment dans la communauté de la reconnaissance audiovisuelle de la parole avec deux voies possibles : contexte estimé en sortie des processus de classification de chaque capteur [1][7] ou directement à partir d'une estimation du RSB sur les données d'entrée [6]. Nous avons exploré ces deux voies, en testant dans chaque cas l'intérêt du dépliage des données acoustiques

[1] Deux facteurs causent un légère différence par rapport aux performances monomodales : d'une part, les termes croisés entre composantes auditives et visuelles, d'autre part les différences $\sqrt{|V_i|}$ pour les classifieurs Audio, Visuel et Audiovisuel.




(a)




(b)

**Figure 2 –** (a) Evolution de l'estimation moyenne de α à partir de l'ambiguïté (pour les dix classes vocaliques) pendant la phase de test vs RSB pour 3 cas (ACP 3D, ACC 3D et sans prétraitement 20D). (b) Evolution du score de reconnaissance audiovisuelle avec une fusion dépendante du contexte pour 3 cas

pour l'estimation de la fiabilité du capteur auditif.

## 5.1 Estimation en sortie du classifieur

Un classifieur fournit une estimation de la conformité d'un stimulus à un ensemble d'apprentissage : ceci peut servir de point d'entrée à l'estimation du contexte, par diverses mesures (entropie, probabilités *a-priori*, ambiguïté). L'estimation du contexte selon la probabilité *a-priori* d'observer un stimulus étant donnée une classe gagnante (distance du stimulus à la moyenne de la classe, pondérée par la matrice de covariance) s'avère inutilisable dans notre cas. En effet, notre modélisation gaussienne produit des effets de queue de distribution symétriques, qui peuvent conduire à des estimations de contexte similaires à très fort et très faible bruit. L'utilisation de "l'ambiguïté" consiste à calculer une information sur le contraste des probabilités $P_A$ en sortie du capteur audio (cette méthode ne peut donc s'appliquer que dans le modèle IS) : si la décision du capteur audio est ambiguë, on met plus de poids sur le capteur visuel. Pour calculer l'ambiguïté du capteur audio, nous avons utilisé la formule suivante :

$$\alpha(x) = \frac{P_A(\omega_i / x) - P_A(\omega_j / x)}{P_A(\omega_i / x)} \qquad (5.1)$$

où $\omega_i$ est la classe gagnante et $\omega_j$ est la seconde classe gagnante du classifeur audio. Nous présentons sur la figure 2 l'estimation du contexte par ambiguïté et les scores de reconnaissance audiovisuelle après introduction de cette estimation pour le modèle IS selon 3 cas : sans prétraitement (dimension audio $n_a$=20), avec prétraitement linéaire (par ACP, $n_a$=3) et enfin avec prétraitement non linéaire (par ACC, $n_a$=3). On remarque que l'estimation du contexte avec ou sans prétraitement linéaire est catastrophique. Ceci est dû à la complexité des trajectoires des voyelles à travers le bruit : à fort niveau de bruit, l'une des classes audio est souvent gagnante avec une très faible ambiguïté. Les scores audiovisuels après introduction du contexte estimé sont alors aussi très mauvais. Dans le cas du prétraitement non linéaire, ce problème est largement diminué : l'estimation du contexte est beaucoup mieux reliée au RSB, grâce à une meilleure modélisation des données due au dépliage des trajectoires. La dynamique de cette estimation est néanmoins trop faible, elle ne permet donc pas d'obtenir notre challenge : le score audiovisuel reste inférieur au score visuel à fort bruit.

## 5.2 Estimation en entrée du classifieur

La position d'un échantillon sur la trajectoire des voyelles en fonction du bruit nous donne une information sur le RSB (donc sur la fiabilité du capteur audio). Dans le cas de prétraitement non linéaire, l'estimation du RSB est calculée à partir de la dernière dimension qui est fortement corrélée avec le bruit grâce au dépliage (voir figure 1.b). Nous définissons d'abord un facteur de bruit $\zeta$ , compris entre 0 et 1 :

$$\zeta(x) = \frac{Power(Signal)}{Power(Signal) + Power(Noise)} \qquad (5.2)$$

$$\zeta(x) = \frac{10^{\frac{SNR(x)}{10}}}{1 + 10^{\frac{SNR(x)}{10}}} \qquad (5.3)$$

Puis nous apprenons sur les 4 niveaux de bruit du corpus d'apprentissage, une régression linéaire (pour une entrée audio 20D) ou quadratique (pour une entrée ACP 3D ou ACC 3D) entre l'entrée et $\zeta$ , pour chaque catégorie. Dans la phase de test, l'estimation de $\zeta$ , $\hat{\zeta}$ , est réalisée une fois déterminée la classe audio gagnante, puis $\alpha$ est déduit par mise en forme avec seuillage à 0 et saturation à 1.

L'évolution du facteur de pondération est décrite sur la figure 3. Avec ou sans prétraitement linéaire, on peut clairement remarquer que l'estimation du contexte n'est pas précise à cause de la complexité des trajectoires. Par contre, lorsque les trajectoires sont dépliées à l'aide du prétraitement non linéaire, l'estimation du contexte est tout à fait correcte. Les figures 4.a et 4.b nous indiquent les scores de reconnaissance audiovisuelle après introduction du contexte estimé pour le modèle ID et pour le modèle IS. Avec ou sans prétraitement linéaire, les scores ne sont pas satisfaisants à cause d'une mauvaise estimation du contexte : le poids du capteur vidéo est trop important à faible RSB et trop faible à fort RSB. Avec un prétraitement non linéaire, la bonne estimation du contexte permet une reconnaissance audiovisuelle efficace. Pour le modèle ID, on remarque que le score audiovisuel est légèrement inférieur au score visuel à fort niveau de bruit : ceci est principalement dû à des

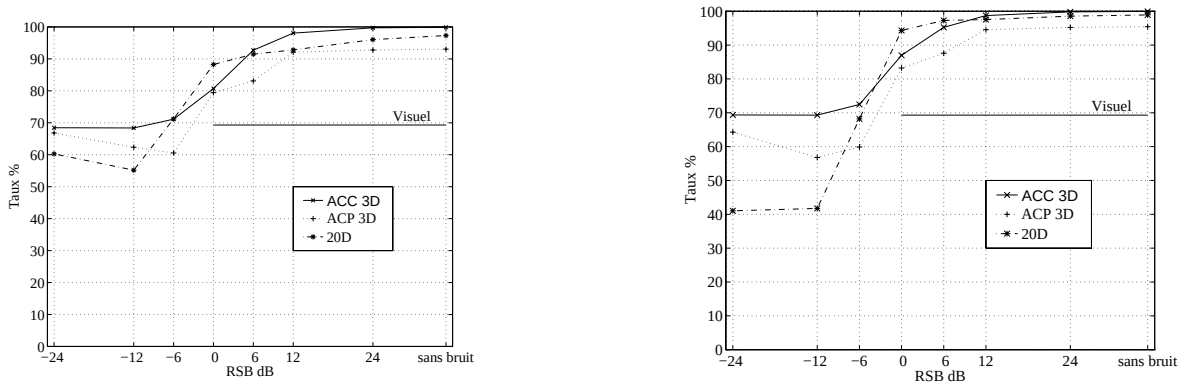



**Figure 4 –** Evolution du score de reconnaissance audiovisuelle avec une fusion dépendante du contexte pour 3 cas (ACC 3D, ACP 3D et sans prétraitement 20D) pour le modèle ID (a) et IS (b)

problèmes techniques d'introduction du contexte dans ce modèle (voir note 1 dans la section 3.1). En revanche, pour le modèle IS, l'objectif est atteint : le score de reconnaissance audiovisuelle converge vers le score de reconnaissance visuelle quand le RSB diminue et il reste supérieur ou égal aux scores visuel et auditif dans toute la dynamique de RSB.

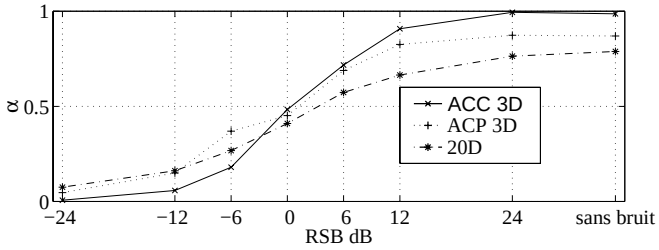



**Figure 3 –** Evolution de l'estimation moyenne de α à partir du facteur de bruit (pour les dix classes vocaliques) pendant la phase de test vs RSB pour 3 cas (ACC 3D, ACP 3D et sans prétraitement 20D).

# 6 Discussion et conclusion

Nous avons présenté deux façons d'estimer le contexte afin de prendre en compte la fiabilité du capteur acoustique pour une application à la reconnaissance audiovisuelle. La première technique est basée sur l'information d'ambiguïté de la décision du capteur acoustique, elle est très facile à implémenter mais la difficulté est d'adapter l'évolution de ce paramètre pour une pondération efficace de toutes les décisions à tous les niveaux de bruit. La deuxième technique exploite la structure intrinsèque des données. Pour cela, on utilise les trajectoires des échantillons dans une configuration bruitée. Pour ces deux techniques, le prétraitement par ACC est bien adapté puisqu'il permet une représentation quasi optimale en révélant la structure intrinsèque des données. Cette représentation facilite donc la modélisation des données et permet une estimation du contexte, et une reconnaissance audiovisuelle efficace par rapport à un prétraitement linéaire. En résumé, prétraitement non linéaire, estimation de la fiabilité du capteur auditif et fusion dépendante du contexte nous permettent de réaliser un système de reconnaissance audiovisuelle efficace à tous les niveaux de bruit.

# 7 Remerciements

Ce travail est supporté par la Fédération de laboratoires ELESA du CNRS-INPG.

# 8 Références

[1] Adjoudani, A. & Benoît, C. *"On the integration of auditory and visual parameters in an HMM-based ASR"*, In Stork, D.G.& Hennecke, M.E. (Eds.) Speechreading by Man and Machine: Models, Systems and Applications. NATO ASI Series, Springer, pp. 461-472, 1996.

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

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

[4] Guérin-Dugué, A. et al. *"Non linear representation for audio-visual fusion in a noisy-vowel recognition task"*, NEURAP'97, Marseille, pp. 31-40, March 1997.

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

[6] Meier, U. et al. *"Adaptive bimodal sensor fusion for automatic speechreading"*, ICAPSS'96, 1996.

[7] Movellan, J.R. & Mineiro, P. *"Modularity and catastrophic fusion : a bayesian approach with application to audiovisual specch recognition"*, Technical Report CogSci. USCD-97.01, Department of Cognitive Science, USCD, San Diego, CA, 92093-0515, 1996.

[8] Schwartz, J.L. et al. (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 aspectes of language processing. Erlbaum/Psychology Press.

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

[10] Teissier, P et al. *"Comparing models for audiovisual fusion in a noisy-vowel recognition task"*, Submitted to *IEEE Trans. Speech and Audio Processing*, 1997.