Séparation de sources audio-visuelles : Formalisation et expérimentation

D. Sodoyer [1], L. Girin [1], C.Jutten [2], J.L. Schwartz [1]

[1]Institut de la Communication Parlée – INPG/Univ. Stendhal/CNRS
46, av Felix Viallet, 38031 Grenoble Cedex 1, France
Tél.: ++33 (0)476 57 47 12 - Fax: ++33 (0)476 57 47 10
Mél: sodoyer@icp.inpg.fr - http://www.icp.inpg.fr
[2]Laboratoire des Images et des Signaux – INPG/UJF/CNRS

ABSTRACT

In this paper, we present a new approach to the source separation problem in the case of multiple speech signals. The method is based on the use of automatic lipreading: the objective is to extract an acoustic speech input from other acoustic signals by exploiting its coherence with the speaker's lips movements. We consider the case of an additive stationary mixture. Firstly we present a theoretical framework showing that it is indeed possible to separate a source when some of its spectral characteristics are provided to the system. Then we address the case of audio-visual sources. We show how, if a statistical model of the joint probability of visual and spectral audio input is learnt to quantify the audio-visual coherence, separation can be achieved by maximising this probability.

1. INTRODUCTION

De nombreux travaux de recherche montrent l'existence d'une cohérence intrinsèque et même d'une complémentarité entre l'audition et la vision pour la perception de la parole [Sum87]. La contribution de la vision dans la perception de la parole bruitée permet d'obtenir un gain d'intelligibilité notable entre les conditions « audio seul » et « audio + visage du locuteur » [Ben94], [Rob98].

Dans leur étude, [Gir01] montrent qu'il est possible de réaliser un système de rehaussement de la parole dans un bruit blanc par filtrage utilisant l'image du locuteur. Dans la lignée de ce travail, nous cherchons maintenant à appliquer cette idée dans le cas d'un mélange de plusieurs sources de parole; le but étant de séparer l'une des sources. Ceci revient en fait à un problème de séparation de sources, problème qui a déjà suscité de nombreuses études en traitement du signal avec un certain nombre d'applications pour la parole. L'enjeu de ce travail est d'exploiter la cohérence entre l'image et le son de la parole. Nous aborderons ce principe en Section 2, les résultats étant exposés en Section 3. Puis en Section 4, nous présenterons quelques perspectives d'améliorations possibles afin de conclure.

2. MODELE

2.1 Présentation du problème

Nous considérons dans notre problème un système possédant N sources inconnues et localement stationnaires $\boldsymbol{s_i(t)}$. La multiplication de ces dernières avec une matrice (P,N) de mélange $\boldsymbol{A}$ fournit P observations $\boldsymbol{x_k(t)}$: $\boldsymbol{x = As}$. L'opération de séparation consiste à estimer la matrice (N,P) de séparation $\boldsymbol{B}$ dont l'opération $\boldsymbol{y = Bx}$ fournit N estimées $\boldsymbol{y_j(t)}$ de $\boldsymbol{s_i(t)}$ (figure 1). Nous prendrons le cas où $N = P$.

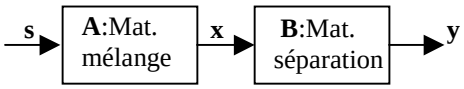

Figure 1: Schéma de la séparation de sources.

Dans les méthodes classiques d'ACI (Analyse en Composantes Indépendantes) la matrice de séparation $\boldsymbol{B}$ est estimée selon un critère de maximisation d'indépendance entre les sorties $\boldsymbol{y_j(t)}$. Ici nous utilisons une information supplémentaire $\boldsymbol{V_1}$ correspondant aux dimensions géométriques des lèvres d'un locuteur fournies par la vidéo synchronisée avec son signal acoustique $\boldsymbol{s_1}$, signal que nous chercherons à extraire d'un mélange de deux ou plusieurs sources (figure 2).

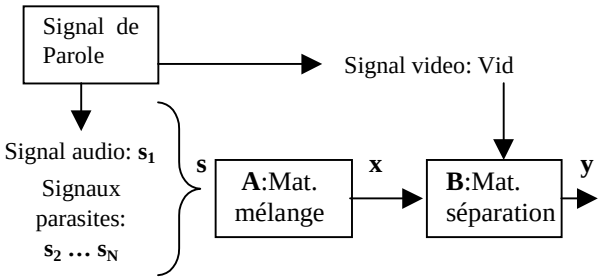

Figure 2 : Schéma de principe de la séparation de sources audio-visuelles.

2.2 Principe théorique

Dans notre étude, nous supposons simplement la décorrélation des sources, et une connaissance supplémentaire sur la source $\boldsymbol{s_1}$ que l'on désire extraire du mélange. Cette information supplémentaire correspond au signal visuel (associé à $\boldsymbol{s_1}$) qui est approximativement lié partiellement au spectre de la fonction de transfert du conduit vocal, donc à

l'enveloppe spectrale du signal $\boldsymbol{s_1}$. Nous pouvons donc associer à cette information spectrale, soit des coefficients d'autocorrélation, soit des coefficients d'énergie résultant d'un banc de filtre.

Critère d'autocorrélation

Nous supposons dans cette sous partie être capable de connaître une valeur d'autocorrélation du signal $\boldsymbol{s_1}$ pour un délai k donné. En normalisant cette valeur d'autocorrélation par l'énergie de $\boldsymbol{s_1}$ nous posons comme coefficient d'autocorrélation :

$$\gamma_k = \frac{R_{s_1 s_1}(k)}{R_{s_1 s_1}(0)} \qquad (1)$$

Si, à la séparation, nous voulons que $\boldsymbol{y_1}$ soit une estimée de $\boldsymbol{s_1}$, nous devons obtenir :

$$\frac{R_{y_1 y_1}(k)}{R_{y_1 y_1}(0)} = \gamma_k \qquad (2)$$

Ainsi nous proposons de minimiser le critère :

$$f_{AC}(\boldsymbol{y}) = C_k(\boldsymbol{y_1 y_1})^2 \qquad (3)$$

avec : $$C_k(\boldsymbol{y_i y_j}) = R_{y_i y_j}(k) - \gamma_k R_{y_i y_j}(0) \qquad (4)$$

Ce critère a la propriété d'être positif ou nul, et minimum (= 0) quand la séparation est achevée ($\boldsymbol{y_1}=\lambda\boldsymbol{s_1}$). Cependant nous devons vérifier que ce critère ne s'annule qu'à la séparation et non dans d'autres configurations. Pour ceci, nous introduisons la matrice $\boldsymbol{G=BA}$ ce qui nous permet de poser :

$$\boldsymbol{y_p} = \sum_n g_{pn} \boldsymbol{s_n} \qquad (5)$$

La question est de savoir si l'annulation du critère annule tous les g_{1n} pour $n \geq 2$. En introduisant (5) dans l'expression de l'autocorrélation nous obtenons :

$$R_{y_1 y_1}(k) = \sum_{m,n} g_{1m} g_{1n} R_{s_m s_n}(k) \qquad (6)$$

Sous l'hypothèse que les sources sont décorrélées :

$$R_{y_1 y_1}(k) = \sum_n g_{1n}^2 R_{s_n s_n}(k) \qquad (7)$$

En introduisant (7) dans (3), il faut pour annuler $f_{AC}(\boldsymbol{y})$:

$$C_k(\boldsymbol{y_1 y_1}) = \sum_n g_{1n}^2 C_k(\boldsymbol{s_n s_n}) = 0 \qquad (8)$$

Par définition $C_k(\boldsymbol{s_1 s_1})=0$, ainsi :

$$C_k(\boldsymbol{y_1 y_1}) = \sum_{n \geq 2} g_{1n}^2 C_k(\boldsymbol{s_n s_n}) = 0 \qquad (9)$$

Dans le cas de 2 sources, en supposant que $C_k(\boldsymbol{s_2 s_2})$ n'est pas nul ,c'est à dire que les 2 sources n'ont pas la même propriété spectrale définit en (1), (9) implique que $g_{12}=0$ et ainsi nous obtenons $\boldsymbol{y_1}=g_{11}\boldsymbol{s_1}$. Mais pour plus de 2 sources, l'éq. (9) n'assure pas l'annulation des g_{1n} $n \geq 2$. Pour N sources il faut connaître (N-1) valeurs de γ_k définies en (1) par (N-1) délais k différents. Nous posons alors un nouveau critère $f_{AC}(\boldsymbol{y})$:

$$f_{AC}(\boldsymbol{y}) = \sum_{k=1}^{N-1} C_k(\boldsymbol{y_1 y_1})^2 \qquad (10)$$

Ici $f_{AC}(\boldsymbol{y})$ est égal à zéro si et seulement si :

$$C_k(\boldsymbol{y_1 y_1}) = 0 \quad \forall k \in \{1 \ldots (N-1)\} \qquad (11)$$

Introduisant (9) dans (11) :

$$\begin{bmatrix} C_1(\boldsymbol{s_2 s_2}) & \cdots & C_1(\boldsymbol{s_n s_n}) & \cdots & C_1(\boldsymbol{s_N s_N}) \\ C_k(\boldsymbol{s_2 s_2}) & & C_k(\boldsymbol{s_n s_n}) & & C_k(\boldsymbol{s_N s_N}) \\ C_{N-1}(\boldsymbol{s_2 s_2}) & & C_{N-1}(\boldsymbol{s_n s_n}) & & C_{N-1}(\boldsymbol{s_N s_N}) \end{bmatrix} \begin{bmatrix} g_{12}^2 \\ g_{1n}^2 \\ g_{1N}^2 \end{bmatrix} = 0 \qquad (12)$$

Si la matrice $\boldsymbol{C}\{C_k(\boldsymbol{s_n s_n})\}$ n'est pas singulière, cela implique que tous les g_{1n} de 2 à N sont nuls et ainsi $\boldsymbol{y_1}=g_{11}\boldsymbol{s_1}$. La non-singularité traduit l'hypothèse que les sources ont des propriétés spectrales différentes.

Critère Spectral

De la même manière, nous pouvons supposer que nous connaissons un coefficient spectral fourni par un banc de K filtres passe-bandes. Posons $\boldsymbol{H_k(f)}$ la réponse fréquentielle du filtre passe-bande RIF de canal spectral k ($1 \leq k \leq K$) et $\boldsymbol{h_k(t)}$ sa réponse impulsionnelle. L'énergie de la source $\boldsymbol{s_1}$ à la sortie du processus de filtrage est fournie par l'autocorrélation de retard nul du signal filtré $\boldsymbol{h_k * s_1(t)}$. Nous pouvons alors supposer connaître l'énergie normalisée du signal $\boldsymbol{s_1}$ dans un certain canal spectral:

$$\gamma_k' = \frac{R_{(h_k * s_1)(h_k * s_1)}(0)}{R_{s_1 s_1}(0)} \qquad (13)$$

Comme dans le cas précédent, nous posons la fonction:

$$C_k'(\boldsymbol{y_i y_j}) = R_{(h_k * y_i)(h_k * y_j)}(0) - \gamma_k' R_{y_i y_j}(0) \qquad (14)$$

Puis comme en (10), nous posons comme critère possible :

$$f_{SC}(\boldsymbol{y}) = \sum_{k=1}^{N-1} C_k'(\boldsymbol{y_1 y_1})^2 \qquad (15)$$

Ce critère, fondé sur la connaissance de (N-1) canaux spectraux , se comporte comme le précédent et conduit à la même équation (12).

En résumé, cette analyse théorique montre qu'il faut *(N-1)* connaissances spectrales (coefficients d'autocorrélation ou coefficients spectraux) du signal $\boldsymbol{s_1}$ afin de pouvoir l'extraire d'un mélange additif à N sources. Notre étude s'est orientée sur l'extraction d'une seule source $\boldsymbol{s_1}$. Pour la minimisation de nos critères [Sod01] nous avons utilisé la technique du gradient relatif introduisant la notion d'équivariance [Car96]. L'extraction d'autres sources supposera que l'on connaisse des informations spectrales du même

type pour chacune d'elles, ou dans le but d'une séparation totale, qu'il sera possible d'utiliser l'indépendance.

Dans notre application nous ne possédons en fait que des approximations de ces connaissances spectrales, celles-ci étant estimées par l'image des lèvres associées au signal de parole $\boldsymbol{s_1}$. La section suivante a pour objet de faire le lien entre cette application et l'analyse théorique qui vient d'être vue.

2.3 Méthode Audio-visuelle

Sachant qu'il existe une relation statistique entre les dimensions labiales et le spectre du signal de parole d'un locuteur [Yeh98], nous construisons un modèle statistique audio-visuel $p(\boldsymbol{S,V})$ comme étant une probabilité conjointe du vecteur visuel $\boldsymbol{V}$ (contenant des dimensions labiales) et d'un vecteur $\boldsymbol{S}$ représentant le spectre du signal audio. Nous définissons cette probabilité conjointe comme une somme de gaussiennes estimée par un algorithme Expectation-Maximisation (EM) itératif exploitant un corpus d'apprentissage afin de déterminer la moyenne et la matrice de covariance de chaque gaussienne.

L'algorithme de séparation consiste à trouver une matrice $\boldsymbol{B}$ pour laquelle la sortie $\boldsymbol{y_1}$ fournit un vecteur spectral le plus cohérent possible avec la vidéo $\boldsymbol{V_1}$ associée à $\boldsymbol{s_1}$. Il faut donc maximiser la probabilité audio-visuelle $p(\boldsymbol{Y_1,V_1})$ ou minimiser le critère:

$$f_{AV}(\boldsymbol{y}) = -\mathrm{Log}(p(\boldsymbol{Y_1,V_1})) \qquad (16)$$

Afin de faire le lien entre les eq. (15) et (17), considérons le cas où $p(\boldsymbol{Y_1,V_1})$ possède une seule gaussienne centrée

$$p(\boldsymbol{Y_1,V_1}) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{C})}} \exp\left[-\frac{1}{2}[\boldsymbol{Y_1 V_1}]^t \boldsymbol{C}^{-1}[\boldsymbol{Y_1 V_1}]\right] \qquad (17)$$

où $\boldsymbol{Y_1}$ est un vecteur spectral défini comme en Sect. 2.1 par K valeurs d'énergies fournies par un banc de filtre à K bandes passantes, $\boldsymbol{V_1}$ est un vecteur contenant des dimensions labiales, et $\boldsymbol{C}$ la matrice de covariance estimée avec le corpus d'apprentissage. Minimiser $\boldsymbol{f_{AV}(y)}$, revient à minimiser le terme de l'exponentielle de (17). En décomposant $\boldsymbol{C}^{-1}$ selon les termes impliquant les composantes auditives, visuelles et croisées, soit :

$$\boldsymbol{C}^{-1} = \begin{bmatrix} \boldsymbol{D} & \boldsymbol{E} \\ \boldsymbol{E}^{\mathbf{t}} & \boldsymbol{F} \end{bmatrix} \qquad (18)$$

nous obtenons un nouveau critère à minimiser :

$$f_{AV2}(\boldsymbol{y}) = \boldsymbol{Y_1}^{\mathbf{t}}\boldsymbol{D}\boldsymbol{Y_1} + 2\boldsymbol{V_1}^{\mathbf{t}}\boldsymbol{E}^{\mathbf{t}}\boldsymbol{Y_1} + \boldsymbol{V_1}^{\mathbf{t}}\boldsymbol{F}\boldsymbol{V_1} \qquad (19)$$

ou de manière équivalente:

$$f_{AV3}(\boldsymbol{y}) = (\boldsymbol{Y_1} - \boldsymbol{H}\boldsymbol{V_1})^{\mathbf{t}}\boldsymbol{D}(\boldsymbol{Y_1} - \boldsymbol{H}\boldsymbol{V_1}) \qquad (20)$$

$$\text{avec } \boldsymbol{H} = -\boldsymbol{D}^{-1}\boldsymbol{E} \qquad (21)$$

$\boldsymbol{H}$ est alors la matrice de régression de $\boldsymbol{V}$ vers $\boldsymbol{S}$. Ainsi nous faisons le lien avec le critère (15), la connaissance du vecteur $\boldsymbol{\gamma'_k}$, contenant les coefficients γ'_k, étant remplacée par $\boldsymbol{HV_1}$, et $\boldsymbol{D}$ remplaçant une simple sommation dans (15) par une sommation plus complexe comprenant des rotations et des poids.

Etude du cas de 2 sources.
Sachant que nous cherchons à extraire la source $\boldsymbol{s_1}$ sur $\boldsymbol{y_1}$, dans le cas d'un mélange de 2 sources, et en imposant le terme b_{11}=1 de la matrice $\boldsymbol{B}$, nous nous intéressons alors à :

$$\boldsymbol{y_1} = \boldsymbol{x_1} + b_{12}\boldsymbol{x_2} \quad \text{avec} \quad b_{12} = \arg\min(f_{AV}(\boldsymbol{y})) \qquad (22)$$

Quand $\boldsymbol{y_1}$ est défini par la 1ère éq. de (22), le spectre de $\boldsymbol{Y_1}$ décrit une courbe dans l'espace spectral en fonction de b_{12}, la 2ème éq. spécifiant la valeur optimale de b_{12} tel que $\boldsymbol{y_1}$ soit le plus cohérent avec $\boldsymbol{V_1}$ donc que $\boldsymbol{y_1}$ soit proche de $g_{11}\boldsymbol{s_1}$. Cependant, l'information vidéo $\boldsymbol{V_1}$ peut être parfois associée à plusieurs spectres ce qui peut aboutir à une mauvaise séparation. C'est pourquoi, en nous permettant de considérant que les données audio et spectrales sont indépendantes trame à trame, nous spécifions la probabilité conjointe intégrée:

$$p(\boldsymbol{Y_1}(t,\ldots,t-T),\boldsymbol{V_1}(t,\ldots,t-T)) = \prod_{l=0}^{T} p(\boldsymbol{Y_1}(t-l),\boldsymbol{V_1}(t-l)) \qquad (23)$$

En remplaçant dans l'eq. (16) la probabilité sur une trame par cette probabilité conjointe intégrée, nous pouvons espérer une meilleure estimation de la matrice $\boldsymbol{B}$.

3. RESULTATS EXPERIMENTAUX

3.1 Methodologie

Le corpus utilisé est un corpus mono locuteur composé des séquences [$V_1CV_2CV_1$] avec V_1 et V_2 une voyelle parmi [a, i, y, u] et C une consonne parmi [p, t, k, b, d, g]. Il y a deux répétitions de chaque séquence, une pour l'apprentissage et l'autre pour le test.
Les signaux audio-visuels, synchrones, sont analysés par trame de 20 ms. Pour chaque trame, le système d'analyse labiale développé à l'ICP fournit deux paramètres labiaux (hauteur et largeur intero labiales) et une analyse spectrale fournit 32 canaux spectraux. Une analyse en composantes principales permet ensuite de réduire le nombre de composantes spectrales en perdant un minimum d'information. Par la suite, nous prendrons 5 ou 8 composantes principales, fournissant respectivement 85.5 % et 92.5 % de variance expliquée.
Le corpus d'apprentissage (96 stimulis de 24 trames en moyenne, soit à peu près 2300 trames au total) est traité par l'algorithme EM qui fournit une modélisation multi-gaussienne pour laquelle nous avons utilisé 6, 8 et 12 gaussiennes.

Critère d'évaluation
Nous considérons un mélange à deux sources :

$$\boldsymbol{x_1} = a_{11}\boldsymbol{s_1} + a_{21}\boldsymbol{s_2} \quad ; \quad \boldsymbol{x_2} = a_{21}\boldsymbol{s_1} + a_{22}\boldsymbol{s_2} \qquad (24)$$

$\boldsymbol{s_2}$ étant le signal perturbateur. Les sources étant normalisées en énergie, les RSBs définis pour $\boldsymbol{s_1}$ sur chaque capteur $\boldsymbol{x_i}$ s'en déduisent comme :

$$RSB_{x1}= 20\log(a_{11}^2/a_{12}^2) \; ; \; RSB_{x2}= 20\log(a_{21}^2/a_{22}^2) \quad (25)$$

Et le *RSB* sur la sortie $\boldsymbol{y_1}$:

$$RSB_{y1}= 20\log((a_{11}+b_{12}a_{12})^2/(a_{12}+b_{12}a_{22})^2) \quad (26)$$

Nous avons pris pour la suite le mélange suivant :

$$a_{11}=2 \quad a_{12}=1 \quad a_{21}=3 \quad a_{22}=5 \quad (27)$$

Ceci fournit théoriquement une solution b_{12}=-0.2, et des valeurs de RSB_{xi} respectivement de 6 et -4.4 dB. L'expérimentation a été faite avec 96 stimuli de test (cf Sect. 3.1) contenant à peu près 2300 trames. La procédure de séparation est la suivante : $\boldsymbol{y_1}$ étant défini par la 1ère partie de (22), on estime son spectre $\boldsymbol{Y_1}$ selon le processus décrit en Sect. 3.1, puis on estime b_{12} selon la 2ème partie de (22). Une première optimisation a été effectuée avec l'algorithme du simplexe [Kli00].

3.2 Résultats

Sur les 2300 trames de test, nous avons comptabilisé le nombre de trames pour lesquelles la valeur de b_{12} estimée appartenait à l'intervalle [–0.15 , –0.25] centrée sur la solution, trames pour lesquelles RSB_{y1} est supérieur à 14 dB donc correctement débruitées. Nous présentons sur la figure 3 le pourcentage de trames correctement débruitées, pour une modélisation à 12 gaussiennes, 5 ou 8 composantes principales, et un nombre de trames *T* pour l'intégration temporelle (23) variant de 1 à 20.

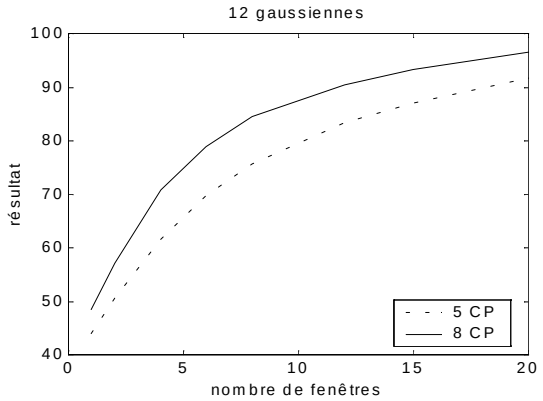

Figure 3 : Pourcentage de l'estimation correcte de b_{12} sur le corpus de test en fonction de *T*.

Au vu de cette figure il est assez facile de se rendre compte de l'importance de l'intégration temporelle. Pour *T*=20, 96 % de trames sont correctement débruitées si l'on utilise 8 composantes principales. L'augmentation du nombre de composantes principales et de gaussiennes pour la modélisation accroît légèrement les performances.

4. CONCLUSION

Il apparaît que cet algorithme audio-visuel possède les caractéristiques d'être théoriquement cohérent et techniquement réalisable. Cependant il aura besoin de plusieurs améliorations fondamentales. D'abord nous tenterons d'appliquer la méthode du gradient relatif vue en Sect. 2.1 à la place de celle du simplexe afin d'améliorer l'optimisation. Deuxièmement notre modèle statistique de probabilité conjointe audio-visuelle reposant sur une hypothèse d'indépendance trame à trame pourrait être remplacé par une probabilité plus complexe introduisant par exemple des modèles de Markov. D'autre part, si l'on compare notre algorithme à d'autres algorithmes classiques de séparation de sources, nous avons des performances semblables en termes de RSB sur la première source extraite [Kli00]. Il est vrai que de tels algorithmes permettent de séparer toutes les sources mais sans pouvoir contrôler et identifier une permutation possible entre les sources et les signaux de sortie de trame à trame. Ainsi notre algorithme dispose ici d'un atout non négligeable dans le cas de signaux audio en étant capable de combler cette carence. Enfin, un objectif à terme est bien sûr d'utiliser conjointement critère audio-visuel et critère d'indépendance. Ceci devrait s'avérer très utile dans des cas plus complexes impliquant des mélanges convolutifs ou des mélanges possédant moins de capteurs que de sources.

BIBLIOGRAPHIE

[Sum87] Summerfield Q. (1987), in Hearing by Eye : The Psychology of Lipreading, edited by Dodd B. and Campbell R (Lawrence Erlbaum Associates, London), pp. 3-51.

[Ben94] Benoît C. Mohamadi T. Kandel S. (1994) "Effects of phonetic context on audio-visual intelligibility of French", JSHR, Vol. 37, pp. 1195-1293.

[Rob98] Robert-Ribes J.Schwartz J.L. Lallouache T. Escudier P. (1998), " Complementary and synergy in bimodal speech : auditory, visual, and audio-visual identification of French oral vowels in noise", JASA, Vol. 103 :6, pp. 3677-3689.

[Gir01] Girin L. Schwartz J.L. & Feng G. (2001) " Audio-Visual enhancement of speech in noise", JASA, Vol. 109:6, pp. 3007-3020.

[Com95] Comon P. (1995), " Independenr Component Analysis, a new concept ?", Sig. Proc, Vol. 36:3, pp. 287-314.

[Sod01] Sodoyer D. (2001), " Séparation de sources Audio-visuelles: de la formalisation à l'expérimentation ", Master, INP Grenoble.

[Car96] Cardoso J.F. & Laheld B. (1996), " Equivariant adaptative source separation", IEEE Trans. on SP, Vol. 44:12, pp. 3017-3030.

[Yeh98] Yehia H. Rubin P. & Vatikiotis-Bateson (1998), " Quantitative association of vocal-tract and facial behavior", Speech Comm. , Vol. 26, pp. 23-43.

[Kli00] Klinkisch J. (2000), " Séparation audio-visuelle pour la reconnaissance de chiffre dans le bruit", Master, INP Grenoble.