Méthode temps-fréquence de séparation aveugle de sources basée sur la fonction de cohérence segmentée

Benoit ALBOUY, Yannick DEVILLE

Laboratoire d'Acoustique, de Métrologie et d'Instrumentation
Université Toulouse III, Bât. 3R1B2, 118 Route de Narbonne, 31062 Toulouse Cedex, France
albouy@cict.fr, ydeville@cict.fr

Résumé – Nous introduisons dans ce papier une nouvelle méthode de séparation aveugle de sources (SAS) concernant les mélanges linéaires instantanés. Cette approche est basée sur l'analyse de la fonction de cohérence fréquentielle réelle des signaux observés, qui est segmentée temporellement et permet de détecter les zones temps-fréquence (TF) où une seule source est active. Par ailleurs, cette méthode suppose seulement que les sources sont non corrélées. L'identification des coefficients de séparation est réalisée par le calcul de rapports de densités spectrales de puissance des signaux mélangés, dans des zones TF mono-sources. Cette approche fournit de très bonnes performances pour des mélanges de signaux de parole et/ou de bruit, avec des améliorations de rapports signal/bruit (SNRI) de 40 à plus de 90 dB et des taux de reconnaissance automatique de la parole de 100 %.

Abstract – In this paper, we introduce a new blind source separation (BSS) method for linear instantaneous mixtures. This approach is based on the time-segmented frequency-dependent real coherence function of the observed signals, which makes it possible to detect time-frequency (TF) zones where only one source is active. In addition, this method only assumes the sources to be uncorrelated. The separating coefficient identification is based on ratios of power spectral densities of the mixed signals, computed in single-sources TF zones. This BSS method yields very high performance for mixtures of speech and/or noise signals, with signal/noise ratios improvements ranging from 40 dB to more than 90 dB and 100 % automatic speech recognition rates.

1 Introduction

La séparation aveugle de sources (SAS) est un problème fondamental en traitement du signal, dont l'objectif est de restaurer des signaux sources à partir seulement de signaux observés, souvent modélisés comme des mélanges linéaires instantanés de ces sources [1, 2]. Son utilité dans des domaines très porteurs tels que les télécommunications, le rehaussement de la parole, la séismologie et même dans le cadre biomédical, a permis à la SAS de connaître un essor important ces dernières années.

La plupart des approches de SAS développées sont basées sur l'Analyse en Composantes Indépendantes (ACI), qui nécessite que les sources soient des signaux aléatoires stationnaires statistiquement indépendants.

Cependant d'autres méthodes existent, comme celles basées sur une analyse Temps-fréquence (TF) [3]-[8]. Nous proposons ici une méthode de SAS, qui suppose uniquement que les sources sont non corrélées et présentent des différences dans le plan TF. Nous exploitons ces différences au moyen de la fonction de cohérence fréquentielle réelle des observations, segmentée temporellement.

2 Méthode proposée

2.1 Configuration pour deux sources et deux capteurs

Dans la version de base de l'approche proposée, nous considérons la configuration de la Fig. 1, où les deux signaux centrés, $x_1(n)$ et $x_2(n)$, sont fournis par un jeu de deux capteurs. Ces signaux sont des mélanges de deux signaux sources aléatoires et réels, notés ici $s_1(n)$ et $s_2(n)$, supposés centrés et non corrélés. Le modèle de mélange considéré est ici linéaire instantané. Par conséquent, les signaux observés s'écrivent :

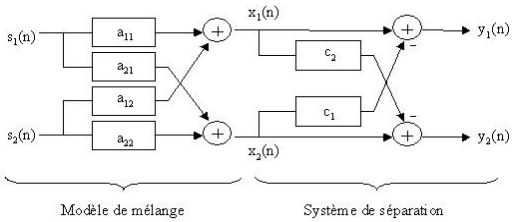

FIG. 1: *Configuration considérée.*

$$\begin{cases} x_1(n) &= a_{11}.s_1(n) + a_{12}.s_2(n) \\ x_2(n) &= a_{21}.s_1(n) + a_{22}.s_2(n) \end{cases} \quad (1)$$

où a_{ij} désigne le coefficient réel du mélange associé à la source j et au capteur i.

Le système de séparation considéré sur la Fig. 1, permet d'obtenir les signaux de sortie suivants :

$$\begin{cases} y_1(n) &= s_1(n).[a_{11} - c_1.a_{21}] + s_2(n).[a_{12} - c_1.a_{22}] \\ y_2(n) &= s_1(n).[a_{21} - c_2.a_{11}] + s_2(n).[a_{22} - c_2.a_{12}] \end{cases} \quad (2)$$

2.2 Identification du système de séparation

L'objectif de l'approche de SAS proposée est de déterminer une valeur du couple de coefficients (c_1, c_2) telle que les

signaux $y_1(n)$ et $y_2(n)$ en sortie du système de séparation ne dépendent respectivement que de chacun des signaux sources, c-à-d telle que :

$$y_1(n) = \beta.s_1(n) \quad et \quad y_2(n) = \eta.s_2(n) \quad (3)$$

ou

$$y_1(n) = \beta'.s_2(n) \quad et \quad y_2(n) = \eta'.s_1(n) \quad (4)$$

D'après (2), deux valeurs de (c_1, c_2) remplissent cette condition, i.e. : $(c_1, c_2) = \left(\frac{a_{12}}{a_{22}}, \frac{a_{21}}{a_{11}}\right)$ ou $(c_1, c_2) = \left(\frac{a_{11}}{a_{21}}, \frac{a_{22}}{a_{12}}\right)$. On notera que ces deux couples de valeurs sont de la forme :

$$(c_1, c_2) = \left(\frac{a_{1l}}{a_{2l}}, \frac{a_{2m}}{a_{1m}}\right) \quad (5)$$

avec respectivement $(l, m) = (2, 1)$ et $(l, m) = (1, 2)$.

La méthode que nous proposons dans cet article pour identifier l'un de ces couples de valeurs est fondée sur les densités auto- et inter-spectrales de puissance des signaux observés, successivement estimées sur diverses fenêtres temporelles repérées par l'indice k. De manière générale, pour les signaux définis en sous-section 2.1, ces densités s'expriment suivant :

$$S_{x_i x_i}(k,\omega) = a_{i1}^2.S_{s_1 s_1}(k,\omega) + a_{i2}^2.S_{s_2 s_2}(k,\omega) \quad (6)$$

$$S_{x_i x_j}(k,\omega) = a_{ii}.a_{ji}.S_{s_i s_i}(k,\omega) + a_{ij}.a_{jj}.S_{s_j s_j}(k,\omega) \quad avec\ i \neq j \in \{1,2\}. \quad (7)$$

Supposons que dans un domaine temps-fréquence (k_1, ω_1) défini par l'intervalle temporel k_1 et la pulsation ω_1, seule la source s_l est "active", i.e. : $S_{s_l s_l}(k_1, \omega_1) \neq 0$ et $S_{s_{l'} s_{l'}}(k_1, \omega_1) = 0$ pour $l \neq l'$. Alors (6) et (7) deviennent :

$$S_{x_i x_i}(k_1,\omega_1) = a_{il}^2.S_{s_l s_l}(k_1,\omega_1)$$

$$S_{x_i x_j}(k_1,\omega_1) = a_{il}.a_{jl}.S_{s_l s_l}(k_1,\omega_1) \quad i \neq j \in \{1,2\}$$

Il en résulte que :

$$\frac{S_{x_1 x_2}(k_1,\omega_1)}{S_{x_2 x_2}(k_1,\omega_1)} = \frac{a_{1l}}{a_{2l}} \quad (8)$$

De manière similaire, si dans un domaine (k_2, ω_2) seule la source s_m est "active", on obtient :

$$\frac{S_{x_2 x_1}(k_2,\omega_2)}{S_{x_1 x_1}(k_2,\omega_2)} = \frac{a_{2m}}{a_{1m}} \quad (9)$$

Le couple de valeurs défini par (8)-(9) avec $l \neq m \in \{1, 2\}$ s'avère alors identique à l'un des couples recherchés, définis par (5).

La méthode finale de SAS que nous proposons en découle directement. Elle suppose que, pour chacune des sources $s_1(n)$ et $s_2(n)$, il existe au moins une zone TF où elle seule est "active" (cette condition est par exemple facilement vérifiée pour des signaux de parole), les deux sources pouvant être conjointement actives dans d'autres zones TF. Notre méthode comporte alors les étapes suivantes :

- On détecte deux zones TF (k_1, ω_1) et (k_2, ω_2) où l'unique source active est respectivement s_l et s_m, avec $l \neq m$. La méthode que nous proposons pour cela est définie dans la sous-section suivante.
- On choisit pour valeurs de coefficients de séparation :

$$c_1 = \frac{S_{x_1 x_2}(k_1,\omega_1)}{S_{x_2 x_2}(k_1,\omega_1)} \quad et \quad c_2 = \frac{S_{x_2 x_1}(k_2,\omega_2)}{S_{x_1 x_1}(k_2,\omega_2)} \quad (10)$$

- On en déduit les signaux de sortie $y_1(n)$ et $y_2(n)$, conformément au système de séparation représenté sur la Fig. 1.

2.3 Détection des zones temps-fréquence mono-sources

La méthode proposée pour détecter des zones temps-fréquence où une seule source est active est basée sur la fonction de cohérence des observations, estimée sur les fenêtres temporelles indexées par k introduites plus haut. La fonction de cohérence dite "complexe" est définie par :

$$\gamma_{x_1 x_2}(k,\omega) = \frac{S_{x_1 x_2}(k,\omega)}{\sqrt{S_{x_1 x_1}(k,\omega).S_{x_2 x_2}(k,\omega)}} \quad (11)$$

et la fonction de cohérence dite "réelle" associée est :

$$\Gamma_{x_1 x_2}(k,\omega) = |\gamma_{x_1 x_2}(k,\omega)|^2.$$

Pour les signaux définis en sous-section 2.1, d'après (6) et (7) on obtient $\gamma_{x_1 x_2}(k,\omega)$ égale à :

$$\frac{a_{11}.a_{21}.S_{s_1 s_1}(k,\omega) + a_{12}.a_{22}.S_{s_2 s_2}(k,\omega)}{\sqrt{\left(a_{11}^2.S_{s_1 s_1}(k,\omega) + a_{12}^2.S_{s_2 s_2}(k,\omega)\right).\left(a_{21}^2.S_{s_1 s_1}(k,\omega) + a_{22}^2.S_{s_2 s_2}(k,\omega)\right)}}$$

La fonction de cohérence réelle des observations vérifie alors les deux propriétés suivantes, indépendamment dans chaque zone temps-fréquence (k, ω) :

Propriété 1 *si une seule source est active dans cette zone, alors :* $\Gamma_{x_1 x_2}(k,\omega) = 1$.

Propriété 2 *si les deux sources sont actives dans cette zone, i.e. si* $S_{s_1 s_1}(k,\omega) \neq 0$ *et* $S_{s_2 s_2}(k,\omega) \neq 0$ *alors :* $\Gamma_{x_1 x_2}(k,\omega) < 1$.

Preuve : Soient deux vecteurs $\overrightarrow{v_1}$ *et* $\overrightarrow{v_2}$ *définis de la façon suivante :*

$$\overrightarrow{v_1} = \begin{pmatrix} a_{11}.\sqrt{S_{s_1 s_1}(k,\omega)} \\ a_{12}.\sqrt{S_{s_2 s_2}(k,\omega)} \end{pmatrix} \ et\ \overrightarrow{v_2} = \begin{pmatrix} a_{21}.\sqrt{S_{s_1 s_1}(k,\omega)} \\ a_{22}.\sqrt{S_{s_2 s_2}(k,\omega)} \end{pmatrix}$$

La fonction de cohérence complexe $\gamma(k,\omega)$ *fournie en (12) est alors égale à :*

$$\gamma(k,\omega) = \frac{\langle \overrightarrow{v_1}, \overrightarrow{v_2} \rangle}{\| \overrightarrow{v_1} \| . \| \overrightarrow{v_2} \|} \quad (12)$$

où $\langle\ \rangle$ *représente le produit scalaire. En utilisant alors le théorème de Cauchy-Schwartz dans l'Eq. (12), on obtient :*

$$0 \leq \Gamma(k,\omega) = |\gamma(k,\omega)|^2 \leq 1 \quad (13)$$

De plus, on a $\Gamma(k,\omega) = |\gamma(k,\omega)|^2 = 1$ *si et seulement si* $det[\overrightarrow{v_1}, \overrightarrow{v_2}] = 0$. *Or le déterminant associé à ces deux vecteurs est égal à :*

$$det[\overrightarrow{v_1}, \overrightarrow{v_2}] = (a_{11}.a_{22} - a_{12}.a_{21})\sqrt{S_{s_1 s_1}(k,\omega) S_{s_2 s_2}(k,\omega)} \quad (14)$$

De plus, le mélange est supposé inversible, ce qui signifie que $(a_{11}.a_{22} - a_{12}.a_{21}) \neq 0$. *Alors la seule solution pour que* $det[\overrightarrow{v_1}, \overrightarrow{v_2}] = 0$ *est* $S_{s_1 s_1}(k,\omega) = 0$ *ou* $S_{s_2 s_2}(k,\omega) = 0$. *Il n'y a donc que dans le cas où une seule source est active dans la zone temps-fréquence considérée que* $\Gamma(k,\omega) = 1$.

Notre méthode pour détecter les zones temps-fréquence mono-sources (k_1, ω_1) et (k_2, ω_2) consiste donc à déterminer deux zones temps-fréquence correspondant aux valeurs parmi les plus élevées de $\Gamma_{x_1 x_2}(k,\omega)$ et fournissant deux valeurs de c_1 et c_2 définies en (10) qui ne sont pas inverses l'une de l'autre (afin de ne pas retenir deux zones correspondant à la même source).

3 Résultats expérimentaux

Toutes les expériences présentées dans cette section concernent des signaux sources (de parole et/ou de bruit), mélangés à l'aide de la matrice suivante :

$$\mathbf{A} = \begin{pmatrix} 1 & 0.8 \\ 0.9 & 1 \end{pmatrix}.$$

Nous obtenons des estimations des densités $S_{x_i x_i}(k,\omega)$ et $S_{x_i x_j}(k,\omega)$ auto- et inter-spectrales de puissance segmentées des observations à l'aide d'une version modifiée du périodogramme moyenné, dans laquelle chaque segment temporel k des signaux observés, de 4096 échantillons, est décomposé en sous-fenêtres se chevauchant (de 75%), à l'aide de fenêtres de pondération (Hamming), de 512 échantillons. Les densités spectrales de puissance sont d'abord calculées sur chaque sous-fenêtre avant d'être moyennées pour en déduire $S_{x_i x_i}(k,\omega)$ ou $S_{x_i x_j}(k,\omega)$. De plus, les segments temporels utilisés pour calculer ces densités moyennées présentent un recouvrement de 50 %.

Les rapports signal/bruit (SNR), mesurés dans les observations sont notés dans la suite : $SNR_{in}(i)$. Les performances sont alors évaluées en termes : i) de SNR mesurés sur chaque sortie i du système de SAS, notés $SNR_{out}(i)$, et ii) d'amélioration du SNR, pour chaque voie i, c-à-d : $SNRI(i)$=$SNR_{out}(i)$ / $SNR_{in}(i)$. Pour les expériences de cocktail-party, nous calculons le SNRI moyen obtenu par la méthode par :

$$SNRI=\sqrt{SNRI(1).SNRI(2)}.$$

La première série d'expériences est réalisée sur des signaux de parole de la base de données Multext. Nous considérons douze signaux répartis en six couples de la façon suivante : (1,2),(3,4),...(11,12). Nous présentons les résultats de ces six tests en cocktail-party dans le Tableau 1. Les couples de coefficients théoriques de séparation à identifier sont : ($c_{1_{theo}} = 0.8$,$c_{2_{theo}} = 0.9$) ou ($c_{1_{theo}} = 1.1111$,$c_{2_{theo}} = 1.25$) qui fournit des signaux de sortie permutés. Or pour quatre tests, nous identifions le couple de coefficients ($c_1 = 0.8$,$c_2 = 0.9$) avec une précision de ± 0.0001, alors que dans deux tests, on obtient ($c_1 = 1.1111$,$c_2 = 1.25$) à ± 0.0004 près. Les performances sont très élevées puisque les valeurs de SNRI sur ces six tests s'étendent de 55.7 dB à 76.5 dB, avec une moyenne de 68.3 dB.

La représentation TF de la fonction de cohérence $\Gamma(k,\omega)$, donnée en Fig. 2, montre la présence de nombreuses zones où $\Gamma_{x_1 x_2}(k,\omega)$ est proche de 1, c-à-d de zones mono-sources. Cela explique les performances élevées et confirme que notre méthode peut s'appliquer à de la parole continue, contrairement à notre approche similaire présentée dans [9] qui nécessite que les signaux présentent de longues zones temporelles de silence. Les représentations TF des sources considérées en Fig. 2 sont données sur la Fig. 3(a), qui représente une voix masculine (premier formant situé dans la bande de fréquence [100-200Hz]) et la Fig. 3(b) pour une voix féminine (premier formant situé à des fréquences supérieures à 200Hz). On retrouve cette différence entre les sources sur la Fig. 2, puisque de nombreuses zones mono-sources sont situées dans la bande [100-200Hz].

Indices des signaux de parole	$SNRI(1)$	$SNRI(2)$	$SNRI$
1, 2	65.4	63.6	64.5
3, 4	78.4	64.5	71.5
5, 6	51.1	60.2	55.7
7, 8	66.8	72.4	69.6
9, 10	77.9	66.2	72.1
11, 12	66.3	86.7	76.5

TAB. 1: *Cocktail-party : SNRI (dB) obtenus pour différents mélanges de signaux de parole de la base Multext.*

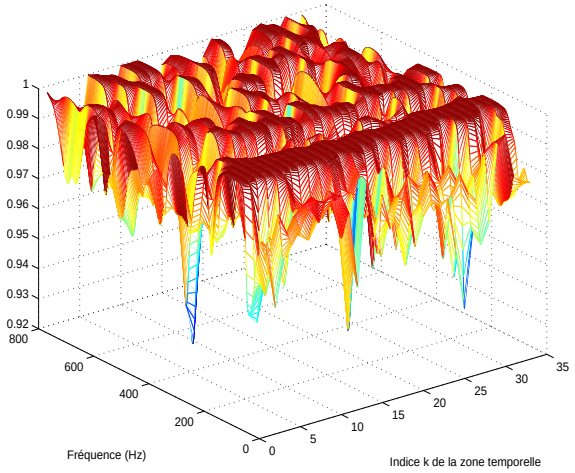

FIG. 2: *Représentation temps-fréquence de la fonction de cohérence des signaux observés.*

Le second jeu de tests proposé consiste à étudier le comportement de la méthode en rehaussement de la parole, pour les précédents signaux. Chacun des douze signaux de parole est mélangé à un bruit blanc, auquel est associé un facteur d'échelle α, pour permettre de traiter différents niveaux de SNR_{in}. Le Tableau 2 contient tous les $SNRI(i)$, correspondant à la sortie sur le canal i, où est extraite la source de parole. Dans le cas le plus bruité ($\alpha = 3$), les valeurs des améliorations obtenues pour ces tests s'étendent de 62.3 dB à 93.3 dB avec une moyenne de 78.2 dB.

D'autres expériences en rehaussement de la parole sont réa-

Indices des signaux de parole	$SNRI(i)$		
	alpha=1	alpha=2	alpha=3
1	65.3	71.5	75.0
2	69.0	75.2	78.8
3	70.5	76.7	80.2
4	81.6	88.1	91.7
5	83.4	89.7	93.3
6	60.2	66.5	70.0
7	62.4	68.6	72.2
8	59.4	66.0	69.6
9	52.2	58.7	62.3
10	82.1	88.0	91.5
11	63.6	69.9	73.5
12	71.0	77.3	80.9

TAB. 2: *Rehaussement de la parole : SNRI(i) (dB) obtenus pour différents signaux de parole et facteurs d'échelle α.*

Indices des signaux de parole	$SNRI(i)$		
	alpha=0.25	alpha=0.5	alpha=1
1	32.7	41.5	48.7
2	38.0	45.5	51.9
3	37.0	50.4	55.1
4	39.1	47.3	54.0
5	31.9	35.4	43.2
6	30.2	31.7	39.1

TAB. 3: *Rehaussement de la parole : SNRI(i) (dB) obtenus pour différents signaux de parole et facteurs d'échelle α.*

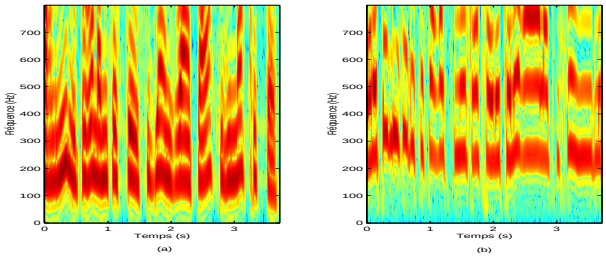

FIG. 3: *Spectrogrammes des signaux sources : (a) parole male et (b) parole femelle.*

lisées avec des signaux de parole, cette fois-ci discontinus car issus d'une base de données notée BD2, contenant des commandes de numérotation téléphonique. Le Tableau 3 représente les performances obtenues lors de ces expériences avec différents facteurs d'échelle α, afin de tester différents niveaux de SNR_{in} (entre environ -5 et 5 dB comme pour la base Multext, avec pour cela $\alpha = 0.25$, 0.5 et 1). On observe des performances en termes de SNRI légèrement moins élevées que les précédentes. Pour $\alpha = 1$, les améliorations de SNR s'étendent de 39.1 dB à 55.1 dB avec une valeur moyenne de 48.7 dB.

En plus des mesures de SNRI, nous avons réalisé des tests de reconnaissance automatique de la parole (RAP) sur les observations et les sorties du système de SAS, recueillies lors des précédents tests avec des sources de la base BD2. Le critère de performance est le pourcentage de mots reconnus (WRR), calculé de la façon suivante : WRR=(N-S-D-I)/N où N désigne le nombre total de mots contenus dans le signal, et S, D, I sont respectivement le nombre d'erreurs de substitution, de suppression et d'insertion (les insertions de silence n'étant pas comptabilisées comme des erreurs). Les valeurs obtenues lors de ces

		Indices des signaux			
alpha	signal	3	4	5	6
0.25	obs 1	84.21	78.95	84.21	73.68
0.25	obs 2	78.95	26.32	84.21	73.68
0.25	sortie	100	100	100	100
0.5	obs 1	21.05	15.79	15.79	15.79
0.5	obs 2	21.05	15.79	15.79	15.79
0.5	sortie	100	100	100	100
1	obs 1	21.05	15.79	15.79	15.79
1	obs 2	21.05	15.79	15.79	15.79
1	sortie	100	100	100	100

TAB. 4: *% de mots reconnus avant (obs) et après (sortie) séparation lors des tests en rehaussement de la parole.*

tests (voir le Tableau 4) prouvent que la méthode de séparation permet d'améliorer de façon idéale le WRR entre les signaux mélangés au niveau des observations, et les sorties du système de SAS : les signaux extraits sont vus identiques aux sources par le système de RAP, puisqu'ils sont reconnus à 100 %.

4 Conclusions

Nous avons présenté ici une méthode Temps-fréquence de SAS pour les mélanges linéaires instantanés basée sur la détection de zones TF mono-sources à l'aide de la fonction de cohérence réelle des observations segmentée temporellement, qui suppose uniquement les sources non corrélées. Des tests en cocktail-party et en rehaussement de la parole sur deux bases de données conduisent à de très fortes améliorations en termes de SNR mais aussi en termes de taux de mots reconnus par un système de traitement automatique de la parole. L'extension à N sources et N capteurs est immédiate et l'extension au cas convolutif envisagée.

Références

[1] J.F. Cardoso. *Blind Signal Separation : Statistical Principles*. Proceedings of the IEEE, vol.86, no.10, 1998.

[2] A. Hyvarinen, J. Karhunen, et E. Oja. *Independent Component Analysis*. John Wiley, 2001.

[3] A. Belouchrani, M.G. Amin. *Blind source separation using time-frequency distributions : algorithm and asymptotic performance*. ICASSP'97, pp. 3469-3472, Munich, Germany, 34-31 Avril, 1997.

[4] A. Holobar, C. Fevotte, C. Doncarli, D. Zazula. *Single autoterms selection for blind source separation in time-frequency plane*. EUSIPCO'2002, Toulouse, France, 3-6 Sept., 2002.

[5] L. Giulieri, N. Thirion-Moreau, P.-Y Arquès. *Blind sources separation using bilinear and quadratic time-frequency representations*. ICA'2001, pp. 486-491, San Diego, E-U, 9-13 Dec., 2001.

[6] A. Jourjine, S. Rickard, O. Yilmaz. *Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures*. ICASSP'2000, vol. 5, pp. 2985-2988, Istanbul, Turkey, 18-22 Juin, 2000.

[7] S. Rickard, R. Balan, J. Rosca. *Real-time time-frequency based blind source separation*. ICA'2001, pp. 651-656, San Diego, E-U, 9-13 Dec., 2001.

[8] F. Abrard, Y. Deville, P. White. *From blind source separation to blind source cancellation in the underdetermined case : a new approach based on time-frequency analysis*. ICA'2001, pp. 734-739, San Diego, E-U, 9-13 Déc., 2001.

[9] B. Albouy, Y. Deville. *Segmentation and separation of speech and/or noise signals, using coherence functions and power spectra*. ICA'2003, Nara, Japon, 1-4 Avril, 2003.