Séparation aveugle sous-déterminée de sources en utilisant la décomposition en paquet d'ondelettes

Abdeldjalil Aïssa-El-Bey, Karim Abed-Meraim, Yves Grenier

Département Traitement du Signal et des Images
ENST, 46 rue Barrault, 75634 Paris Cedex 13, France
elbey@enst.fr, abed@enst.fr, yves.grenier@enst.fr

Résumé – Dans le cadre de la séparation aveugle de sources, on cherche à effectuer la séparation de mélanges instantanés de sources audio en exploitant leurs parcimonie dans des domaines transformés. Cette approche nous permet, en particulier de traiter le cas sous-déterminé (c'est à dire le cas où l'on a moins de capteurs que de sources). Récemment, de nombreux travaux ont proposé d'exploiter la parcimonie des signaux audio dans le plan temps-fréquence (TF). Ces approches permettent d'obtenir de bons résultats dans le cas où les sources sont disjointes dans le plan temps-fréquence en utilisant des techniques de classification. Cependant, dans le cas où les sources sont non-disjointes, la reconstruction des signaux nécessite une interpolation des points de recouvrement dans le plan TF. Dans cet article, nous proposons une nouvelle technique qui combine la décomposition en paquets d'ondelettes et la projection en sous-espace afin de traiter explicitement le cas des sources non-disjointes. Nous présenterons quelques résultats de simulation qui permettent d'évaluer les performances de cette nouvelle méthode.

Abstract – This paper considers the blind separation of audio sources in the underdetermined case, where we have more sources than sensors. Recently, several time-frequency (TF) based solutions have been proposed using the sparsity representation of audio sources in the TF domain. These methods achieve good separation performance in the case where sources are disjoint in the TF plane. However, in the non-disjoint case, the reconstruction of the signals requires some interpolation at the intersection points in the TF plane. In this paper, we propose a new algorithm that combines the wavelet packet decomposition with subspace projection in order to explicitly treat non-disjoint sources. Finally, we will present some simulation results that illustrate the effectiveness of our algorithm.

1 Introduction

La séparation aveugle de sources est un problème qui consiste à retrouver des signaux indépendants à partir de leurs mélanges (observations) et cela sans connaissance a priori de la structure des mélanges ou des signaux sources. La séparation de sources intervient dans des applications diverses [1] telles que la localisation et poursuite de cibles en radar et sonar, la séparation de locuteurs (problème dit de "cocktail party"), la détection et séparation dans les systèmes de communication à accès multiple, l'analyse en composantes indépendantes de signaux biomédicaux (e.g., EEG ou ECG), etc. Ce problème a été intensément étudié dans la littérature et beaucoup de solutions efficaces ont déjà été proposées [1]. Néanmoins, le cas sous-déterminé où le nombre de sources est supérieur à celui des capteurs (observations) reste à ce jour un problème ouvert de la séparation aveugle de sources. Dans le cas de signaux non-stationnaires (incluant en particulier les signaux audio), certaines solutions utilisant la transformée temps-fréquence (TF) des observations existent pour le cas sous-déterminé [2, 3, 4, 5]. Ces approches permettent d'obtenir de bons résultats dans le cas où les sources sont disjointes (faible taux de recouvrement) dans le plan TF en utilisant des techniques de classification. On trouve aussi dans la littérature plusieurs techniques de séparation de sources exploitant d'autre domaine transformé notamment le domaine temps-échelle [6, 7].

Dans cet article, nous proposons une nouvelle technique basée sur une approche combinant la décomposition en paquet d'ondelettes et la projection en sous-espace, qui apporte un traitement explicite aux points d'intersection dans le cas de sources non-disjointes dans le domaine transformé. L'hypothèse utilisée dans cette méthode est que le nombre de sources présentes simultanément dans un point donné du domaine transformé, est inférieur au nombre de capteurs. Un autre avantage de cette méthode est qu'elle permet de pallier l'inconvénient majeur de la transformée de Fourier à court terme qui est la résolution fixe. Ceci est un handicap important lorsqu'on veut traiter des signaux dont les variations peuvent avoir des ordres de grandeur très variable tels que les signaux audio.

2 Formalisation du problème et hypothèses

Le modèle de séparation aveugle de sources suppose l'existence de N signaux $s_1(t), \ldots, s_N(t)$ et M observations $x_1(t), \ldots, x_M(t)$ qui représentent les mélanges. Ces mélanges sont supposés instantanés,

$$x_i(t) = \sum_{j=1}^{N} a_{ij} s_j(t) \qquad i = 1, \ldots, M \ . \tag{1}$$

Ceci peut être représenté par l'équation de mélange :

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t) \ , \tag{2}$$

où $\boldsymbol{s}(t) \triangleq [s_1(t), \ldots, s_N(t)]^T$ est le vecteur colonne de dimension $N \times 1$ qui regroupe les signaux sources, le vecteur $\boldsymbol{x}(t)$ regroupe de la même manière les $M \leq N$ signaux observés, et $\boldsymbol{A} \triangleq [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N]$ est la matrice de mélange de taille $M \times N$ où $\boldsymbol{a}_i = [a_{1i}, \ldots, a_{Mi}]^T$ contient les coefficients du mélange. Nous supposerons que tout sous ensemble de M vecteurs colonnes de $\boldsymbol{A}$ forme une famille de vecteurs linéairement indépendants.
Dans cet article, nous utiliserons comme domaine transformé le domaine temps-échelle (TE). Pour ce faire, nous utiliserons la décomposition en paquet d'ondelettes [8] qui est définie comme suit :

$$\mathcal{W}_x(t,\lambda) = \frac{1}{\sqrt{|\lambda|}} \sum_{m=-\infty}^{m=+\infty} x(m)\psi^*\left(\frac{m-t}{\lambda}\right) \qquad (3)$$

où $\psi(t)$ représente la fonction d'ondelette et $(\cdot)^*$ le complexe conjugué. En appliquant (3) au modèle des observations dans (2), on obtient l'expression suivante :

$$\mathcal{W}_{\boldsymbol{x}}(t,\lambda) = \boldsymbol{A}\mathcal{W}_{\boldsymbol{s}}(t,\lambda)\ , \qquad (4)$$

où $\mathcal{W}_{\boldsymbol{x}}(t,\lambda)$ (respectivement $\mathcal{W}_{\boldsymbol{s}}(t,\lambda)$) est le vecteur d'ondelette des mélanges (respectivement des sources).
Soit Ω_1 et Ω_2 les supports TE (c'est à dire les points du domaine transformé en ondelettes où l'énergie locale de la source considérée est non-nulle ou non-négligeable) de deux sources $s_1(t)$ et $s_2(t)$. Si $\Omega_1 \cap \Omega_2 \neq \emptyset$, les sources sont dites non-disjointes dans le plan TE. Dans ce travail, nous supposerons que pour tout point du plan TE, au plus $(M-1)$ sources se recouvrent. Cependant, on suppose aussi qu'il existe pour chaque source, des régions dans le plan TE où elle est présente seule.

3 Algorithme de séparation TE

3.1 Algorithme utilisant la classification vectorielle

Dans cette section, nous introduirons une version de l'algorithme présenté dans [2] utilisant la décomposition en paquet d'ondelettes et un nombre de capteurs $M \geq 2$. A partir de l'équation (4) obtenue après application de la décomposition en paquet d'ondelettes et sous l'hypothèse que toutes les sources sont disjointes dans le domaine TE, l'équation (4) est réduite à

$$\mathcal{W}_{\boldsymbol{x}}(t,\lambda) = \boldsymbol{a}_i \mathcal{W}_{s_i}(t,\lambda),\ \forall (t,\lambda) \in \Omega_i,\ \forall i = 1, \cdots, N. \quad (5)$$

On observe de cette équation que deux points TE (t,λ) et (t',λ') de Ω_i sont tels que $\mathcal{W}_{\boldsymbol{x}}(t,\lambda)$ et $\mathcal{W}_{\boldsymbol{x}}(t',\lambda')$ sont colinéaires. Ceci suggère l'utilisation de la classification vectorielle pour grouper les points des ensembles $\Omega_i,\ i = 1, \cdots, N$, et permettre ainsi la séparation des sources dans le domaine TE.
Au préalable, on propose de sélectionner tous les points d'énergie significative en utilisant un seuillage du bruit afin de réduire l'effet du bruit et la complexité calculative. Plus précisément, pour chaque trame d'échelle (t,λ_p) de la représentation TE, on applique le critère suivant pour tous les points de temps t_q appartenant à cette trame d'échelle :

$$\text{Si} \quad \frac{\|\mathcal{W}_{\boldsymbol{x}}(t_q,\lambda_p)\|}{\max\limits_t\{\|\mathcal{W}_{\boldsymbol{x}}(t,\lambda_p)\|\}} > \epsilon_1, \quad \text{garder}\,(t_q,\lambda_p)\ , \qquad (6)$$

où ϵ_1 est un seuil de faible valeur (typiquement $\epsilon_1 = 0.01$). Alors, l'ensemble des points sélectionnés Ω, est exprimé par $\Omega = \bigcup_{i=1}^{N} \Omega_i$, où Ω_i est le support TE de la source $s_i(t)$. Notons que, l'effet d'étalement de l'énergie du bruit et la localization de celle des sources dans le domaine TE permet d'augmenter la robustesse de la méthode proposée vis à vis du bruit. Ainsi, en utilisant l'équation (6), nous souhaitons garder uniquement les points TE où l'énergie du signal est significative. De plus, en raison de l'étalement de l'énergie du bruit dans le plan TE, la contribution du bruit dans les points TE des sources est relativement négligeable du moins pour des valeurs modérées et hautes du rapport signal à bruit (RSB).
Par la suite, nous proposons d'utiliser une procédure de classification vectorielle dans le but de reformer les supports TE des signaux sources. Premièrement, on estime les vecteurs de direction spatiale par :

$$\boldsymbol{v}(t,\lambda) = \frac{\mathcal{W}_{\boldsymbol{x}}(t,\lambda)}{\|\mathcal{W}_{\boldsymbol{x}}(t,\lambda)\|}, \quad (t,\lambda) \in \Omega\ , \qquad (7)$$

et on forcera leur premier élément à être réel et positif, et ceci sans perte de généralité.
Ensuite, on classifiera ces vecteurs en N classes $\{\mathcal{C}_i | i = 1, \cdots, N\}$, en utilisant l'algorithme de classification k-means [9]. La collecte de tous les points correspondant à tous les vecteurs de la classe $\mathcal{C}_i$, formera le support TE Ω_i de la source $s_i(t)$. Alors, le vecteur colonne $\boldsymbol{a}_i$ de la matrice de mélange $\boldsymbol{A}$ est estimé comme le centroïde de cet ensemble de vecteurs :

$$\widehat{\boldsymbol{a}}_i = \frac{1}{\#\mathcal{C}_i} \sum_{(t,\lambda)\in\Omega_i} \boldsymbol{v}(t,\lambda)\ , \qquad (8)$$

où $\#\mathcal{C}_i$ est le nombre de vecteurs dans cette classe.
Donc, on peut estimer la transformée temps-échelle de chaque source $s_i(t)$ par :

$$\widehat{\mathcal{W}}_{s_i}(t,\lambda) = \begin{cases} \widehat{\boldsymbol{a}}_i^H \mathcal{W}_{\boldsymbol{x}}(t,\lambda), & \forall\,(t,\lambda) \in \Omega_i, \\ 0, & \text{sinon}\ . \end{cases} \qquad (9)$$

En effet, à partir de l'équation (5), on a :

$$\widehat{\boldsymbol{a}}_i^H \mathcal{W}_{\boldsymbol{x}}(t,\lambda) = \widehat{\boldsymbol{a}}_i^H \boldsymbol{a}_i \mathcal{W}_{s_i}(t,\lambda) \approx \mathcal{W}_{s_i}(t,\lambda), \quad \forall\,(t,\lambda) \in \Omega_i\ .$$

3.2 Algorithme utilisant la projection en sous-espace

Nous proposons dans cette section de traiter le cas des sources non-disjointes dans le plan TE, en utilisant une projection en sous-espace appropriée et sous les hypothèses qu'au plus $(M-1)$ sources se recouvrent en chaque point TE. Plus précisément, on considère le point TE où les sources $s_{\alpha_1}, \ldots, s_{\alpha_{\mathcal{J}}}$ sont présentes ($\mathcal{J} < M$), et définissons les termes suivants :

$$\widetilde{\boldsymbol{s}}(t) = [s_{\alpha_1}(t), \ldots, s_{\alpha_{\mathcal{J}}}(t)]^T, \qquad (10a)$$

$$\widetilde{\boldsymbol{A}} = [\boldsymbol{a}_{\alpha_1}, \ldots, \boldsymbol{a}_{\alpha_{\mathcal{J}}}]\ . \qquad (10b)$$

Alors au point TE (t,λ), l'équation (4) est réduite à :

$$\mathcal{W}_{\boldsymbol{x}}(t,\lambda) = \widetilde{\boldsymbol{A}}\mathcal{W}_{\widetilde{\boldsymbol{s}}}(t,\lambda)\ . \qquad (11)$$

Soit $\boldsymbol{\mathcal{Q}}$ la matrice de projection orthogonale sur le sous-espace bruit de la matrice $\widetilde{\boldsymbol{A}}$. Alors, $\boldsymbol{\mathcal{Q}}$ peut être calculée par :

$$\boldsymbol{\mathcal{Q}} = \boldsymbol{I}_M - \widetilde{\boldsymbol{A}} \left(\widetilde{\boldsymbol{A}}^H \widetilde{\boldsymbol{A}} \right)^{-1} \widetilde{\boldsymbol{A}}^H \ . \tag{12}$$

Nous avons par conséquent l'observation suivante :

$$\begin{cases} \boldsymbol{\mathcal{Q}} \boldsymbol{a}_i = \boldsymbol{0}, & i \in \{\alpha_1, \ldots, \alpha_{\mathcal{J}}\} \\ \boldsymbol{\mathcal{Q}} \boldsymbol{a}_i \neq \boldsymbol{0}, & \text{sinon} \end{cases} . \tag{13}$$

En supposant que la matrice $\boldsymbol{A}$ est connue ou correctement estimée, l'exploitation de l'équation (13) nous permet d'identifier les indices $\alpha_1, \ldots, \alpha_{\mathcal{J}}$, et donc, les sources présentes au point (t, λ). En pratique, en tenant compte de l'effet du bruit, on peut détecter les vecteurs colonnes de $\widetilde{\boldsymbol{A}}$ en minimisant :

$$\{\alpha_1, \ldots, \alpha_{\mathcal{J}}\} = \arg \min_{\beta_1, \ldots, \beta_{\mathcal{J}}} \left\{ \| \boldsymbol{\mathcal{Q}} \mathcal{W}_{\boldsymbol{x}}(t, \lambda) \| \mid \widetilde{\boldsymbol{A}}_\beta = [\boldsymbol{a}_{\beta_1}, \ldots, \boldsymbol{a}_{\beta_{\mathcal{J}}}] \right\} . \tag{14}$$

Puis, les valeurs des $\mathcal{J}$ sources au point TE (t, λ) sont estimées par :

$$\widehat{\mathcal{W}_{\widetilde{\boldsymbol{s}}}}(t, \lambda) \approx \widetilde{\boldsymbol{A}}^{\#} \mathcal{W}_{\boldsymbol{x}}(t, \lambda) \ . \tag{15}$$

Dans les étapes précédentes, nous avons supposé que $\boldsymbol{A}$ est connue ou estimée a priori. Nous proposons d'estimer $\boldsymbol{A}$ comme suit : on effectue une classification de l'ensemble des vecteurs $\mathcal{W}_{\boldsymbol{x}}(t, \lambda)$ en N classes en utilisant des techniques de classification vectorielle existantes dans la littérature [9]. Dans ce travail, nous avons utilisé l'algorithme k-means. Les vecteurs colonnes de $\boldsymbol{A}$ sont estimés comme les N centroïdes des N classes.

4 Discussion

Le nombre de sources

Le nombre de sources N est supposé connu pour la méthode de classification vectorielle k-means que nous avons utilisé. Cependant, il existe des méthodes de classification [9] qui effectuent l'estimation des classes aussi bien que l'estimation du nombre de classes N. Dans nos simulations, nous avons observé que la plupart du temps le nombre de classes est sur-estimé, ce qui conduit à une mauvaise qualité de séparation. C'est pourquoi, une estimation robuste du nombre de sources dans le cas de la séparation aveugle de sources sous-déterminée demeure un problème ouvert et difficile qui mérite une attention particulière pour nos travaux futurs.

Le recouvrement des sources

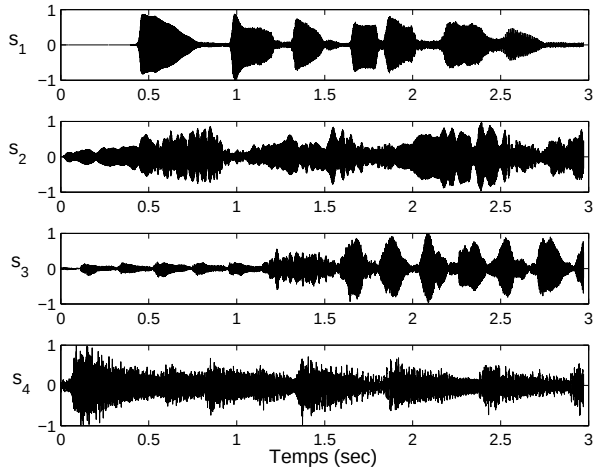
Dans l'approche basée sur la projection en sous-espace, nous devons évaluer le nombre de sources $\mathcal{J}$ présentes à un point TE donné. Il est possible de considérer une valeur fixe (maximum) de $\mathcal{J}$ qui sera utilisée pour tous les points TE. En effet, si le nombre de sources réellement présentes au point TE est inférieur à $\mathcal{J}$, nous estimerons des valeurs de la distribution TE des sources proche de zéro. Par exemple, si on suppose que $\mathcal{J} = 2$ sources sont présentes à un point TE donné, où une seule source contribue effectivement, alors on estimera une des deux valeurs de la distribution TE proche de zero. Cette approche augmente légèrement l'erreur d'estimation des signaux sources (en particulier pour des RSB faibles), mais a comme avantage la simplicité de mise en oeuvre. D'autre part l'optimization du critère (14) a été effectuée dans notre cas par une recherche exhaustive. En pratique, celle-ci peut être onéreuse si jamais $\mathcal{J}$ et M sont grands auquel cas d'autre techniques d'optimization devrait être envisagées.

5 Simulations

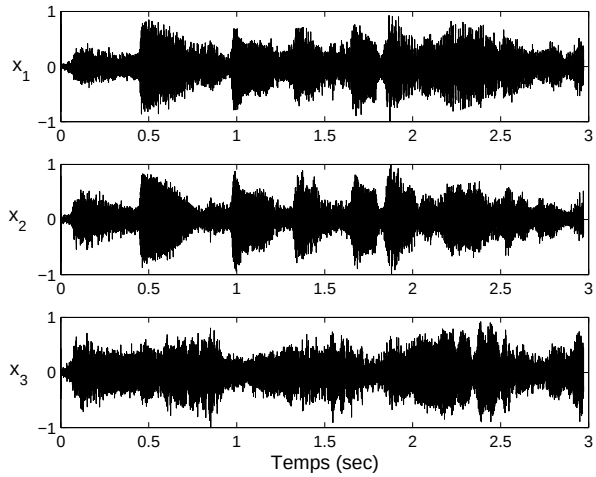
Dans cette section, nous présentons quelques résultats de simulation pour illustrer l'efficacité de notre algorithme de séparation. Pour cela, nous considérons une antenne composée de $M = 3$ capteurs recevant $N = 4$ signaux sources audio arrivant suivant les angles d'arrivés $\theta_1 = 15$, $\theta_2 = 30$, $\theta_3 = 45$ et $\theta_4 = 75$ degrés respectivement. La taille des observations $T = 25000$ échantillons (les signaux sources sont échantillonnés à une fréquence de 44.1 KHz). Les signaux observés sont corrompus par un bruit blanc additif de covariance $\sigma^2 \boldsymbol{I}_M$ (σ^2 étant la puissance du bruit). La qualité de la séparation est mesurée par l'erreur quadratique moyenne normalisée (EQMN) des sources estimées pour $N_r = 100$ réalisations aléatoires du bruit. Dans la figure 1, on représente un exemple d'exécution de notre algorithme de séparation. Les quatres premières lignes représentent les signaux sources originaux, les trois suivantes représentent les mélanges, et les quatres dernières donnent les estimées des sources en utilisant notre algorithme. La figure 2 représente la variation de l'erreur quadratique moyenne normalisée (EQMN) des estimées des signaux sources en fonction du RSB. Sur cette figure nous comparons les résultats obtenus par notre méthode où nous combinons la décomposition en paquet d'ondelettes (nous avons utilisé les paquets d'ondelettes de Daubechies) avec la projection sous-espace et des méthodes qui utilisent une transformation de Fourier à court terme (TFCT) avec une projection sous-espace (avec $\mathcal{J} = 2$) ou bien des techniques de classification [10]. Nous observons que l'utilisation de notre méthode apporte un gain significatif en terme d'EQMN sur pratiquement toute la plage du RSB ce qui reflète une amélioration de la qualité de la séparation.

6 Conclusion

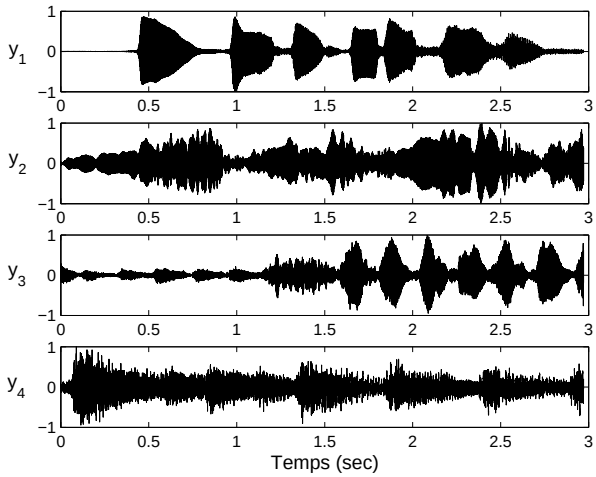
Dans cette article nous avons présenté une nouvelle technique de séparation aveugle de source dans le cas sous-déterminé pour des sources non-disjointes dans le domaine transformé temps-échelle. Les principaux avantages des algorithmes de séparation proposés sont : premièrement, une hypothèse plus réaliste sur la parcimonie des sources, c'est à dire que les sources ne sont pas nécessairement disjointes dans le domaine temps-échelle, deuxièmement, un traitement explicite des points de recouvrement en utilisant une projection en sous-espace, et troisièmement l'utilisation de la décomposition en paquet d'ondelettes mieux adaptée à la nature des signaux audio que la TFCT à résolution fixe. Ces avantages consistent à une amélioration significative de la qualité de séparation. Des résultats de simulations



(a) Les signaux originaux



(b) Les mélanges



(c) Les signaux estimés

FIG. 1 – (a) Les signaux sources originaux, (b) Les mélanges, (c) Les estimées des signaux sources.

illustrent l'efficacité de nos algorithmes comparés à ceux existant dans la littérature.

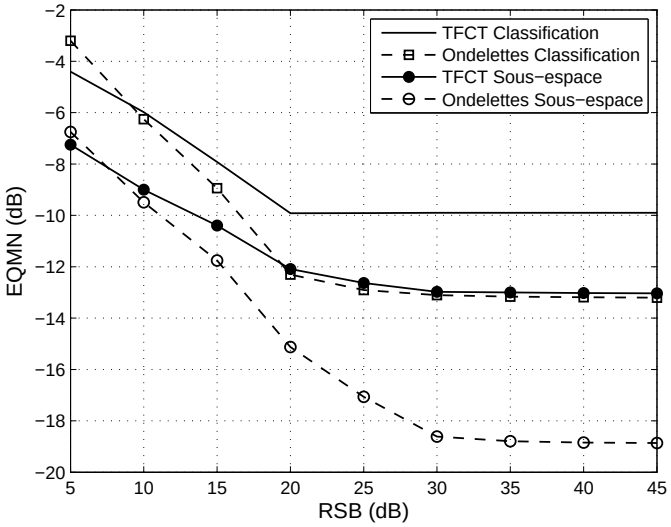


FIG. 2 – Performances de l'algorithme de séparation de 4 sources audio pour 3 capteurs : les courbes représentent la valeur moyenne de l'erreur quadratique de l'estimation des signaux sources en fonction du rapport signal à bruit.

Références

[1] A.K. Nandi editor, "Blind estimation using higher-order statistics." *Kluwer Acadimic Publishers*, Boston 1999.

[2] O. Yilmaz, S. Rickard, "Blind separation of speech mixtures via time-frequancy masking", *IEEE Transactions on Signal Processing*, Vol. 52, no. 7, pp 1830-1847, July 2004.

[3] N. Linh-Trung, A. Belouchrani, K. Abed-Meraim and B. Boashash, "Separating more sources than sensors using time–frequency distributions," *EURASIP Journal of Applied Signal Processing*, vol. 2005, no. 17, pp. 2828-2847, 2005.

[4] F. Abrard and Y. Deville, "A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources," *Signal Processing*, vol. 85, no. 7, pp. 1389-1403, July 2005.

[5] S. Arberet, R. Gribonval and F. Bimbot, "A robust method to count and locate audio sources in a stereophonic linear instantaneous mixture," *in Proc. ICA*, pp 536-543, March 2006.

[6] Y. Li, S.I. Amari, A. Cichocki, D.W.C. Ho and S. Xie, "Underdetermined blind source separation based on sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 2, pp. 423-437, February 2006.

[7] Y. Deville, D. Bissessur, M. Puigt, S. Hosseini and H. Carfantan, "A time-scale correlation-based blind separation method applicable to correlated sources," *in Proc. ESANN*, Belgium, April 2006.

[8] Stéphane Mallat, "Une exploration des signaux en ondelettes", *Les éditions de l'école polytechnique*, 2000.

[9] I.E. Frank and R. Todeschini, "The data analysis handbook", *Elsevier,Sci. Pub. Co.*, 1994.

[10] A. Aïssa-El-Bey, N. Linh-Trung, K. Abed-Meraim, A. Belouchrani, and Y. Grenier, "Underdetermined blind separation of nondisjoint sources using time-frequency representation", *IEEE Transactions on Signal Processing*, vol. 55, no. 3, pp. 897-907 March 2007.