Séparation de sources non-stationnaires par la parcimonie pour un mélange linéaire instantané

Bertrand Rivet

GIPSA-Lab, Département Parole & Cognition,
CNRS UMR-5216, Grenoble INP,
46, avenue Félix Viallet, 38000 Grenoble, France
bertrand.rivet@gipsa-lab.inpg.fr
www.icp.inpg.fr/~rivetber

ABSTRACT

We propose a method to estimate non-stationnary sources with non activity periods in a determined linear instantaneous mixture. Our method is based on the assumption that in some unknown temporal periods speech signals are inactive leading thus to an overdetermined mixture. Such silence periods allow to estimate the rows of the demixing matrix by a new algorithm called Direction Estimation of Separating Matrix (DESM). The periods of sources inactivity are estimated by a generalized eigen decomposition of covariance matrices of the mixtures, and the separating matrix is then estimated by a kernel principal component analysis. Experiments are provided with determined mixtures, and shown to be efficient.

Keywords : blind source separation, speech signals, sparsity, kernel PCA, DESM algorithm.

1. Introduction

La séparation aveugle de sources consiste à retrouver des signaux de sources à partir de mélanges de ces signaux, sans connaissance *a priori* sur la nature du mélange ou des sources. Pour les signaux de parole, la séparation peut ne pas être complètement aveugle puisque qu'il est possible de s'appuyer sur des propriétés spécifiques de ce signal. Ainsi, leur non-stationnarité [8, 11] ou leur parcimonie dans une certaine représentation [15, 1, 2] ont pu être exploitées. Parallèlement, il a également été envisagé de profiter de la bimodalité (audio et visuelle) de la parole [12, 14, 9, 10]. La séparation de source audiovisuelle repose sur les liens forts qui existent entre le son produit par le locuteur et les signaux visuels de la parole, notamment le mouvement des lèvres : ainsi ces méthodes recourent à la complémentarité et à la redondance de ces informations.

Notre nouvelle approche repose sur l'hypothèse de parcimonie du signal de parole : celui-ci est hautement non-stationnaire, il existe notamment de nombreuses périodes de temps au cours desquelles la puissance du signal de parole est négligeable vis-à-vis de sa puissance moyenne, citons par exemple les intervalles entre le mots. Cette nouvelle méthode s'inspire de [10] qui exploite les moments de silence d'un locuteur particulier pour identifier la fonction permettant ensuite d'extraire ce même locuteur lorsque celui parle, les instants de silence ayant été estimés par un détecteur purement visuel d'activité vocale reposant sur le mouvement des lèvres. Si cette approche audiovisuelle a pu montrer son efficacité même dans le cadre complexe de mélanges convolutifs, elle n'en demeure pas moins complexe à mettre en oeuvre puisqu'il est nécessaire de recueillir non seulement l'information audio grâce aux microphones mais aussi l'information visuelle obtenue par des caméras. Nous proposons donc ici de remplacer la détection visuelle d'activité vocale par un procédé purement acoustique, conduisant également à un nouveau principe d'estimation de la matrice de séparation.

Ce papier est organisé de la façon suivante. Le paragraphe 2 présente notre approche de la parcimonie tandis que le paragraphe 3 décrit l'algorithme DESM pour estimer les sources. Le paragraphe 4 regroupe les expériences numériques et les résultats obtenus avant de conclure cette étude au paragraphe 5.

2. Exploitation de la parcimonie

Dans ce paragraphe, nous présentons notre approche de la parcimonie du signal de parole après avoir rappelé les bases de la séparation de source dans un cadre linéaire instantané.

Soit $\mathbf{s}(t) \in \mathbb{R}^{N_s}$ le vecteur colonne regroupant les N_s sources $s_j(t)$ à l'instant t. Dans le cadre linéaire instantané, les N_m mélanges $x_i(t)$ s'expriment comme une combinaison linéaire des sources $s_j(t)$: $x_i(t) = \sum_j a_{i,j} s_j(t)$, ou de façon matricielle

$$\mathbf{x}(t) = A\,\mathbf{s}(t), \tag{1}$$

où $A \in \mathbb{R}^{N_m \times N_s}$ est la matrice de mélange dont le $(i,j)^{\text{ème}}$ terme est $a_{i,j}$ et $\mathbf{x}(t) \in \mathbb{R}^{N_m}$ le vecteur colonne regroupant les N_m mélanges $x_i(t)$. Dans la suite de cette étude, nous nous placerons dans le cas dit déterminé où le nombre de capteurs N_m est égal au nombre de sources N_s. L'estimation des sources est alors équivalente à estimer une matrice de séparation $B \in \mathbb{R}^{N_s \times N_s}$ telle que

$$\mathbf{y}(t) = B\,\mathbf{x}(t) \tag{2}$$

soit un vecteur dont les composantes soient les estimées des sources $s_i(t)$.

Si l'analyse en composantes indépendantes [4, 3] est un bon candidat pour résoudre ce problème en exploitant l'indépendance mutuelle, la parcimonie a récemment été introduite en séparation de sources [6]. Ainsi, par exemple, les méthodes proposées dans [15, 1, 2] font l'hypothèse que, dans une certaine base de représentation, il existe des zones où une seule source

est présente à la fois dans les mélanges permettant ainsi l'identification de la matrice de mélange. En effet, si à l'instant τ seule la $n^{\text{ème}}$ source est active (*i.e.* $\forall i \neq n, s_i(\tau) = 0$) alors $\mathbf{x}(\tau) = \mathbf{a}_n s_n(\tau)$. En d'autres termes, les mélanges $\mathbf{x}(\tau)$ sont colinéaires à la $n^{\text{ème}}$ colonne $\mathbf{a}_n$ de la matrice de mélange A. Il est ainsi possible d'obtenir toutes les colonnes de la matrice de mélange et d'exprimer ensuite la matrice de séparation B comme l'inverse de la matrice de mélange estimée.

L'approche que nous proposons ici est quelque peu différente puisqu'elle repose sur l'hypothèse qu'il existe des zones où au moins une source est inactive : *i.e.* pour $t = \tau$, $\exists n \ / \ s_n(\tau) = 0$. Supposons ainsi que toutes les sources soient stationnaires excepté une qui sera arbitrairement la première. Soient R_1 la matrice de covariance des observations $\mathbf{x}(t)$ calculée pour l'ensemble des instants t et R_2 la matrice de covariance des observations calculée pendant une période où la source $s_1(t)$ est inactive. La méthode proposée reprend l'idée de [13] sur la décomposition propre généralisée du couple (R_2, R_1). Ainsi, le couple (R_2, R_1) n'admet que deux valeurs propres généralisées distinctes : 1 dégénérée $N_s - 1$ fois dont le sous-espace propre $\mathcal{E}$ est un hyperplan complémentaire à $\mathbf{a}_1$ et une valeur propre généralisée nulle dont le vecteur propre généralisé associé $\mathbf{v}$ est orthogonal à $\mathcal{E}$. Ainsi, la projection des observations $\mathbf{x}(t)$ sur le vecteur propre généralisé $\mathbf{v}$ permet extraire la source $s_1(t) = \mathbf{v}^T\mathbf{x}(t)$ en annulant la contribution des autres sources : $\forall i \neq 1, \ \mathbf{v}^T\mathbf{a}_i = 0$ puisque $\mathbf{a}_i$ pour $i \neq 1$ vit dans l'hyperplan $\mathcal{E}$.

Cette méthode permet donc d'une part de détecter si la source $s_1(t)$ s'annule en testant les valeurs propres généralisées et d'autre part d'extraire cette même source lorsque celle-ci n'est plus inactive en projetant les observations sur le vecteur propre généralisé associé à la valeur propre généralisée nulle.

3. L'algorithme DESM

Dans le paragraphe précédent, nous avons fait l'hypothèse qu'une seule source est non-stationnaire présentant de plus des périodes inactives. Or un mélange peut contenir plusieurs de ces sources, par exemple si celui-ci est fait de plusieurs sources de parole. De plus, les périodes inactives des différentes sources sont inconnues. Dans ce paragraphe, nous allons donc expliquer comment mettre en oeuvre, par l'algorithme DESM ("Direction Estimation of Separating Matrix" en anglais), le principe décrit au paragraphe précédent pour extraire des observations les sources non-stationnaires présentant des périodes inactives.

De façon à détecter les périodes où au moins une source est inactive, nous proposons de calculer les structures propres généralisées des couples $\{(R_2(\tau), R_1)\}_\tau$ où R_1 est la matrice de covariance des observations $\mathbf{x}(t)$ calculée sur l'ensemble des échantillons temporels et $R_2(\tau)$ la matrice de covariance des observations $\mathbf{x}(t)$ calculée pour les échantillons temporels voisins de τ (typiquement, la fenêtre de calcul de ces matrices de covariance est de l'ordre de 100 millisecondes). Les décompositions en valeurs propres généralisées des couples $\{(R_2(\tau), R_1)\}_\tau$ fournissent donc

$$R_2(\tau)\,\Phi(\tau) = R_1\,\Phi(\tau)\,\Lambda(\tau), \tag{3}$$

où $\Lambda(\tau)$ est une matrice diagonale dont les termes diagonaux $\lambda_1(\tau) \leq \cdots \leq \lambda_{N_s}(\tau)$ sont les valeurs propres généralisées et $\Phi(\tau)$ une matrice orthonormale dont les colonnes sont les vecteurs propres généralisés. Ainsi à l'instant τ, si N sources sont inactives alors N valeurs propres généralisées sont nulles dont les vecteurs propres généralisés associés définissent un sous-espace orthogonal au sous-espace engendré par les $N_s - N$ autres sources actives.

L'algorithme DESM fonctionne donc sur le principe suivant. Il s'agit tout d'abord de détecter les périodes où au moins une source est inactive en testant les valeurs propres généralisées $\lambda_1(\tau)$: si $\lambda_1(\tau) \leq \eta$, où η est un seuil fixé *a priori*, alors on décide qu'un au moins une source est inactive dans la fenêtre temporelle centrée sur τ. Soit $\Theta = \{\tau \mid \lambda_1(\tau) \leq \eta\}$, de cardinal N_τ, l'ensemble des indices temporels où au moins une source est inactive. Ainsi, on obtient un ensemble de vecteurs $\{\phi_1(\tau)\}_{\tau\in\Theta}$ défini comme l'ensemble du premier vecteur propre généralisé pour les instants temporels τ où au moins une source est inactive. Ces vecteurs sont principalement alignés dans les directions permettant d'extraire les sources correspondantes (*cf* Fig. 2).

Il s'agit ensuite d'estimer ces directions. Pour cela, nous proposons d'utiliser une analyse en composantes principales à noyaux ("Kernel PCA" en anglais) [7, 5], où le noyau est choisi de la forme

$$k\big(\mathbf{x}(t), \mathbf{x}(t')\big) = k_{t,t'} \stackrel{\triangle}{=} \begin{cases} \frac{\mathbf{x}^T(t)\mathbf{x}(t') - \cos\theta_0}{1-\cos\theta_0}, \text{ si } \mathbf{x}^T(t)\mathbf{x}(t') \leq \cos\theta_0 \\ 0, \text{ sinon} \end{cases} \tag{4}$$

pour t et t' appartenant à Θ et où θ_0 est un angle fixé *a priori*. La Kernel PCA consiste donc à faire une décomposition propre de la matrice $K \in \mathbb{R}^{N_\tau \times N_\tau}$ dont le $(i,j)^{\text{ème}}$ terme est $k_{i,j}$:

$$K = \Psi\Delta\Psi^T, \tag{5}$$

où Δ est une matrice diagonale regroupant les valeurs propres de K et Ψ est une matrice orthonormale dont les colonnes sont les vecteurs propres de K. Soit $W = [\psi_1, \cdots, \psi_{N_s}]$ la matrice regroupant les N_s vecteurs propres ψ_i de K associés aux N_s valeurs propres maximales. La matrice de séparation B est alors obtenue par

$$B = W^T\,K\,V, \tag{6}$$

où $V = [\phi_1(t \in \Theta)]$ est la matrice obtenue par la concaténation du vecteur propre généralisé associé à la valeur propre généralisée minimale $\lambda_1(t)$ (3) pour $t \in \Theta$. Les sources sont alors estimées grâce à

$$\hat{s}(t) = B\mathbf{x}(t), \tag{7}$$

pour tous les instants temporels t, y compris ceux où les sources sont actives.

Finalement, l'algorithme DESM qui permet d'extraire les sources non-stationnaires présentant des périodes d'inactivité est résumé par l'algorithme 1.

Algorithme 1 Algorithme DESM.

1: Calculer la matrice de covariance R_1 avec l'ensemble des échantillons temporels
2: **Pour** chaque instant τ **faire**
3: Calculer la matrice de covariance $R_2(\tau)$ sur une fenêtre temporelle centrée sur τ
4: Calculer la décomposition propre généralisée (3) du couple $(R_2(\tau), R_1) \Rightarrow (\Phi(\tau), \Lambda(\tau))$
5: **Fin**
6: Estimer $\Theta = \{\tau \mid \lambda_1(\tau) \leq \eta\}$
7: Calculer la matrice K définie par (4)
8: Faire la décomposition propre (5) de $K \Rightarrow (\Psi, \Delta)$
9: Calculer $W = [\psi_1, \cdots, \psi_{N_s}]$ et $V = [\phi_1(t \in \Theta)]$
10: Calculer $B = W^T K V$ (6)
11: Estimer les sources par $\hat{\mathbf{s}}(t) = B\mathbf{x}(t)$

Notez que l'emploi des valeurs propres généralisées du couple $(R_1, R_2(\tau))$ à l'étape 4, au lieu des simples valeurs propres de $R_2(\tau)$ permet de s'affranchir du problème de la différence de puissance des sources, notamment si certaines d'entre elles sont nettement moins puissantes que les autres, l'algorithme risquant alors de considérer ces sources comme inactives.

4. Résultats numériques

Dans ce paragraphe, nous illustrons le principe d'extraction de sources par l'algorithme DESM. Pour cela, les sources sont issues d'un corpus de 18 phrases françaises lues par huit locuteurs différents (hommes et femmes). Les signaux sont échantillonnés à 16kHz. Pour les différentes configurations testées, les sources sont choisies de façon aléatoire et les coefficients de la matrice de mélange sont issus d'une variable aléatoire uniforme indépendante et identiquement distribuée entre -1 et 1. Nous avons tester plusieurs configurations de mélanges en faisant varier le nombre de sources, mais nous ne présentons ici qu'un simple exemple contenant trois sources pour illustrer le principe de l'algorithme. Une des sources est un signal de musique ne présentant pas de période de silence. Dans cet exemple, il s'agit de la deuxième.

Tout d'abord, la figure 1 montre la détection des périodes d'inactivité des sources par la décomposition propre généralisée. Comme on peut le constater sur le tracé du haut, qui représente la puissance des trois sources sur une fenêtre glissante de 100ms, les deux sources de parole présentent des périodes d'inactivité pouvant se recouvrir, tandis que la source musicale voit sa puissance moyenne à court terme presque constante. Il est alors intéressant de remarquer que la valeur propre généralisée $\lambda_1(t)$ (tracé du bas) permet effectivement de bien détecter les périodes de silence, sans pour autant indiquer laquelle des sources est inactive. Ceci est obtenu grâce aux vecteurs propres généralisés $\phi_1(t)$ correspondant (Fig. 2) qui sont alignés dans deux directions principales correspondant alors aux lignes de la matrice de séparation permettant d'extraire les sources de parole. Finalement, cet exemple de trois sources et trois capteurs est donné à la figure 3 où l'on constate que le principe proposé permet effectivement d'extraire du mélange les sources de parole ($\hat{s}_1(t)$ et $\hat{s}_2(t)$). La troisième source estimée reste un mélange des trois sources puisque la

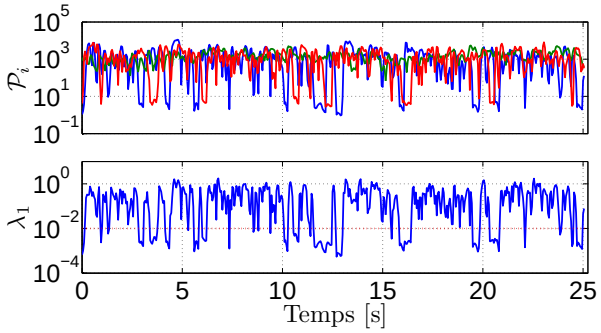


Fig. 1: Estimation des périodes d'inactivité des sources par l'algorithme DESM. La figure du haut représente la puissance des trois sources (Fig. 3(a)) pour une fenêtre glissante de 100ms (bleu, vert et rouge pour la 1ère, 2ème et 3ème source respectivement). La figure du bas représente la valeur propre généralisée minimale (3) $\lambda_1(\tau)$ (bleu) ainsi que le seuil choisi η (pointillés rouges). Les courbes sont tracées en échelle logarithmique.

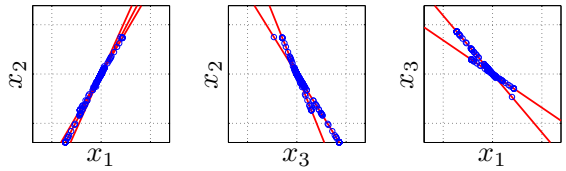
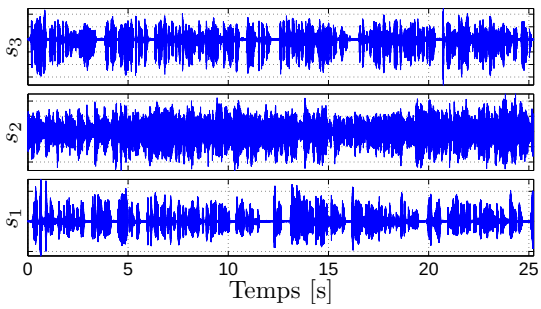


Fig. 2: Exemple de l'estimation des directions de la matrice de séparation B (6) par l'algorithme DESM dans le cas de trois sources. Sont représentés les projections, sur les plans (x_1, x_2), (x_2, x_3) et (x_1, x_3), des directions estimées (droites rouges) ainsi que les vecteurs propres généralisés $\phi_1(t)$, pour $t \in \Theta$ (ronds bleus). Ces vecteurs propres généralisés ont été multipliés par l'inverse de $\lambda_1(t)$.

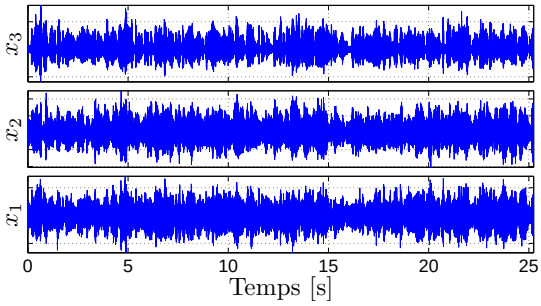
"kernel PCA" de la matrice K_1 ne présente que deux valeurs propres significatives. De façon plus générale, le nombre de valeurs propres significatives de la matrice K_1 permettrait d'estimer le nombre de sources de parole dans le mélange pour n'extraire que celles-ci.

5. Conclusions et perspectives

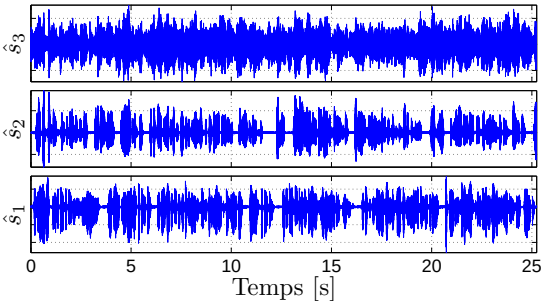
Dans ce papier, nous avons présenté un nouvel algorithme nommé DESM pour l'extraction des sources non-stationnaires présentant des périodes d'inactivité d'un mélange linéaire instantané. La détection de ces zones d'inactivité permet d'estimer la matrice de séparation pour ensuite extraire les sources correspondantes lorsque celles-ci sont présentes dans les mélanges. Cet algorithme a été testé dans différentes configurations et a montré son efficacité. Si dans cette étude seul le cas de mélanges linéaires instantanés a été envisagé, la méthode proposée pourrait s'étendre au cadre de mélanges convolutifs en transformant, de façon classique, le mélange convolutif en autant de problèmes instantanés que de fréquences de calcul de la transformée de Fourier discrète. Cette méthode poserait alors classiquement le problème de l'indétermi-



(a) Sources



(b) Mélanges



(c) Sources estimées

Fig. 3: Séparation de sources par l'algorithme DESM.

nation des permutations [3] qui pourrait être résolu en recourant à l'une des nombreuses méthodes de la littérature. Enfin, une alternative à cette solution sera d'étudier la possibilité d'appliquer ce principe au cas des mélanges linéaires convolutifs anéchoïdes de façon à mieux modéliser les mélanges acoustiques devant prendre en compte la propagation des signaux.

Références

[1] Frédéric Abrard and Yannick Deville. A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources. *Signal Processing*, 85(7) :1389–1403, July 2005.

[2] Simon Arberet, Rémi Gribonval, and Frédéric Bimbot. A robust method to count and locate audio sources in a stereophonic linear instantaneous mixture. In *Proc. ICA*, pages 536–543, Charleston, USA, March 2006.

[3] Jean-François Cardoso. Blind signal separation : statistical principles. *Proceedings of the IEEE*, 86(10) :2009–2025, October 1998.

[4] Pierre Comon. Independent component analysis, a new concept ? *Signal Processing*, 36(3) :287–314, April 1994.

[5] Frédéric Desobry and Cédric Févotte. Kernel PCA based estimation of the mixing matrix in linear instantaneous mixtures of sparse sources. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, pages 669–672, Toulouse, France, May 2006.

[6] Rémi Gribonval and Sylvain Lesage. A survey of Sparse Component Analysis for Blind Source Separation : principles, perspectives, and new challenges. In *Proc. European Symposium on Artificial Neural Networks (ESANN)*, pages 323–330, Bruges, April 2006.

[7] K.-R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2) :181–201, March 2001.

[8] Lucas Parra and Clay Spence. Convolutive blind separation of non stationary sources. *IEEE Transactions on Speech and Audio Processing*, 8(3) :320–327, May 2000.

[9] Bertrand Rivet, Laurent Girin, and Christian Jutten. Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures. *IEEE Transactions on Speech and Audio Processing*, 15(1) :96–108, January 2007.

[10] Bertrand Rivet, Laurent Girin, and Christian Jutten. Visual voice activity detection as a help for speech source separation from convolutive mixtures. *Speech Communication*, 49(7-8) :667–677, 2007.

[11] Christine Servière and Dinh-Tuan Pham. A novel method for permutation correction in frequency-domain in blind separation of speech mixtures. In *Proc. ICA*, pages 807–815, Granada, Spain, 2004.

[12] David Sodoyer, Laurent Girin, Christian Jutten, and Jean-Luc Schwartz. Developing an audio-visual speech source separation algorithm. *Speech Communication*, 44(1–4) :113–125, October 2004.

[13] Antoine Souloumiac. Blind source detection and separation using second order non-stationarity. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 1912–1915, Detroit, USA, May 1995.

[14] Wenwu Wang, Darren Cosker, Yulia Hicks, Saied Sanei, and Jonathon A. Chambers. Video assisted speech source separation. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, USA, March 2005.

[15] Özgür Yilmaz and Scott Rickard. Blind Separation of Speech Mixtures via Time-Frequency Masking. *IEEE Transactions on Signal Processing*, 52(7) :1830–1847, July 2004.