Texte

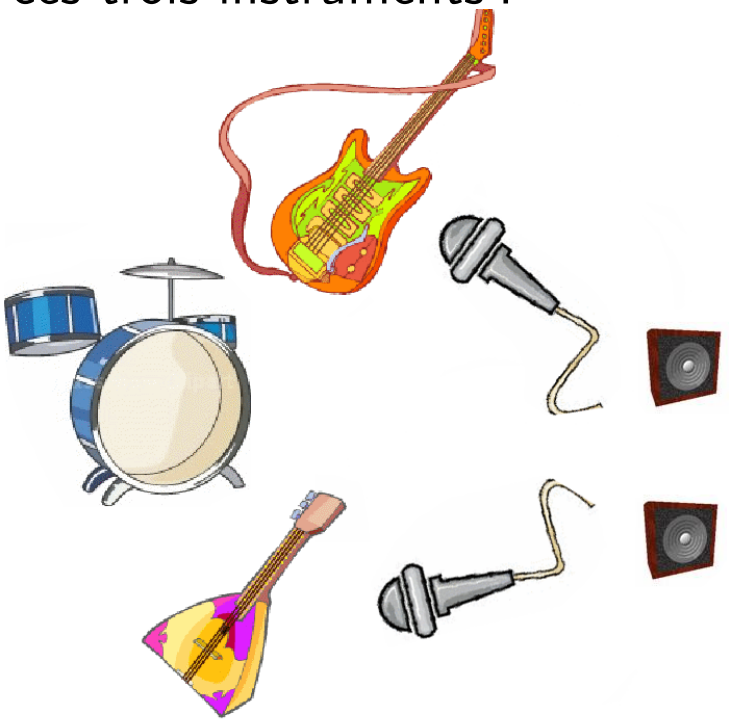
# Séparation (non-aveugle) de sources dans le cas sous-déterminé

Matthieu Kowalski

L2S – CNRS - Supelec - Univ. Paris Sud

# Problème et notations

Problème : On “observe” un mélange stéréo (bruité) de trois instruments. Comment retrouver ces trois instruments ?



Plus généralement, comment retrouver les signaux composant un mélange lorsqu'on dispose de moins d'oberservations que de signaux ?

# Problème et notations

## Mélange instantané sous déterminé

On observe $M$ signaux issus d'un mélange de $N$ signaux, avec $M < N$

$$X = A S$$

$$\begin{pmatrix} \mathbf{x}^1 \\ \vdots \\ \mathbf{x}^M \end{pmatrix} = A \begin{pmatrix} \mathbf{s}^1 \\ \vdots \\ \mathbf{s}^N \end{pmatrix}$$

$$\begin{pmatrix} x_1[t_1] & \dots & x_1[t_T] \\ \vdots & & \\ x_M[t_1] & \dots & x_M[t_T] \end{pmatrix} = \begin{pmatrix} a_{11} & \dots & \dots & a_{1N} \\ \vdots & & & \\ & & & \\ a_{M1} & \dots & \dots & a_{MN} \end{pmatrix} \begin{pmatrix} s_1[t_1] & \dots & s_1[t_T] \\ \vdots & & \\ & & \\ s_N[t_1] & \dots & s_N[t_T] \end{pmatrix}$$

$X \in \mathbb{R}^{M \times T}$, $A \in \mathbb{R}^{M \times N}$, $S \in \mathbb{R}^{N \times T}$

Problème (séparation *aveugle* de source) : comment estimer $A$ et $S$ ?

# Difficultés

Séparation Aveugle de Source (SAS ou BSS) : estimer $S$ et $A$ à partir de la seule observation $X$

$$X = AS$$

avec $X \in \mathbb{R}^{M \times T}$, $A \in \mathbb{R}^{M \times N}$, $S \in \mathbb{R}^{N \times T}$.

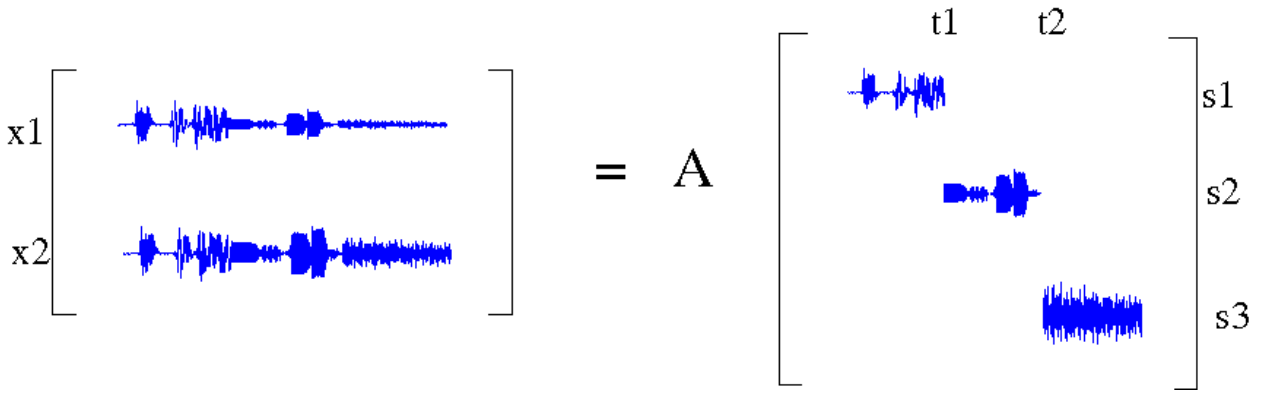
**Difficultés : problème mal posé dans le sens où**

1. On ne connait pas forcément le nombre de sources intervenant dans le mélange ;
2. On ne connait pas forcément la matrice de mélange $A$ ;
3. Même si on connait $A$, il y a plus d'inconnues que d'équations.

Quelles(s) hypothèse(s) raisonnable(s) peut-on faire pour inverser le système d'équation ?

# Un premier cas (très) simple - (1)

On considère un mélange stéréo particulier :




Dans ce cas :

$$x_1[t] = a_{11}s_1[t] + a_{12}s_2[t] + a_{13}s_3[t]$$
$$x_2[t] = a_{21}s_1[t] + a_{22}s_2[t] + a_{23}s_3[t]$$

Mais, si l'on se restreint à l'intervalle $[0, t_1]$

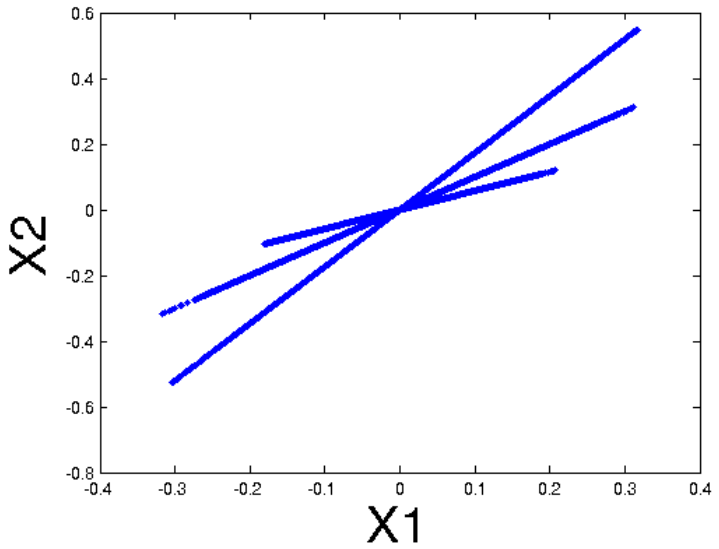
$$\begin{cases} x_1[t] &= a_{11}s_1[t] \\ x_2[t] &= a_{21}s_1[t] \end{cases} \quad \Longrightarrow \quad x_2[t] = \frac{a_{21}}{a_{11}} x_1[t]$$

De même pour les autres intervalles de temps.

# Un premier cas (très) simple - (2)

Comme pour tout $i \in 1, 2, 3$, $x_2[t] = \frac{a_{1i}}{a_{2i}} x_1[t]$, on peut en déduire $A$.

Graphiquement : on représente $x_2$ en fonction de $x_1$

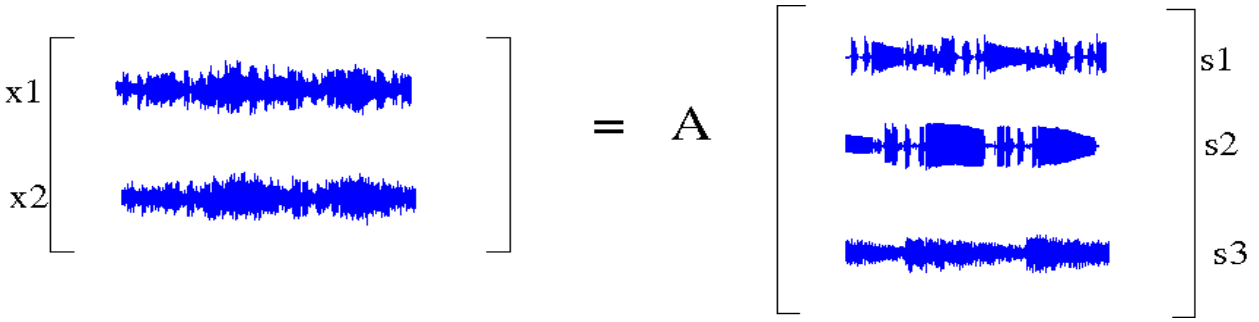



On peut alors estimer la matrice $A$ et les sources $S$ par un simple clustering :

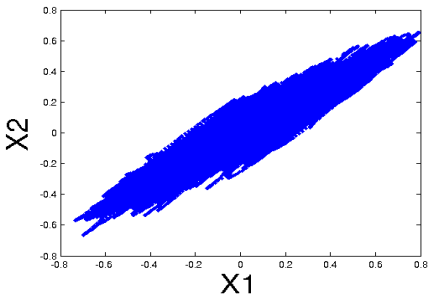
$$\hat{s}_i = \{x_i[t] : x_i[t] \in \mathcal{C}_i\}$$

# Mais en réalité...

Les mélanges sont du type




Si l'on dessine $x_2$ en fonction de $x_1$ on obtient :




clustering impossible !

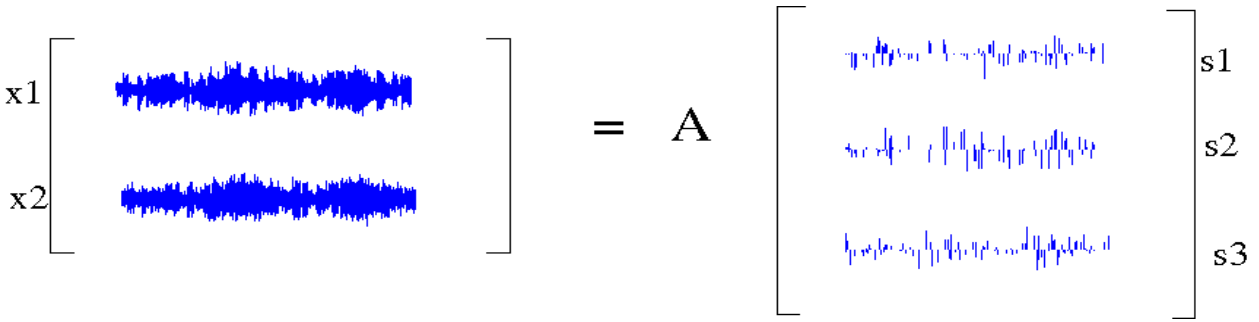
La méthode ne fonctionne que si "une seule source est active pour un $t$ donné"

Autrement dit : les sources doivent être *parcimonieuses*

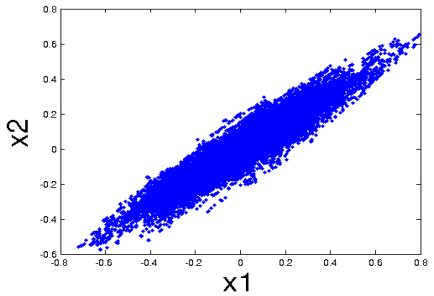
# Quelle parcimonie ?

Hypothèse de séparabilité (1) : un seul signal actif à un instant $t$ donné
$\Rightarrow$ les signaux doivent être parcimonieux.

Si la parcimonie n'est "que" sur les sources : cas particuliers où les sources sont parcimonieuses, mais actives **en même temps**




On dessine $x_2$ en fonction de $x_1$ :




clustering impossible !

Comment rendre les signaux parcimonieux à travers les canaux ?

# Résumons le problème

## Pour l'instant

Le problème est

$$X = AS$$

Mais la séparation est impossibe car l'hypothèse de séparabilité (1) n'est pas vérifiée.

## But

Pour vérifier l'hypothèse (1), *il faut* trouver une représentation de $S$ parcimonieuse.
ie : trouver une transformation $\Phi$ telle que $S = \underline{\mathbf{s}}\Phi^T$ où $\underline{\mathbf{s}}$ est "parcimonieux"

Le problème devient alors

$$X = A\,\underline{\mathbf{s}}\Phi^T$$

# Concentration de l'information

On représente les densités de probabilité des échantillons d'un signal et des coefficients MDCT de ce même signal

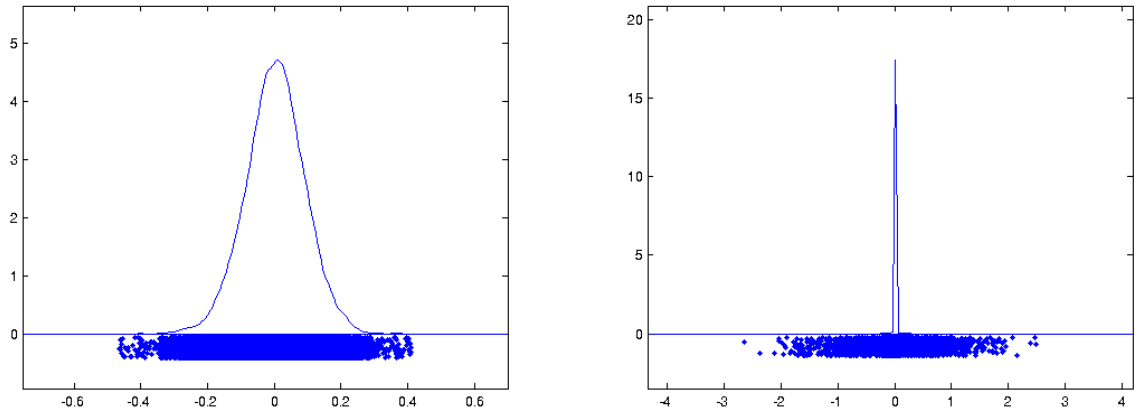



Figure : *Gauche : pdf des échantillons. Droite : pdf des coefficients MDCT.*

Une "bonne transformée" permet de concentrer l'information dans quelques coefficients

Examples de transformée

- Fourier ;
- Gabor ou STFT ;
- MDCT ;
- Ondelettes ;
- *-lettes ;

# Masquage temps-fréquence - (1)

Si l'on trouve une représentation parcimonieuse, on espère que dans ce nouveau repère l'hypothèse (1) est vérifiée.

## Exemple avec une transformation dans une base orthonormée

Soit $\Phi$ une matrice dont les colonnes sont formées par les atomes d'une BON. On a $S = \underline{\mathbf{s}}\Phi$

$$X = AS = A\underline{\mathbf{s}}\Phi \Longleftrightarrow X\Phi^T = A\underline{\mathbf{s}} \Longleftrightarrow \underline{\mathbf{x}} = A\underline{\mathbf{s}}$$

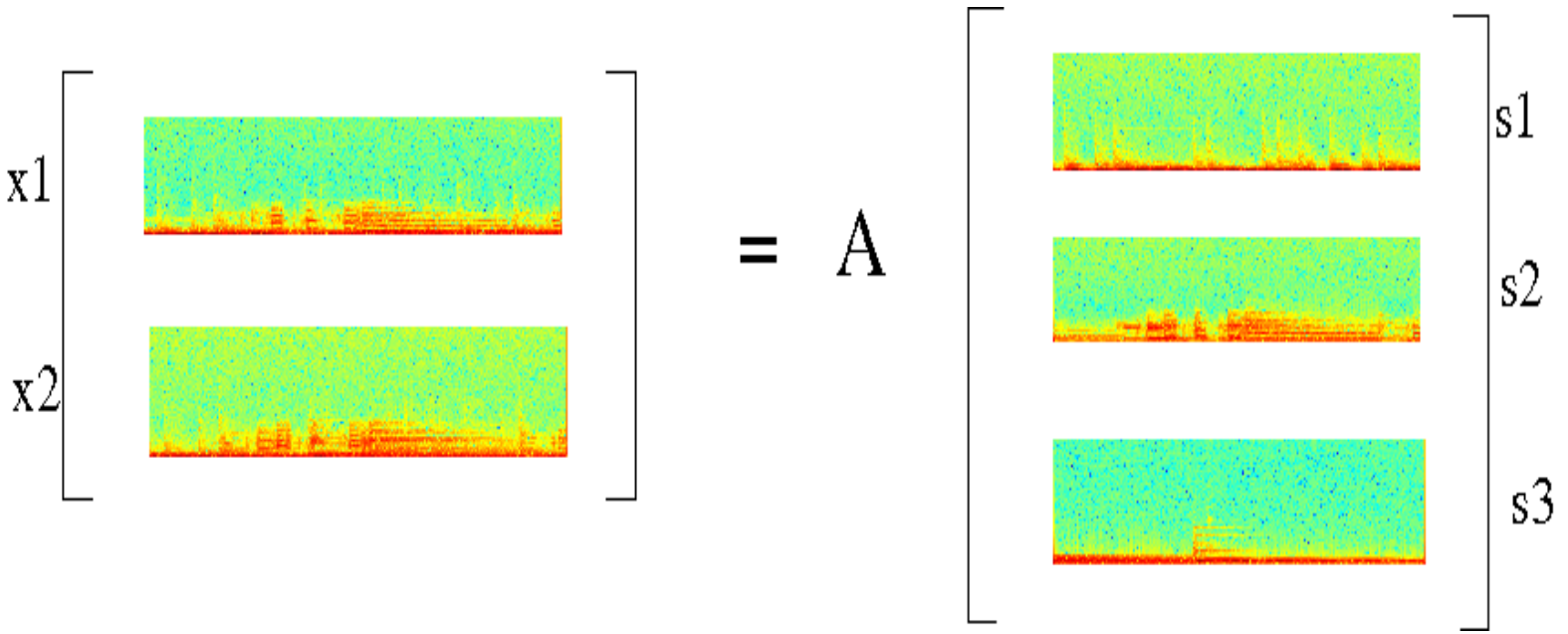
$$\Leftrightarrow \begin{pmatrix} \underline{x}_1[k_1, f_1] & \dots & \underline{x}_1[k_K, f_F] \\ \vdots & & \\ \underline{x}_M[k_1, f_1] & \dots & \underline{x}_M[k_k, f_F] \end{pmatrix} = A \begin{pmatrix} \underline{s}_1[k_1, f_1] & \dots & \underline{s}_1[k_K, f_F] \\ \vdots & & \\ \vdots & & \\ \underline{s}_N[k_1, f_1] & \dots & \underline{s}_N[k_K, f_F] \end{pmatrix}$$

Si l'on choisi un repère non orthogonal (ex : de Gabor), on écrit en général l'équation du mélange directement dans ce repère :

$$\underline{\mathbf{x}} = A\underline{\mathbf{s}} \quad \text{où} \quad \underline{\mathbf{x}} = X\Phi^T, \quad \Phi \quad \text{non orthogonale}$$

# Masquage temps-fréquence - (2)

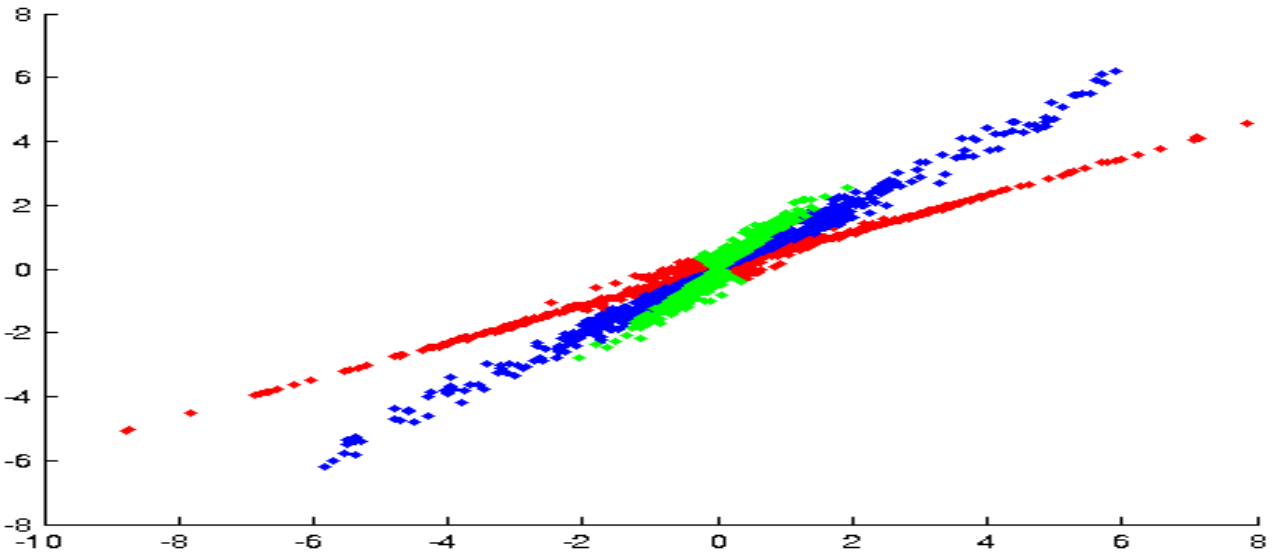
En image :




On espère que l'hypothèse "une seule source active par case temps fréquence" est vérifiée

# Masquage temps-fréquence - (3)

On dessine $\underline{\mathbf{x}}_2$ en fonction de $\underline{\mathbf{x}}_1$




Les directions de la matrice de mélange apparaissent !

Le clustering et la séparation deviennent possible.

$$\hat{\underline{\mathbf{s}}}_i = \{\underline{x}_i[k,f] : \underline{x}_i[k,f] \in \mathcal{C}_i\}$$
$$\hat{\mathbf{s}}_i = \hat{\underline{\mathbf{s}}}_i \Phi$$

# Avantages et inconvénients du masquage temps-fréquence

## Avantages

- Très simple à mettre en oeuvre ;
- Très rapide ;
- Permet l'estimation de la matrice de mélange $A$ **et** de sources $S$.

## Inconvénients

- Le clustering n'est pas évident ;
- Suppose **une seule** source active par case temps-fréquence, les estimations sont *très* (voire trop) parcimonieuses ;
- La "sur"-parcimonie limite les perfromances théoriques de la qualité des sources estimées.

Comment utiliser plusieurs sources actives par case temps-fréquence ?

# Une autre idée...

Plutôt que de supposer une seule source active et se "limiter" à l'équation

$$\underline{x}_1[k,f] = a_{1,i}\ \underline{s}_i[k,f] \Longrightarrow \underline{s}_i[k,f] = \underline{x}_1[k,f]/a_{1,i}$$

On peut supposer qu'il y a exactement autant de sources actives qu'il y a d'observations.

**Supposant la matrice de mélange $A$ connue**, cette hypothèse permet d'inverser localement $A$ :

$$\begin{pmatrix} \underline{x}_1[k,f] \\ \underline{x}_2[k,f] \end{pmatrix} = \begin{pmatrix} a_{1i} & a_{1j} \\ a_{2i} & a_{2j} \end{pmatrix} \begin{pmatrix} \underline{s}_i[k,f] \\ \underline{s}_j[k,f] \end{pmatrix} \Longrightarrow \begin{pmatrix} \hat{\underline{s}}_i[k,f] \\ \hat{\underline{s}}_j[k,f] \end{pmatrix} = A_{ij}^{-1} \begin{pmatrix} \underline{x}_1[k,f] \\ \underline{x}_2[k,f] \end{pmatrix}$$

Comment choisir ces sources ?

# Basis Pursuit

L'hypothèse de parcimonie se traduit ici par $M$ sources actives pour $M$ observations.

Choisir autant de sources qu'il y a d'observations revient à choisir une base parmis les vecteurs $\{a_{:,1}, \ldots, a_{:,M}\}$ dans laquelle représenter les observations $\{\underline{x}_1[k,f], \ldots, \underline{x}_M[k,f]\}$.

On va donc choisir la "meilleure base" au sens où l'on cherche à minimiser une mesure de diversité $E$ des sources $\underline{\mathbf{s}}[k,f]$. Avec $E = \|.\|_p^p$, $p < 2$ :

$$\min_{\underline{\mathbf{s}}[k,f]} \|\underline{\mathbf{s}}[k,f]\|_p^p$$
$$\text{sous contrainte} \quad \underline{\mathbf{x}}[k,f] = A\underline{\mathbf{s}}[k,f]$$

Quand $p = 1$, on reconnait le problème du *Basis Pursuit* (Chen et Donoho, 1998)

# Résolution par FOCUSS

L'algorithme FOCUSS (Rao *et al*, 1999) permet de résoudre le problème précédent

On a $\nabla_{\underline{s}_i[k,f]} \|\underline{\mathbf{s}}_{k,f}\|_p^p = |p| \Pi(\underline{\mathbf{s}}_{k,f}) \underline{\mathbf{s}}_{kf}$ avec $\Pi(\underline{\mathbf{s}_{kf}}) = \mathrm{diag}(|\underline{s_{kf}}_i|^{p-2})$

## FOCUSS

En partant de $\underline{\mathbf{s}}^{(0)}[k,f] = A^T(AA^T)^{-1}\underline{\mathbf{x}}[k,f]$

$$\underline{\mathbf{s}}_{kf}^{(r+1)} = \Pi^{-1}(\underline{\mathbf{s}}_{kf}^{(r)}) A^T \left( A \Pi^{-1}(\underline{\mathbf{s}}_{kf}^{(r)}) A^T \right)^{-1} \underline{\mathbf{x}}[k,f]$$

# Avantages et inconvénients

## Avantages

- Permet d'améliorer les estimations ;
- L'inversion locale permet une "remise à l'échelle" mieux adaptée ;
- Le choix de deux sources actives au lieu d'une seule permet théoriquement d'améliorer significativement les estimations.

## Inconvénients

- Plus lent que le masquage simple (recherche des $M$ "meilleures" sources actives au lieu d'une seule) ;
- Ne permet pas d'estimer $A$
- Il y a plus d'artefact dus aux sources voisines dans les estimations (mais compensé par le fait que l'estimation est meilleure...) ;

Et dans le cas où le mélange n'est plus instantané ?

# mélange convolutif : problème

On considère le cas où $A$ est un "sytème de filtres" :

$$X = A \star S$$

$$X = \begin{pmatrix} \overrightarrow{a}_{11} & \dots & \dots & \overrightarrow{a}_{1N} \\ \vdots & & & \\ \overrightarrow{a}_{M1} & \dots & \dots & \overrightarrow{a}_{MN} \end{pmatrix} \star S$$

où $X \in \mathbb{R}^{M \times T_x}$, $S \in \mathbb{R}^{N \times T_s}$ et $\overrightarrow{a}_{ij} \in \mathbb{R}^K$. On a $T_x = T_s + K - 1$.
Dans le cas "anéchoïque" $A$ est du type $\overrightarrow{a}_{ij} = (0, \dots, 0, 1, 0 \dots, 0)$.

On considère toujours l'hypothèse de parcimonie, on choisi donc une matrice $\Phi$ telle que :

$$X = A \star (\underline{s}\Phi)$$

Mais on ne peut plus écrire

$$X\Phi^T = A \star \underline{\mathbf{s}} \quad (1)$$

$$\underline{\mathbf{x}} = A \star \underline{\mathbf{s}} \quad (2)$$

Comment exploiter la parcimonie ?

# Une première approximation

On considère la transformée de Fourier à court terme, qu'on applique à $X$ et à $A \star S$. On peut alors considérer, $\forall k, f$ :

$$\underline{\mathbf{x}}[k, f] \simeq A[f]\ \underline{\mathbf{s}}[k, f]$$

On peut alors résoudre :

$$\begin{aligned} \min_{\underline{s}} \quad & \|\underline{s}[t, f]\|_1 \\ s.c. \quad & \|\underline{x}[t, f] - A[f]\underline{s}[t, f]\| \leq \varepsilon \end{aligned}$$

Mais ça reste une approximation + Problème s'il y a du bruit...

# Vers une fonctionnelle convexe

Remarque : la parcimonie est "à travers les canaux". On peut alors écrire le problème

$$\min_{\underline{s}} \quad \|\underline{s}[t,f]\|_1 \quad s.c. \quad \|\underline{x}[t,f] - A[f]\ \underline{s}[t,f]\|_2^2 \leq \varepsilon$$

comme

$$\min_{\underline{s}} \sum_{t,f} \sum_{c} |\underline{x}[t,f] - A[f]\ \underline{s}[t,f]|^2 + \lambda \sum_{t,f} \left( \sum_{c} |\underline{s}_c[t,f]| \right)$$

Problèmes :

- $\lambda$ doit être "assez" petit pour avoir la contrainte d'égalité $\underline{x}[t,f] = A\ \underline{s}[t,f]$.
- Si on "relache" cette contrainte, le risque est de trop pénaliser les hautes fréquences.

Solution : utiliser une norme mixte

$$\min_{\underline{s}} \sum_{t,f} \sum_{c} |\underline{x}[t,f] - A[f]\ \underline{s}[t,f]|^2 + \lambda \sum_{t,f} \left( \sum_{c} |\underline{s}_c[t,f]| \right)^2$$

# Vers une fonctionnelle convexe

Solution : utiliser une norme mixte

$$\min_{\underline{s}} \sum_{t,f} \sum_{c} |\underline{x}[t,f] - A[f]\ \underline{s}[t,f]|^2 + \lambda \sum_{t,f} \left( \sum_{c} |\underline{s}_c[t,f]| \right)^2$$

Problème : le terme d'attache aux données traduit mal le bruit supposé gaussien et utilise toujours la même approximation.

Solution : écrire une attache aux données dans le domaine "temporel"

$$\min_{\underline{s}} \|X - A\ \star (\underline{s}\Phi)\|^2 + \lambda \|\underline{s}\|_{12}^2$$

# Algorithmes

Pour résoudre le problème d'optimisation :

- Algorithmes de seuillage itératifs classiques : beaucoup trop lent ! (surtout avec la convolution)
- Utiliser les shéma de Nesterov (algo du 1er ordre optimal).

Les deux sont basés sur la notion d'opérateur de proximité :

**definition : Proximity operator**

Let $\varphi : \mathbb{C}^I \to \mathbb{C}$ be a lower semicontinuous, convex function. The proximity operator associated with $\varphi$ denoted by $\mathrm{prox}_\varphi : \mathbb{C}^I \to \mathbb{C}^I$ is given by

$$\mathrm{prox}_\varphi(\mathbf{z}) = \frac{1}{2} \underset{\mathbf{u}\in\mathbb{C}^I}{\mathrm{argmin}} \|\mathbf{z}-\mathbf{u}\|_2^2 + \varphi(\mathbf{u}). \tag{3}$$

# Algorithmes

## Prox du LASSO $\min_{\mathbf{u}} \|\mathbf{z}-\mathbf{u}\|_2^2 + \lambda\|\mathbf{u}\|_1$

$$\hat{u}_{t,f,c} = \mathrm{sgn}(z_{t,f,c})\left(|z_{t,f,c}| - \lambda\right)^+$$

## Prox du E-LASSO $\min_{\mathbf{u}} \|\mathbf{z}-\mathbf{u}\|_2^2 + \lambda\|\mathbf{u}\|_{1,2}$

$$\hat{u}_{t,f,c} = \mathrm{sgn}(z_{t,f,c})\left(|z_{t,f,c}| - \frac{\lambda}{1+\lambda L_{t,f}}\|\mathbf{z}_{t,f}\|\right)^+$$

# Algorithmes

But : trouver $\operatorname{argmin} \mathcal{L}(\mathbf{s}) + \lambda \mathcal{P}(\mathbf{s})$, avec $\mathcal{L}$ $L$-Lipshitz différentiable

## ISTA (en $O(1/k)$)

Initialization : $\tilde{\mathbf{s}}^{(0)} \in \mathbb{C}^{N \times B}$, $k = 1$.

Repeat :

$$\tilde{\mathbf{s}}^{(k)} = \operatorname{prox}_{\frac{\lambda}{L}\mathcal{P}}\left(\tilde{\mathbf{s}}^{(k-1)} - \frac{\nabla \mathcal{L}(\tilde{\mathbf{s}}^{(k-1)})}{L}\right)$$

## FISTA (en $O(1/k^2)$)

Initialization : $\tilde{\mathbf{s}}^{(0)} \in \mathbb{C}^{N \times B}$, $\mathbf{z}^{(0)} = \tilde{\mathbf{s}}^{(0)}$, $t^{(0)} = 1$, $k = 1$.

Repeat :

- $\tilde{\mathbf{s}}^{(k)} = \operatorname{prox}_{\frac{\lambda}{L}\mathcal{P}}\left(\mathbf{z}^{(k-1)} - \frac{\nabla \mathcal{L}(\mathbf{z}^{(k-1)})}{L}\right)$
- $\tau^{(k)} = \frac{1+\sqrt{1+4\tau^{(k-1)2}}}{2}$
- $\mathbf{z}^{(k)} = \tilde{\mathbf{s}}^{(k)} + \frac{\tau^{(k-1)}-1}{\tau^{(k)}}(\tilde{\mathbf{s}}^{(k)} - \tilde{\mathbf{s}}^{(k-1)})$
- $k = k + 1$

# Results

Table : *Average SDR in decibels as a function of* $\mathrm{RT}_{60}$ *and d over speech mixtures with* $N = 4$ *sources.*

| $\mathrm{RT}_{60}$ | $d$ | narrowband | | | | wideband | |
|---|---|---|---|---|---|---|---|
| | | DUET | $\ell_1$ min. | Lasso | E-Lasso | Lasso | E-Lasso |
| anechoic | 5 cm | 5.4 | 4.5 | **6.4** | **6.7** | 5.7 | **6.5** |
| | 1 m | 3.6 | 7.7 | 7.8 | 8.0 | 7.6 | **8.1** |
| 50 ms | 5 cm | 5.1 | 4.2 | 4.3 | 4.3 | 4.4 | 4.5 |
| | 1 m | 3.1 | 6.3 | 6.4 | 6.4 | **7.0** | **7.4** |
| 250 ms | 5 cm | 2.4 | 1.8 | 2.1 | 2.1 | **5.9** | 5.0 |
| | 1 m | 1.0 | 2.8 | 3.0 | 3.4 | **7.6** | **7.2** |

Table : *Average SDR, SIR and SAR in decibels over speech mixtures with* $N = 4$ *sources,* $\mathrm{RT}_{60} = 250$ *ms and* $d = 1$ *m.*

| method | narrowband | | | | wideband | |
|---|---|---|---|---|---|---|
| | DUET | $\ell_1$ min. | Lasso | E-Lasso | Lasso | E-Lasso |
| SDR | 1.0 | 2.8 | 3.0 | 3.4 | **7.6** | **7.2** |
| SIR | 8.2 | 6.4 | 6.8 | 7.7 | **14.0** | **13.9** |
| SAR | 2.8 | 6.5 | 6.4 | 6.4 | **9.1** | 8.5 |

# Results

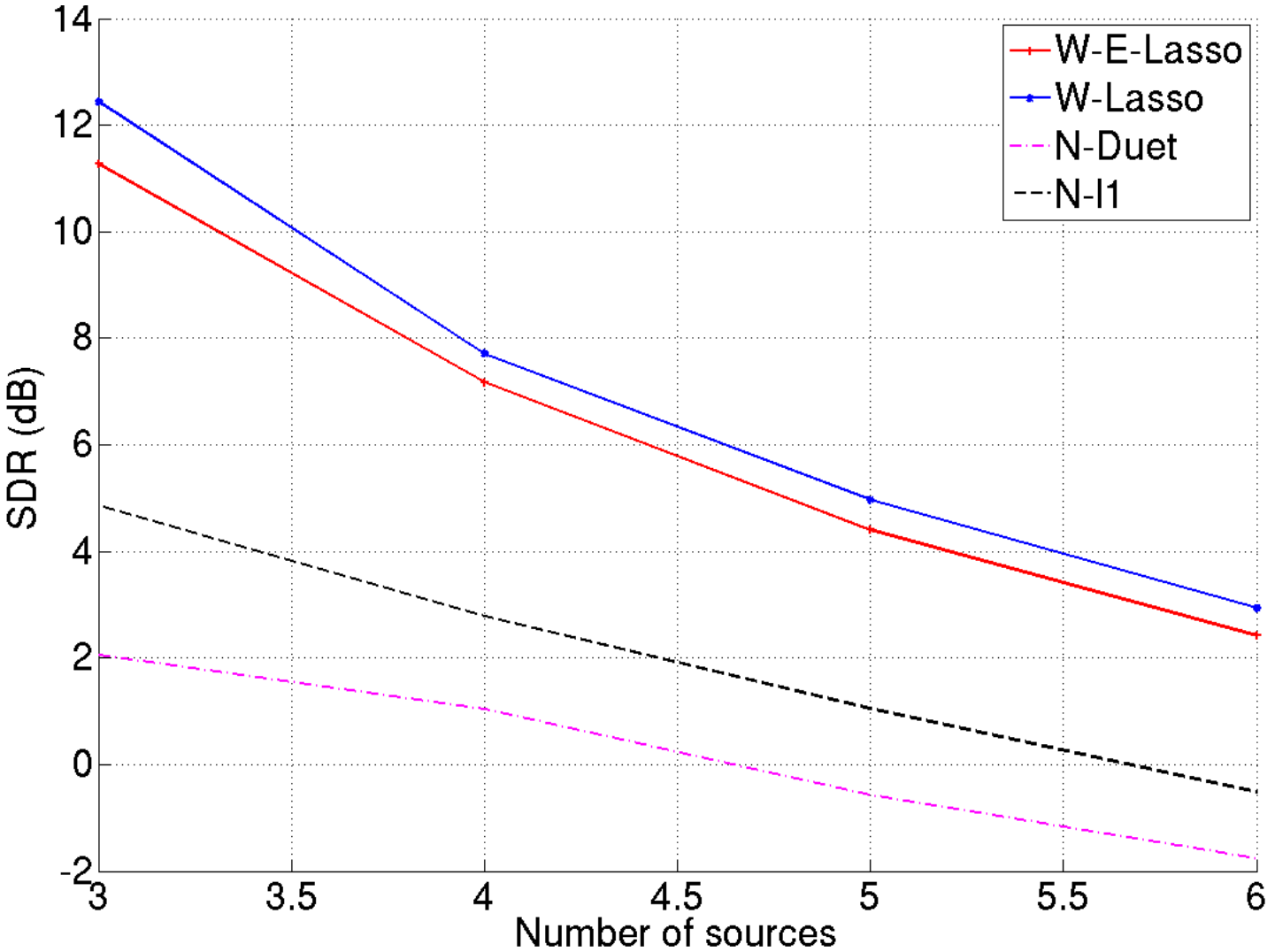



Figure : *Variation of the average SDR as a function of N over speech mixtures with* $\mathrm{RT}_{60} = 250$ *ms and* $d = 1$ *m.*

# Résultats

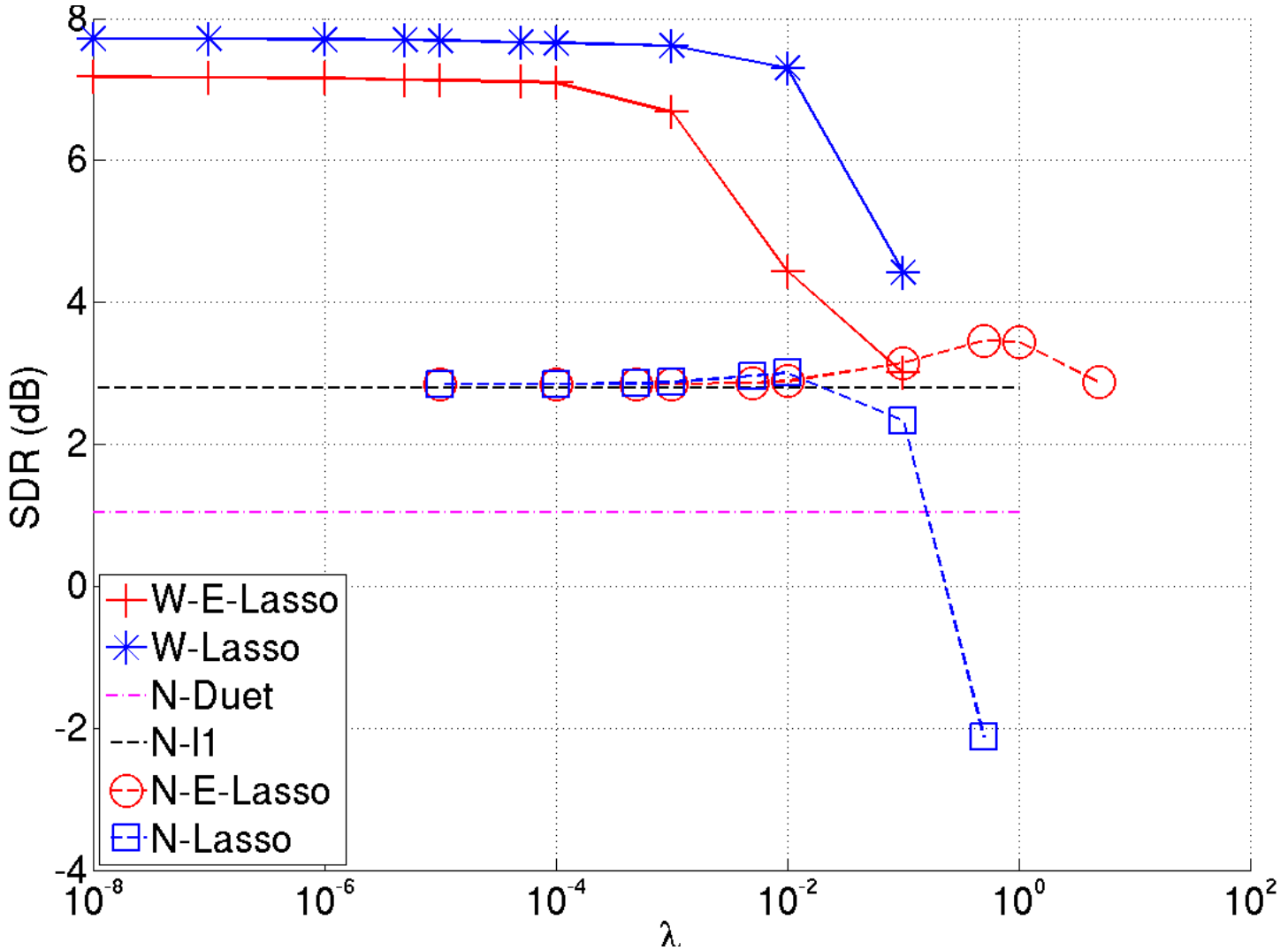



Figure : *Variation of the average SDR as a function of* $\lambda$ *over speech mixtures with* $N = 4$ *sources,* $\mathrm{RT}_{60} = 250$ *ms and* $d = 1$ *m.*

# Results

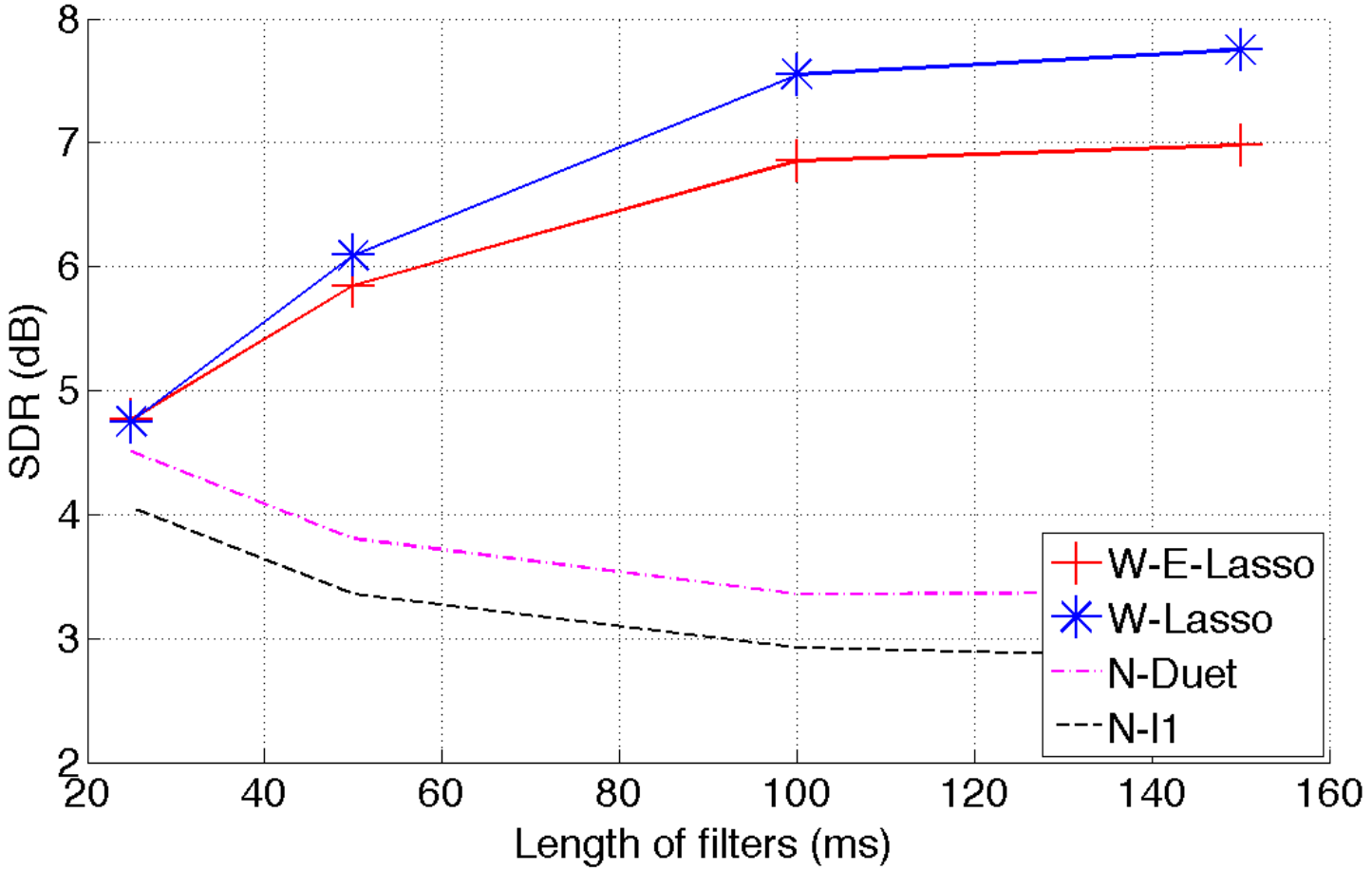



Figure : *Variation of the average SDR as a function of the length of the filters over speech mixtures with* $\mathrm{RT}_{60} = 250$ *ms and* $d = 1$ *m.*

# Résultats

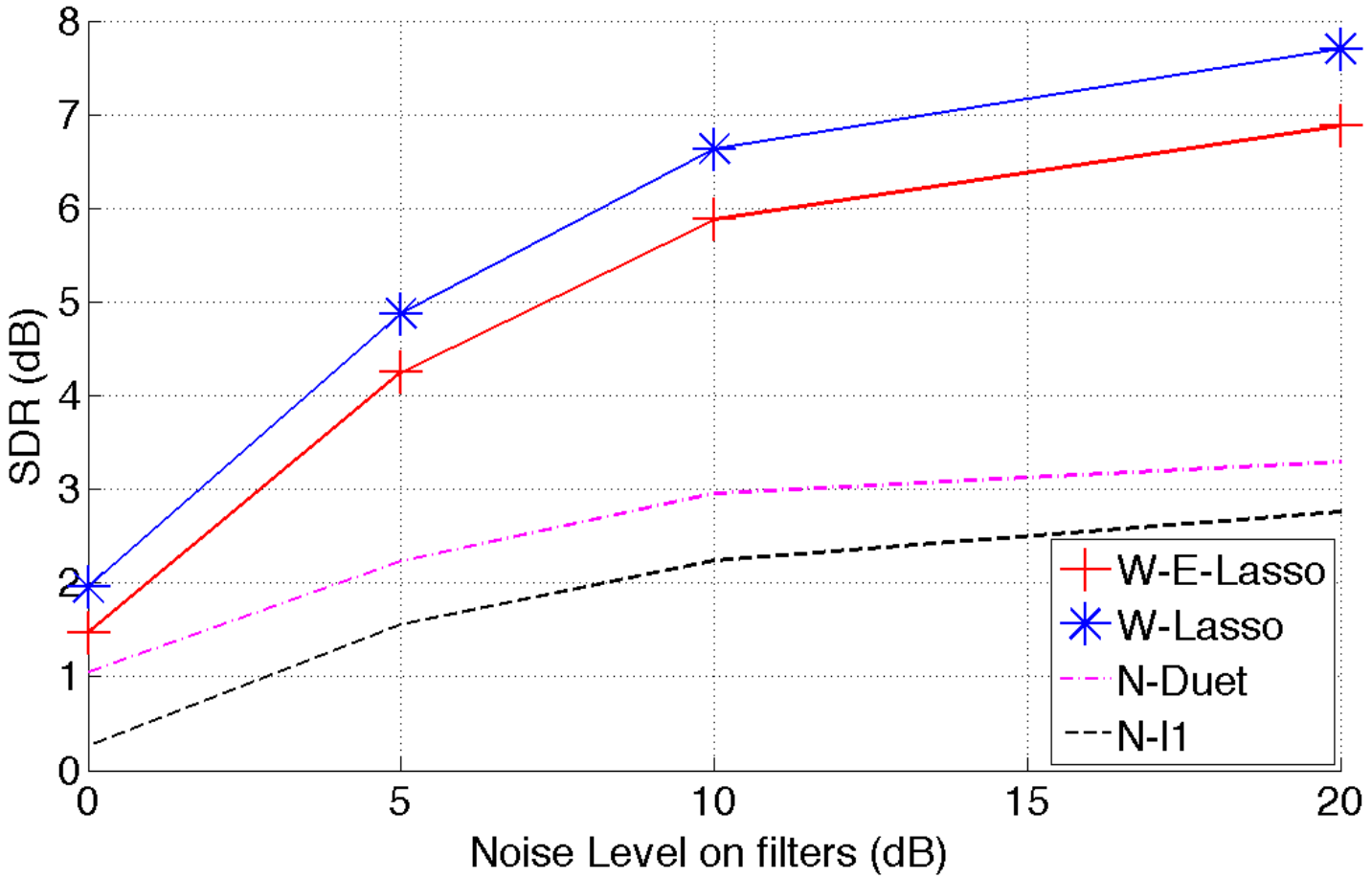



Figure : *Variation of the average SDR as a function of the noise level over speech mixtures with* $\mathrm{RT}_{60} = 250$ *ms and* $d = 1$ *m.*

# Résultats

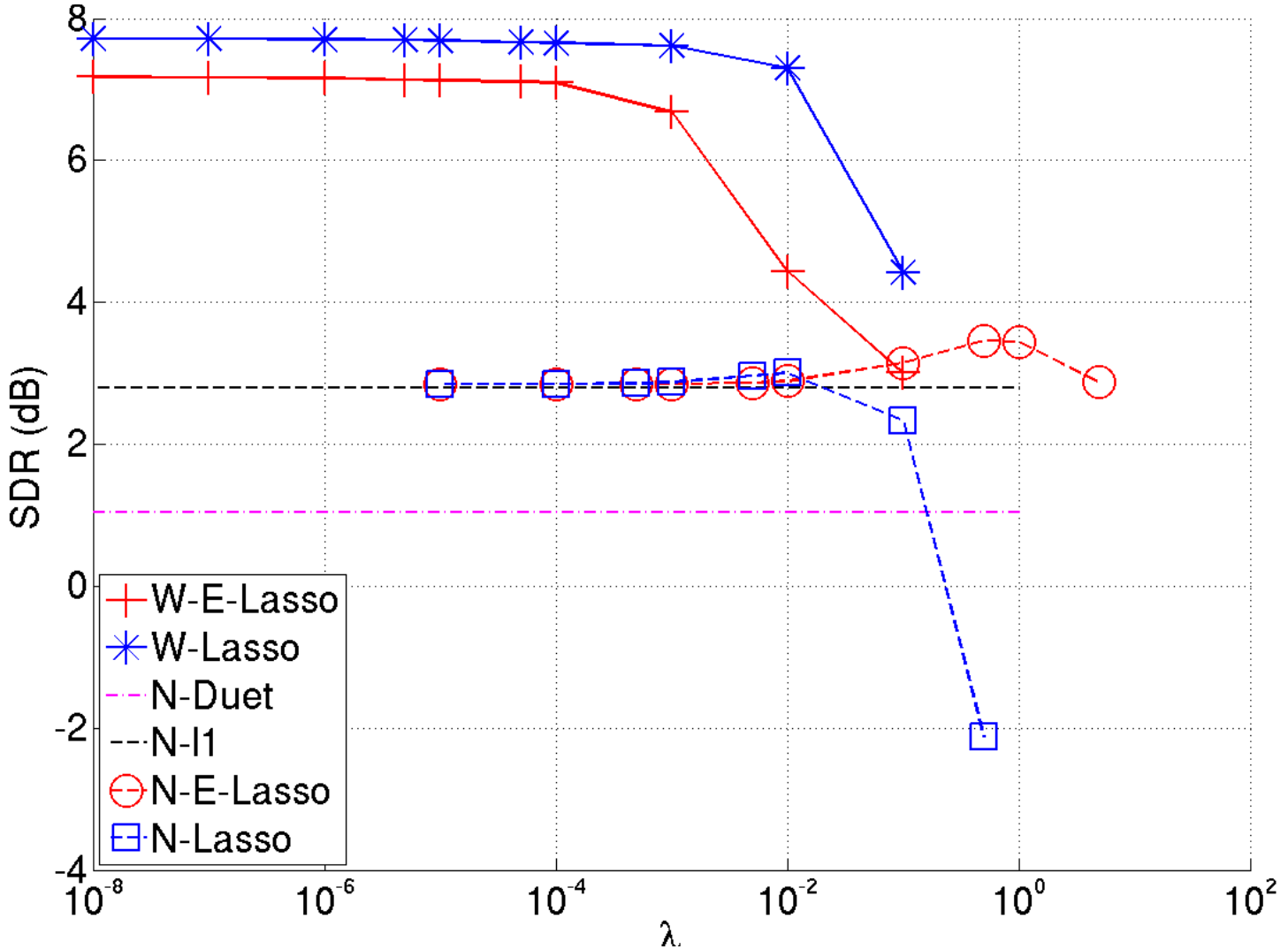



Figure : *Variation of the average SDR as a function of* $\lambda$ *over speech mixtures with* $N = 4$ *sources,* $\mathrm{RT}_{60} = 250$ *ms and* $d = 1$ *m.*