Construction d'estimateurs oracles pour la séparation de sources

Emmanuel VINCENT[1], Rémi GRIBONVAL[2]

[1]Centre for Digital Music, Queen Mary, University of London
Mile End Road, London E1 4NS, United Kingdom

[2]Projet METISS, IRISA-INRIA
Campus de Beaulieu, 35042 Rennes Cedex, France
emmanuel.vincent@elec.qmul.ac.uk, remi.gribonval@irisa.fr

Résumé – La séparation de sources pour des mélanges sous-déterminés et/ou convolutifs est un problème difficile pour lequel de nombreux algorithmes ont été proposés. Afin d'étudier leur performance, nous définissons des estimateurs oracles permettant de calculer la performance maximale théorique de différentes classe d'algorithmes de séparation dans un cadre d'évaluation où les sources de référence sont disponibles. Nous implémentons ces estimateurs pour deux classes (les algorithmes de séparation par filtrage stationnaire et ceux procédant par masquage temps-fréquence) et nous étudions leur performance sur quelques exemples de mélanges audio.

Abstract – Source separation of under-determined and/or convolutive mixtures is a difficult problem that has been addressed by many algorithms. In order to study their performance, we define oracle estimators that compute the maximal theoretical performance achievable for various classes of algorithms in an evaluation framework where the reference sources are available. We implement these estimators for two classes (stationary filtering separation algorithms and time-frequency masking separation algorithms) and we study their performance on a few audio mixture examples.

1 Introduction

La majorité des signaux audio, vidéo et biomédicaux sont des mélanges de plusieurs sources actives simultanément, acquis directement ou par mélange synthétique de signaux sources. En toute généralité, l'opération de mélange est une transformation non linéaire et non stationnaire des signaux sources. Cependant, la plupart des mélanges sont modélisables par un filtrage linéaire stationnaire. Le $i$ème canal du mélange observé s'exprime alors sous la forme $x_i = \sum_j a_{ij} \star s_j$, où (s_j) sont les signaux sources, (a_{ij}) les filtres de mélange et $\star$ représente la convolution. Le mélange est dit instantané lorsque les filtres de mélange sont de simples gains et convolutif dans le cas contraire. Il est aussi dit sur-déterminé lorsque le nombre de canaux observés est supérieur au nombre de sources et sous-déterminé sinon. L'étude des signaux mélangés soulève le problème de la séparation de sources, qui consiste à estimer le signal de chaque source avec la meilleure qualité possible (*i.e.* sans interférences provenant des autres sources ni distortion supplémentaire).

De nombreux algorithmes ont été proposés pour résoudre ce problème. Les mélanges sur-déterminés instantanés ou convolutifs sont généralement séparés par convolution des canaux de mélange par des filtres de démixage stationnaires à réponse impulsionnelle finie [1]. L'Analyse en Composantes Indépendantes (ACI) estime ces filtres de démixage en modélisant les sources par des processus i.i.d. non gaussiens [2, 1] ou des processus gaussiens de variance non stationnaire [3] indépendants. D'autres méthodes utilisent des modèles de sources plus complexes spécifiques à un type de sources donné [4, 5]. La séparation de mélanges sous-déterminés est plus souvent effectuée par masquage binaire temps-fréquence [6] ou par filtrage de Wiener adaptatif [7]. Les masques temps-fréquence sont alors estimés en fonction de la différence d'intensité [6] et de phase [8] entre les canaux du mélange observé en chaque point temps-fréquence, ou bien à l'aide de méthodes plus complexes prenant en compte la structure du spectre de puissance à court terme des sources [9, 7, 10].

La performance de ces diverses méthodes dépend beaucoup du type de mélange rencontré, voire des propriétés des sources et des filtres de mélange dans le mélange particulier étudié. Par exemple la performance de l'ACI sur des mélanges sur-déterminés convolutifs chute généralement lorsque la taille des filtres de mélange dépasse quelques centaines d'échantillons. Trois raisons sont en mesure d'expliquer cette baisse de performance. Premièrement, l'algorithme utilisé pour estimer les filtres de démixage peut échouer à trouver les filtres optimaux en pratique (problèmes de maxima locaux par exemple). Deuxièmement, les filtres de démixage optimaux en terme de performance peuvent ne pas correspondre aux filtres optimaux estimés en fonction du modèle de sources de l'ACI. Troisièmement, les contraintes imposées par le choix de la méthode de séparation (par exemple la relativement faible taille des filtres de séparation) peuvent limiter dans l'absolu la performance optimale possible. Pour améliorer les algorithmes d'ACI existants, il est indispensable de comprendre laquelle de ces raisons est prépondérante. En effet, un tel diagnostic permettrait de savoir s'il faut prioritairement construire de meilleurs algorithmes d'optimisation, de meilleurs modèles de sources, ou modifier l'hypothèse de séparation par filtrage stationnaire. Le même type de raisonnement s'applique aux autres méthodes de séparation.

Dans cet article, nous apportons une réponse partielle à cette question en construisant des estimateurs oracles des sources,

c'est-à-dire des estimateurs idéaux au sein d'une classe d'estimateurs possibles. Ces estimateurs ne peuvent par nature être utilisés que dans un contexte d'évaluation où les sources de référence sont disponibles. L'étude de la performance des estimateurs oracles permet de prédire l'adéquation d'une classe d'algorithmes de séparation à un mélange donné, ainsi que de situer les performances d'un algorithme donné par rapport à l'optimum possible dans la classe dont il fait partie. On aboutit ainsi également à une sorte d'indice mesurant la difficulté de séparation d'un mélange.

La structure de l'article est la suivante. Nous définissons les estimateurs oracles dans la partie 2. Puis dans la partie 3 nous proposons des algorithmes pour calculer les oracles de séparation par filtrage stationnaire et par masquage temps-fréquence. Enfin nous étudions leur performance sur des exemples audio dans la partie 4.

2 Définition d'un estimateur oracle

Supposons que nous disposons des canaux d'un mélange observé $(x_i)_{1\leq i\leq I}$ et de la $j$ième source de référence s_j. Les nombreuses méthodes de séparation de sources citées dans la partie 1 se ramènent à deux classes générales : filtrage stationnaire et masquage temps-fréquence. Au sein d'une classe donnée, tous les algorithmes estiment la $j$ième source sous la forme

$$\widehat{s_j} = f_{\theta_j}(x_1, \ldots, x_I), \tag{1}$$

où θ_j représente les paramètres de séparation permettant de calculer la source séparée en fonction des canaux observés, *i.e.* les coefficients des filtres de démixage ou des masques temps-fréquence utilisés. La différence entre les algorithmes tient à la façon dont ces paramètres sont estimés à partir des données observées $x_1, \ldots, x_I$.

Puisque la source de référence est supposée connue, il est possible d'évaluer la performance de séparation en mesurant la qualité de la source estimée à l'aide d'une mesure $q(\widehat{s_j}, s_j)$. L'estimateur oracle de la $j$ième source est alors la fonction qui calcule la meilleure estimation de la source parmi un ensemble de paramètres de séparation possibles Θ :

$$\widetilde{s_j}(\Theta) = \arg\max_{\theta_j\in\Theta} q(f_{\theta_j}(x_1, \ldots, x_I), s_j). \tag{2}$$

L'ensemble Θ fixe des contraintes sur les paramètres de séparation, comme la taille des filtres de démixage ou la largeur des trames utilisées pour le masquage temps-fréquence. L'étude de l'oracle consiste à calculer sa performance

$$\tilde{q}(\Theta) = q(\widetilde{s_j}(\Theta), s_j) = \max_{\theta_j\in\Theta} q(f_{\theta_j}(x_1, \ldots, x_I), s_j) \tag{3}$$

en fonction du mélange observé et des contraintes posées.

Notons que le calcul de l'oracle ne nécessite pas de connaître toutes les données du problème, mais seulement les données à estimer (dans ce cas la $j$ième source de référence mais pas les autres sources ni les filtres de mélange). Il est possible de définir de la même façon des estimateurs oracles pour d'autres quantités, par exemple pour l'image de la $j$ième source sur le $i$ème capteur $a_{ij} \star s_j$.

3 Calcul de l'estimateur

La première question qui se pose en pratique pour calculer un estimateur oracle est de savoir comment mesurer la qualité d'une source estimée. Généralement les sources séparées sont destinées à l'écoute et la quantité $q(\widehat{s_j}, s_j)$ devrait donc être proportionnelle à la distortion auditive perçue entre $\widehat{s_j}$ et s_j [11]. Cette quantité reste difficile à calculer actuellement, c'est pourquoi nous utilisons par la suite un simple Rapport Signal-à-Distortion (RSD) quadratique exprimé en déciBels (dB) par

$$q(\widehat{s_j}, s_j) = 10\log_{10}\frac{\|s_j\|^2}{\|\widehat{s_j} - s_j\|^2}, \tag{4}$$

où $\|a\|^2 = \sum_{t=0}^{T-1} a(t)^2$ est l'énergie totale d'un signal a de taille T. L'estimateur oracle de la $j$ième source est alors défini par $\widetilde{s_j}(\Theta) = \arg\min_{\theta_j\in\Theta} \|f_{\theta_j}(x_1, \ldots, x_I) - s_j\|^2$.

3.1 Oracle de filtrage stationnaire

Dans le cas d'une séparation par des filtres de démixage stationnaires (causaux) $(w_{ij})_{1\leq i\leq I}$ de taille L, l'estimée de la $j$ième source vaut $\widehat{s_j}(t) = \sum_{i=1}^{I}\sum_{\tau=0}^{L-1} w_{ij}(\tau)x_i(t-\tau)$. En notant $(x_{i\tau})_{1\leq i\leq I,\, 0\leq\tau\leq L-1}$ les versions retardées des canaux observés définies par $x_{i\tau}(t) = x_i(t-\tau)$ et en représentant le couple d'indices (i,τ) par un seul indice η variant entre 1 et $D = IL$, l'estimée prend aussi la forme

$$\widehat{s_j} = \sum_{\eta=1}^{D} w_{\eta j}\, x_\eta. \tag{5}$$

Le calcul des coefficients $(w_{\eta j})_{1\leq\eta\leq D}$ qui maximisent le RSD est un problème de moindre carrés linéaire dont la solution est donnée classiquement par les coefficients de la projection orthogonale de s_j sur le sous-espace vectoriel engendré par les signaux observés et leurs versions décalées $(x_\eta)_{1\leq\eta\leq D}$. Plus précisément, en notant par $\langle a, b\rangle = \sum_{t=0}^{T-1} a(t)b(t)$ le produit scalaire de deux signaux a et b de taille T, le vecteur des coefficients optimaux $\widetilde{\mathbf{w}_j} = [\widetilde{w_{\eta j}}]_{1\leq\eta\leq D}$ vaut

$$\widetilde{\mathbf{w}_j} = \mathbf{G}^{-1}\mathbf{d}, \tag{6}$$

où $\mathbf{G} = [\langle x_\eta, x_{\eta'}\rangle]_{1\leq\eta\leq D,\, 1\leq\eta'\leq D}$ est la matrice de Gram des canaux retardés et $\mathbf{d} = [\langle s_j, x_\eta\rangle]_{1\leq\eta\leq D}$.

3.2 Oracle de masquage temps-fréquence

L'estimation d'une source (par exemple la $j$ième) par masquage temps-fréquence consiste à représenter l'un des canaux observés (par exemple le $i$ème) dans une base de signaux localisés en temps-fréquence, puis à multiplier chaque coefficient de la représentation par un coefficient de masquage réel compris entre 0 et 1, et enfin à estimer la source désirée par inversion de la représentation au sens des moindres carrés. Dans la suite nous utilisons une base orthonormale de cosinus locaux (MDCT), qui présente l'avantage d'être facilement inversible. En appelant $(\phi_n)_{1\leq n\leq N}$ les éléments de la base et $(w_{nj})_{1\leq n\leq N}$ les coefficients de masquage, l'estimée de la $j$ième source vaut $\widehat{s_j}(t) = \sum_{n=1}^{N} w_{nj}\langle x_i, \phi_n\rangle\phi_n(t)$. Puisque la base est orthonormale, l'erreur quadratique sur la source estimée se décompose sous la forme d'une somme

$$\|\widehat{s_j} - s_j\|^2 = \sum_{n=1}^{N} |w_{nj}\langle x_i, \phi_n\rangle - \langle s_j, \phi_n\rangle|^2. \tag{7}$$

Cette somme est minimale lorsque chacun de ses termes est minimal. En notant $r_{nj} = \langle s_j, \phi_n \rangle / \langle x_i, \phi_n \rangle$, le $n$ième coefficient de masquage optimal vaut par conséquent

$$\widetilde{w_{nj}} = \begin{cases} 0 & \text{si } r_{nj} < 0, \\ r_{nj} & \text{si } 0 \leq r_{nj} \leq 1, \\ 1 & \text{si } r_{nj} > 1. \end{cases} \quad (8)$$

4 Étude de performance

Nous étudions la performance des deux estimateurs oracles développés sur quelques mélanges audio musicaux en fonction des contraintes posées (taille des filtres de démixage ou taille de fenêtre de MDCT). Les mélanges utilisés ont été choisis pour rendre compte de divers niveaux de difficulté liés soit à la nature du mélange (déterminé ou sous-déterminé, instantané ou convolutif à filtres courts ou longs) soit à celle des sources (indépendantes ou synchrones et en harmonie).

4.1 Données expérimentales

Nous considérons deux ensembles de sources échantillonnées à 22050 Hz : d'une part un ensemble de sources indépendantes provenant d'extraits solo de CD différents (un violon, un violoncelle, une clarinette), d'autre part un ensemble de sources synchrones et en harmonie synthétisées à partir des pistes d'un même fichier MIDI [1] (un violon, un alto, un violoncelle). Bien que le son des instruments MIDI soit de pauvre qualité, le deuxième ensemble est beaucoup plus réaliste car les sources suivent les règles de l'harmonie musicale.

Pour chaque ensemble de sources, nous effectuons créons deux séries de mélanges stéréo. La première série est composée de mélanges déterminés de deux sources (violon et violoncelle pour l'ensemble de sources indépendantes, violon et alto pour les sources MIDI synchrones). La seconde série est composée de mélanges sous-déterminés des trois sources de l'ensemble considéré.

Dans chaque série, trois mélanges différents de durée égale à 10 s sont engendrés à partir de chaque ensemble de sources par mélange instantané, convolutif court et convolutif long. Les mélanges convolutifs longs sont effectués par convolution avec des réponses impulsionnelles de salle enregistrées à l'IRCAM (temps de réverbération $T_{60} \simeq 800$ ms, soit environ 18000 échantillons) et les mélanges convolutifs courts par convolution avec ces mêmes réponses tronquées à leurs 512 premiers échantillons.

4.2 Résultats

La plupart des algorithmes de séparation de sources à partir de mélanges convolutifs estiment plutôt la contribution de chaque source sur chaque capteur que les sources elles-mêmes. Afin d'obtenir des bornes de performance pour ces algorithmes, nous utilisons les estimateurs oracles de filtrage stationnaire et de masquage temps-fréquence pour estimer l'image de chaque source sur le premier canal. Pour ce faire, nous appliquons la formule (6) (respectivement (8)) avec $\mathbf{d} = [\langle a_{1j} \star s_j, x_\eta \rangle]_{1 \leq \eta \leq D}$ (resp. $r_{nj} = \langle a_{1j} \star s_j, \phi_n \rangle / \langle x_i, \phi_n \rangle$), de façon à optimiser la mesure de qualité $q(\widehat{s_j}, a_{1j} \star s_j)$.

La performance de l'oracle, c'est-à-dire le RSD calculé sur chaque source estimée puis moyenné entre les sources, est calculée pour différentes tailles L des filtres de démixage (oracle de filtrage stationnaire) et différentes tailles L' du support des éléments de la base MDCT (oracle de masquage temps-fréquence). Les résultats pour les mélanges déterminés sont présentés dans le tableau 1, ceux pour les mélanges sous-déterminés dans le tableau 2.

Sans surprise, nous constatons que la performance de tous les oracles considérés est systématiquement meilleure sur les mélanges déterminés que sur les mélanges sous-déterminés : la différence entre les meilleurs oracles est d'environ 4 dB pour les mélanges convolutifs longs, 15 dB pour les mélanges convolutifs courts, et la performance des meilleurs oracles est parfaite pour les mélanges instantanés déterminés.

De même, les oracles ont toujours plus de difficulté à séparer des sources synchrones et en harmonie que des sources indépendantes, la différence de performance étant de l'ordre de 2 dB pour les ensembles de sources considérés.

En général, les mélanges convolutifs sont plus difficiles à séparer que les mélanges instantanés, et ce d'autant plus que les filtres de mélange sont longs. La différence de performance entre les meilleurs oracles pour les mélanges convolutifs courts et longs est importante (de l'ordre de 20 dB) dans le cas déterminé avec des sources indépendantes, plus faible mais significative (de l'ordre de 10 dB) dans le cas déterminé avec des sources synchrones, et négligeable dans le cas sous-déterminé.

Globalement, les seules conditions dans lesquelles une séparation de qualité semble réalisable avec les classes d'algorithmes considérés correspondent à des mélanges déterminés où les filtres de mélange sont courts (ou bien instantanés). Dans ce cas la séparation par filtrage stationnaire peut fournir de meilleurs résultats que le masquage temps-fréquence. Dans les autres cas, les facteurs de difficulté semblent avoir un effet cumulatif sur la dégradation des performances des meilleurs oracles, et le masquage temps-fréquence idéal est aussi bon ou meilleur que le filtrage stationnaire idéal.

5 Conclusion

Les estimateurs oracles proposés dans cet article fournissent un cadre pour calculer les performances idéales de séparation réalisables en théorie avec différentes classes d'algorithmes de séparation de sources. Ils sont disponibles, accompagnés d'exemples sonores, sous forme d'une boîte à outils MATLAB distribuée sous license GPL à l'adresse `http://www.irisa.fr/metiss/bss_oracle/`

En pratique, les algorithmes ne peuvent approcher ces performances que si les modèles de sources sont pertinents, les estimateurs (InfoMax, Maximum A Posteriori, ...) bien choisis et les techniques numériques d'optimisation (descente de gradient, ...) efficaces. D'après les expériences réalisées, le masquage temps-fréquence est potentiellement plus performant que la séparation par filtrage stationnaire sur des mélanges réalistes qui sont souvent sous-déterminés et convolutifs. Le filtrage stationnaire, qui correspond aux méthodes les plus classiques basées sur l'ACI, n'est potentiellement le meilleur que

[1] Il s'agit du fichier *Classical Music n^o 15* de la base de données *RWC* [12].

TAB. 1 – Performance moyenne (RSD en dB) des estimateurs oracles sur des mélanges stéréo déterminés de deux sources.

Type d'oracle	Mélange de sources indépendantes			Mélange de sources synchrones		
	Instantané	Convolutif court	Convolutif long	Instantané	Convolutif court	Convolutif long
Filtrage $L = 128$	$+\infty$	19	10	$+\infty$	15	7
Filtrage $L = 512$	$+\infty$	31	13	$+\infty$	25	10
Filtrage $L = 2048$	$+\infty$	42	15	$+\infty$	28	13
Masquage $L' = 128$	12	12	13	10	10	10
Masquage $L' = 512$	16	15	17	15	14	14
Masquage $L' = 2048$	21	20	20	19	15	16

TAB. 2 – Performance moyenne (RSD en dB) des estimateurs oracles sur des mélanges stéréo sous-déterminés de trois sources.

Type d'oracle	Mélange de sources indépendantes			Mélange de sources synchrones		
	Instantané	Convolutif court	Convolutif long	Instantané	Convolutif court	Convolutif long
Filtrage $L = 128$	12	9	6	9	9	6
Filtrage $L = 512$	15	11	7	11	12	8
Filtrage $L = 2048$	16	13	9	12	13	10
Masquage $L' = 128$	10	9	9	7	8	7
Masquage $L' = 512$	15	12	13	10	10	10
Masquage $L' = 2048$	19	16	17	12	11	12

dans les cas déterminés instantanés ou convolutifs courts.

La performance de séparation par filtrage stationnaire a déjà été étudiée à l'aide d'un estimateur oracle supposant connus les filtres de mélange [13] sous l'hypothèse que les sources sont des bruits blancs indépendants. Cependant l'oracle que nous proposons en supposant connues les sources met en évidence un effet non négligeable de la nature des sources (selon qu'elles sont indépendantes ou synchrones). Une comparaison poussée des deux approches est donc nécessaire, et il faut notamment chercher dans les deux cas à mesurer et prédire la dégradation de performance constatée pour des estimateurs réels proches de l'estimateur oracle. Enfin, nous réfléchissons au calcul effectif des oracles pour d'autres mesures de qualité rendant mieux compte de la qualité perçue des résultats.

Références

[1] S. Amari, S. Douglas, A. Cichocki, and H. Yang, "Novel online algorithms for blind deconvolution using natural gradient approach," in *Proc. SYSID*, 1997, pp. 1057–1062.

[2] J.-F. Cardoso, "Blind source separation : statistical principles," *Proceedings of the IEEE*, vol. 9, no. 10, pp. 2009–2025, 1998.

[3] D.-T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of non stationary sources," *IEEE Trans. on Signal Processing*, vol. 49, no. 9, pp. 1837–1848, 2001.

[4] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, 2001.

[5] M. Reyes-Gomez, B. Raj, and D. Ellis, "Multi-channel source separation by factorial HMMs," in *Proc. ICASSP*, 2003, pp. 664–667.

[6] Ö. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[7] L. Benaroya, "Séparation de plusieurs sources sonores avec un seul microphone," Ph.D. dissertation, Université Rennes I, 2003.

[8] N. Roman, D. Wang, and G. Brown, "Speech segregation based on sound localization," *Journal of the ASA*, vol. 114, no. 4, pp. 2236–2252, 2003.

[9] S. Roweis, "One microphone source separation," in *Proc. NIPS*, 2000, pp. 793–799.

[10] E. Vincent and X. Rodet, "Underdetermined source separation with structured source priors," in *Proc. ICA*, 2004, pp. 327–332.

[11] E. Vincent, C. Févotte, R. Gribonval, L. Benaroya, X. Rodet, A. Röbel, É. LeCarpentier, and F. Bimbot, "Comment évaluer les algorithmes de séparation de sources audio ?" in *Proc. GRETSI*, 2003, pp. 27–32.

[12] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database : Popular, classical, and jazz music databases," in *Proc. ISMIR*, 2002, pp. 287–288. [Online]. Available : http ://staff.aist.go.jp/m.goto/RWC-MDB/

[13] R. Balan, J. Rosca, and S. Rickard, "Robustness of parametric source demixing in echoic environments," in *Proc. ICA*, 2001, pp. 144–148.