Estimateurs oracles pour la séparation de sources monocapteur par approches spectrales à états discrets

Valentin EMIYA, Emmanuel VINCENT, Rémi GRIBONVAL

Équipe-projet METISS, IRISA-INRIA
Campus de Beaulieu, 35042 Rennes Cedex, France
{valentin.emiya,emmanuel.vincent,remi.gribonval}@irisa.fr

Résumé – Dans cet article, des bornes de performances oracles sont déterminées pour la séparation de sources monocapteur sous contrainte d'un nombre fini d'états discrets. En fixant des contraintes qui sont à la base de systèmes existants, les bornes de performances obtenues sont plus réalistes qu'avec une contrainte de masquage temps-fréquence seule. Dans ce contexte, l'efficacité théorique des approches par mélanges de gaussiennes est quantifiée et comparée à des résultats provenant d'un système de l'état de l'art. De futures approches sont envisagées en faisant évoluer ces modèles vers des méthodes discriminantes à états conjoints.

Abstract – In this article, oracle performance bounds are established for single-sensor source separation with discrete, finite-state constraints. When imposing some constraints used in existing systems, we obtain performance bounds that are more realistic than the values related to the time-frequency masking constraint only. In this context, the performance bounds for systems based on Gaussian mixture models are computed and compared to state-of-the-art results. Possible future approaches are proposed, based on joint-state, discriminative methods.

1 Introduction

La séparation de sources consiste à analyser un mélange de signaux et à en estimer chacune des composantes, de façon approchée, les performances atteintes se mesurant par un critère tel que le rapport signal à distortion (RSD) à l'aide des signaux des sources originaux. Si nombre d'approches reposent sur l'analyse en composantes indépendantes [4], ces techniques s'avèrent peu concluantes voire impossible à utiliser dans le cas de mélanges audio monocapteur. Ce dernier problème, qui sous-tend des applications telles que le réhaussement de parole en téléphonie ou dans les dispositifs d'aide aux malentendants, est plutôt traité en considérant des modèles spectraux à états discrets – modèles de mélanges de gaussiennes (MMG) [2], modèles de Markov cachés (MMC) [6] – ou continus [8, 3]. Nous nous intéressons ici aux approches à états discrets qui ont montré leur efficacité, en particulier dans le cas gaussien [5].

Nous supposerons que la séparation de sources est obtenue via une fonction paramétrique f sous la forme $\widehat{s} = f_\lambda(x, \theta)$, où $\widehat{s}$ est l'estimation des sources, x le mélange observé, λ un vecteur de paramètres estimés sur une base d'apprentissage et θ un vecteur de paramètres à estimer sur le signal de test x. Les paramètres λ et θ, que nous appellerons respectivement *paramètres appris* et *paramètres décodés*, prennent leurs valeurs dans des ensembles Λ et Θ, qui peuvent être considérés comme des contraintes sur λ et θ. La description d'un système de séparation de sources se fait alors en trois étapes : le choix de la fonction f et des contraintes Λ et Θ ; l'apprentissage de $\lambda \in \Lambda$; et le choix d'un critère à optimiser et d'une méthode d'estimation du paramètre $\theta \in \Theta$ en fonction du signal de test x.

L'objet de cet article est de diagnostiquer l'effet de ces étapes dans la performance globale d'un système, au moyen d'un estimateur dit *oracle*. Celui-ci est défini comme un estimateur optimal lorsque le critère d'optimisation choisi est le critère de performance lui-même – le RSD dans notre cas –, en supposant en particulier les signaux sources connus. Alors que de précédentes études ont établi les bornes de performance oracles liées à différentes fonctions de masquage [7] et au filtrage dans le domaine spectral pour les mélanges convolutifs [1], nous nous intéressons ici à l'estimation oracle pour évaluer les performances optimales liées aux contraintes de modèles à états discrets pour les mélanges monocapteurs. Nous considérons en particulier des fonctions de séparation reposant sur le masquage temps-fréquence et sur des modèles de type MMG, ainsi que des problématiques d'apprentissage génératif ou discriminant, non-adaptatif (*a priori*) ou adaptatif (en ligne).

Dans la suite du document, nous considérons un mélange $x_\tau = \sum_{j=1}^{J} s_{j\tau}$ de J sources $s_{j\tau}$ échantillonnées à la fréquence f_s, aux instants τ sur une durée T. Les transformées de Fourier à court terme (TFCT) du mélange et des sources sont respectivement notées X_{tf} et S_{jtf}, $f \in [\![0; F-1]\!]$ désignant la fréquence et t la trame. Pour une fonction de séparation donnée, on notera $\widehat{S}_{jtf}$ l'estimation de S_{jtf}, qui est obtenue dans le cas oracle en minimisant le RSD, ou de façon quasi-équivalente [7] l'erreur quadratique

$$e \triangleq \sum_{jtf} \left| S_{jtf} - \widehat{S}_{jtf} \right|^2 \qquad (1)$$

Pour plus de lisibilité, les bornes de variation des sommes seront désormais omises.

2 Estimation oracle

De nombreuses fonctions de séparation reposent sur le masquage temps-fréquence, qui consiste à définir la TFCT de la j^{e} source estimée sous la forme $\widehat{S}_{jtf} = \alpha_{jtf} X_{tf}$, où α_{jtf} est le masque temps-fréquence à déterminer. Les α_{jtf} étant les paramètres décodés, les performances oracles du masquage temps-fréquence ont été établies [7] avec différentes contraintes sur les masques, en particulier : le cas très général $\alpha_{jtf} \in \mathbb{R}$; le cas $\alpha_{jtf} \geq 0$ avec $\sum_j \alpha_{jtf} = 1$ que l'on retrouve dans les systèmes à base de filtrage de Wiener ; et le cas du masquage binaire $\alpha_{jtf} \in \{0; 1\}$, considéré lorsque les sources ne se recouvrent pas dans le plan temps-fréquence. Les bornes de performances obtenues sont alors bien supérieures aux performances des systèmes de séparation, qui introduisent des contraintes supplémentaires que nous allons décrire dans cette partie.

2.1 Approches factorielles par MMG

Dans les approches par MMG, le modèle de chaque source j est composé de Q états. L'état $q \in [\![1; Q]\!]$ de la source j est caractérisé par une probabilité *a priori* π_{jq} et une densité spectrale de puissance (DSP) $\sigma^2_{qj}(f)$ telle que la vraisemblance associée à l'état q est

$$S_{jtf}|q \sim \mathcal{N}\left(0, \sigma^2_{jqf}\right) \tag{2}$$

2.1.1 Apprentissage des modèles

L'apprentissage des paramètres des sources consiste donc à appliquer l'algorithme suivant sur une base de sources isolées :

Algorithme 1 Apprentissage du MMG de la source j

ENTRÉES: ensemble d'apprentissage $\{S_j\}_j$
Boucler
mise à jour des probabilités *a posteriori* :
$\gamma_{jqt} \propto p\left(\{S_{jtf}\}_f | q\right) \pi_{jq}$ (via éq.(2))
mise à jour des probabilités *a priori* : $\pi_{jq} \propto \sum_t \gamma_{jqt}$
mise à jour des DSP : $\sigma^2_{jqf} \leftarrow \frac{\sum_t \gamma_{qt} |S_{jtf}|^2}{\sum_t \gamma_{qt}}$
Fin boucle
RÉSULTAT: $\{\pi_{jq}\}$ and $\left\{\sigma^2_{jqf}\right\}$

2.1.2 Phase de séparation

Dans la phase de séparation, le mélange observé à l'instant t dépend d'un état parmi $K \triangleq Q^J$ états sous-jacents, dits factoriels car s'écrivant $k = (k_1, \ldots, k_J) \in [\![1; Q]\!]^J$ où k_j désigne l'état de la source j. La vraisemblance du mélange associé à un tel état conjoint est alors

$$X_{tf}|k \sim \mathcal{N}\left(0, \sum_j \sigma^2_{jk_jf}\right) \tag{3}$$

On déduit l'estimation $\widehat{S_{jtf}}$ de la TFCT de la source j par masquage temps-fréquence en définissant le masque

$$\alpha_{jtf} \triangleq \sum_k \gamma_{kt} w_{jkf} \tag{4}$$

$$\text{avec } w_{jkf} \triangleq \frac{\sigma^2_{jk_jf}}{\sum_{j'} \sigma^2_{j'k_{j'}f}} \tag{5}$$

$\gamma_{kt} \triangleq p\left(k | \{X_{tf}\}_f\right)$ étant la probabilité *a posteriori* de l'état k, et w_{jkf} correspondant à un filtre de Wiener. L'algorithme de séparation est alors

Algorithme 2 Séparation de sources par MMG

ENTRÉES: mélange X, paramètres appris $\left\{\sigma^2_{jqf}\right\}$ et $\{\pi_{jq}\}$.
Calcul des probabilités *a posteriori* :
$\gamma_{kt} \propto p\left(\{X_{tf}\}_f \mid k\right) \prod_j \pi_{jk_j}$ via eq.(3)
Calcul des masques via éq. (4).
RÉSULTAT: Estimation des sources $\left\{\widehat{S_j}\right\}$

Ces deux algorithmes illustrent les principes de base de l'approche par MMG, tels que proposée par [2, 6], où les DSP σ^2_{jqf} sont des paramètres appris et les probabilités *a posteriori* γ_{kt} des états conjoints sont les paramètres décodés. Les systèmes dits adaptatifs [5] sont des variantes qui considèrent les DSP comme des paramètres à décoder, et effectuent une adaptation « en ligne ». Par ailleurs, l'extension aux MMC s'obtient avec une hypothèse supplémentaire de dépendance temporelle sur les probabilités *a priori* des états des sources.

2.1.3 Algorithmes d'estimation oracle

Le nombre d'états Q et l'hypothèse de gaussianité de l'approche par MMG peuvent être vus comme des contraintes susceptibles de limiter les résultats de séparation. Pour évaluer ces limites, nous proposons de dresser les bornes de performance de ce type d'approche dans les cas non-adaptatif et adaptatif.

Le cas non-adaptatif (NA) consiste à effectuer la phase d'apprentissage, puis à réaliser l'estimation oracle γ^{NA}_{kt} des probabilités *a posteriori* sur les signaux de test. Cette estimation résulte de la sélection des γ_{kt} qui minimisent l'erreur (1). Ceux-ci sont obtenus par la mise à jour multiplicative (détaillée dans la partie 2.2)

$$\gamma_{kt} \leftarrow \gamma_{kt} \frac{\sum_{jf} \text{Re}\left(S_{jtf} X^*_{tf}\right) w_{jkf}}{\sum_{jf} |X_{tf}|^2 w_{jkf} \sum_{k'} \gamma_{k't} w_{jk'f}} \tag{6}$$

Dans le cas adaptatif (A), les signaux de test sont utilisés pour estimer à la fois les DSP oracles $\sigma^2_{\text{A}jqf}$ et les probabilités *a posteriori* oracles γ^{A}_{kt}. L'estimation oracle consiste à apprendre les $\sigma^2_{\text{A}jqf}$ sur les signaux de test via l'algorithme 1, à en déduire les filtres $w^{\text{A}}_{jk'f}$ (éq. (5)), puis à estimer les γ^{A}_{kt} (éq. (6)).

Les deux approches précédentes reposent sur un apprentissage génératif des paramètres des modèles : le critère optimisé vise à modéliser au mieux les sources. Alternativement,

dans la perspective de futurs systèmes de séparation, une approche adaptative discriminante (AD) est également proposée, en optimisant un critère de séparation des sources telle que l'erreur (1). Cette approche est mise en oeuvre de manière approchée en réalisant les étapes suivantes de manière itérative :

– minimisation de l'erreur (1) par rapport aux DSP σ^2_{jqf} (en utilisant la fonction de minimisation sous contraintes `fmincon` de Matlab) ;
– minimisation de l'erreur (1) par rapport aux γ_{kt} via l'éq. (6).

2.2 Approches généralisées avec états conjoints

2.2.1 Principe

En considérant Q états par source, l'approche factorielle modélise leurs formes spectrales avec environ $J \times Q \times F$ paramètres. La phase de décodage implique de considérer les $K = Q^J$ états conjoints, avec en particulier $J \times K$ formes spectrales conjointes w_{jkf}, d'où une complexité calculatoire importante lorsque que les nombres de sources et d'états augmentent. Nous proposons ici une solution qui conserve la notion d'états conjoints, tout en s'affranchissant de la modélisation des sources par MMG.

Nous considérons un masquage temps-fréquence ayant une structure plus générale que précédemment : les masques sont définis selon l'expression (4)), où $k \in [\![1; K]\!]$ est l'indice des K états conjoints, où γ_{kt} est le coefficient d'activation de l'état k à l'instant t tel que $\sum_k \gamma_{kt} = 1$ et où le filtre de séparation $w_{jkf} \in [0; 1]$ associé à l'état k et à la source j n'est plus contraint que par $\sum_j w_{jkf} = 1$. Le masquage défini par (4) est donc un cas particulier où w_{jkf} s'exprime en fonction des DSP des sources. Dans le cas général, les états conjoints k ne sont plus considérés comme des états factoriels et les coefficients w_{jkf} sont des paramètres libres, naturellement discriminants, qui ne sont pas associés à une notion de variance des sources.

2.2.2 Algorithmes d'estimation oracle

Les paramètres optimaux $\gamma^{\text{CONJ}}_{kt}$ et w^{CONJ}_{jkf} s'obtiennent en minimisant l'erreur (1) :

$$\left\{\gamma^{\text{CONJ}}_{kt}, w^{\text{CONJ}}_{jkf}\right\} = \underset{\{\gamma_{kt}, w_{jkf}\}}{\arg\min} \sum_{jtkf} \left| S_{jtf} - \sum_k \gamma_{kt} w_{jkf} X_{tf} \right|^2 \tag{7}$$

qui peut se réécrire comme un problème de factorisation en matrices à termes positifs (NMF) pondérée [1] :

$$\underset{\{\gamma_{kt}, w_{jkf}\}}{\arg\min} \sum_{jtf} \left(\frac{\text{Re}\left(S_{jtf} X^*_{tf}\right)}{|X_{tf}|} - [\Gamma W_j]_{tf} |X_{tf}| \right)^2 \tag{8}$$

avec $[W_j]_{kf} \triangleq w_{jkf}$ et $[\Gamma]_{tk} \triangleq \gamma_{kt}$. En utilisant le gradient de (8), les paramètres sont estimés itérativement par les mises à jour multiplicatives

$$\gamma_{kt} \leftarrow \gamma_{kt} \frac{\sum_{jf} \text{Re}\left(S_{jtf} X^*_{tf}\right) w_{jkf}}{\sum_{jf} |X_{tf}|^2 w_{jkf} \sum_{k'} \gamma_{k't} w_{jk'f}} \tag{9}$$

$$w_{jkf} \leftarrow w_{jkf} \frac{\sum_{t} \text{Re}\left(S_{jtf} X^*_{tf}\right) \gamma_{kt}}{\sum_{t} |X_{tf}|^2 \gamma_{kt} \sum_{k'} w_{jk'f} \gamma_{k't}} \tag{10}$$

Par ailleurs, les numérateurs pouvant prendre des valeurs négatives dans (9) et (10), la décomposition NMF est assurée en ajoutant un seuillage positif minimal sur ces numérateurs. Les bornes de performances oracles sont ainsi obtenues après un certain nombre d'itérations des mises à jour ci-dessus. L'initialisation des paramètres γ_{kt} et w_{jkf} peut se faire de façon aléatoire ou en les déduisant d'un apprentissage de MMG sur les sources isolées par l'algorithme 1.

3 Étude des performances

Une première expérimentation a été menée sur des mélanges musicaux dans lesquels le chant et l'accompagnement doivent être séparés ($J = 2$ sources). Huit séquences d'environ 11 s extraits de six morceaux différents provenant de [5] ont été analysés [2], en utilisant des trames de 185 ms avec $f_s = 11025$ Hz et se recouvrant de moitié. La séparation est réalisée par les estimateurs oracles adaptatifs – factoriel génératif (A), factoriel discriminant (AD), états conjoints généralisés (CONJ) avec $K = Q^J$ –, par le système (factoriel génératif adaptatif) existant [5] et par le masquage temps-fréquence positif oracle [7]. Les résultats sont représentés sur la figure 1 en fonction du nombre d'états par source [3].

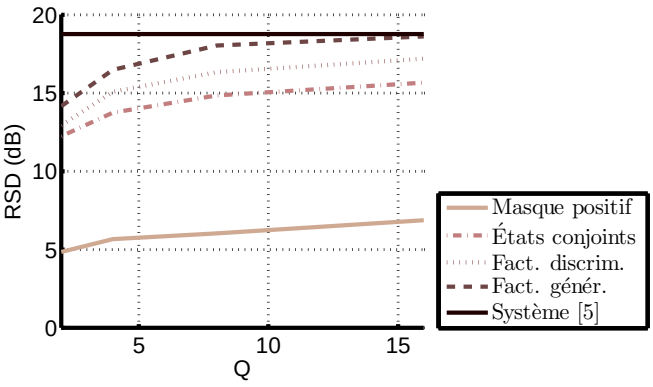

FIGURE 1 – Performances (RSD moyen), en fonction du nombre d'états Q, pour la séparation chant / accompagnement.

1. Une approche similaire a été proposée dans [9], avec quelques différences dans le contexte considéré, le critère à optimiser, les modèles utilisés et les valeurs des poids.

2. Les auteurs remercient A. Ozerov pour leur avoir fourni ses données ainsi que les résultats obtenus avec sa méthode.

3. Des exemples sonores sont disponibles à l'url `http://www.irisa.fr/metiss/vemiya/GRETSI09/`.

Dans une seconde expérimentation, des mélanges de deux ou trois sources de parole ont été analysés, avec 30 mélanges de 10 s environ dans chaque cas. L'analyse est réalisée en utilisant des trames de 128 ms avec $f_s = 16$ kHz et les résultats sont représentés sur la figure 2. Une approche factorielle générative avec apprentissage non-adaptatif (NA) a été ajoutée par rapport à l'expérience précédente, en effectuant l'apprentissage de chaque modèle sur des signaux du même locuteur.

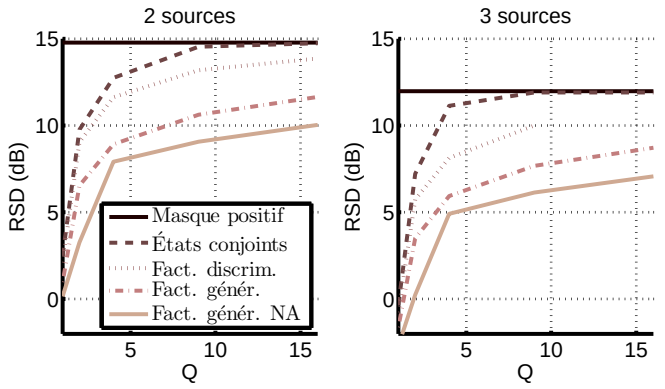

FIGURE 2 – Bornes de performances oracles en fonction du nombre d'états Q, pour la séparation de signaux de parole.

Les tendances observées sur les performances oracles peuvent se résumer ainsi :

- passage d'un apprentissage non-adaptatif à adaptatif des MMG (parole) : env. $+1$ à 2 dB ;
- passage d'un apprentissage génératif à discriminant des MMG : env. $+1$ dB ;
- passage d'une structure de MMG à des états conjoints généralisés : env. $+1$ à 2 dB ;
- passage de 2 à 3 sources de parole : env. -3 dB.

On observe également que les performances du système de l'état de l'art se situent bien en-dessous de l'ensemble des performances oracles. L'écart provient de l'estimation des paramètres par ce système, qui doit en particulier estimer les DSP des deux sources dans des séquences où l'une des sources – la voix – n'apparaît seule à aucun moment. Par ailleurs, les performances oracles avec des contraintes d'états discrets sont proches des performances oracles du masquage temps-fréquence positif non-contraint. On peut ainsi quantifier à environ 1 à 5 dB la dégradation due aux contraintes structurelles que l'on impose en choisissant un modèle de type MMG. On notera par ailleurs que dans le cas $Q = 16$, le nombre d'états conjoints $K = Q^J$ devient comparable au nombre de trames analysées : on atteint ici les limites de validité des expérimentations sur des sons d'environ 10 s.

Enfin, d'un point de vue algorithmique, il est à noter que les estimateurs oracles proposés sont relativement lourds en temps d'exécution (quelques heures pour les plus coûteux sur des PC actuels) et que la convergence des algorithmes itératifs a été observée avec un nombre maximal d'itérations fixé à 500.

4 Conclusion

L'élaboration d'estimateurs oracles a permis d'établir les performances oracles de systèmes de séparation de sources monocapteurs à états discrets, en fonction des contraintes introduites par les structures des modèles et par le type d'apprentissage. Ces résultats ont notamment été mis en évidence dans le cas d'un apprentissage non-adaptatif ou adaptatif, génératif ou discriminant. Il apparait ainsi qu'un gain de performance peut être envisagé en adoptant une approche à états conjoints généralisés. La réalisation d'un tel système de séparation de sources constitue alors la suite logique des travaux présentés ici.

Références

[1] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari. The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech. *IEEE Trans. Speech Audio Process.*, 11(2) :109–116, March 2003.

[2] L. Benaroya and F. Bimbot. Wiener based source separation with HMM/GMM using a single sensor. In *Proc. of ICA*, pages 957–961, Nara, Japan, April 2003.

[3] R. Blouet, G. Rapaport, and C. Fevotte. Evaluation of several strategies for single sensor speech/music separation. In *Proc. of ICASSP*, pages 37–40, 2008.

[4] S. Makino, T.W. Lee, and H. Sawada. *Blind speech separation*. Springer, 2007.

[5] A. Ozerov. *Adaptation de modèles statistiques pour la séparation de sources mono-capteur. Application à la séparation voix / musique dans les chansons*. PhD thesis, Univ. de Rennes 1, France, 2006.

[6] S.T. Roweis. One Microphone Source Separation. *NIPS*, 13 :793–799, 2001.

[7] E. Vincent, R. Gribonval, and M.D. Plumbley. Oracle estimators for the benchmarking of source separation algorithms. *Signal Processing*, 87(8) :1933–1950, 2007.

[8] T. Virtanen. Unsupervised Learning Methods for Source Separation in Monaural Music. In A. Klapuri and M. Davy, editors, *Signal Processing Methods for Music Transcription*, pages 267–296. Springer-Verlag, 2006.

[9] T. Virtanen. Monaural sound source separation by perceptually weighted non-negative matrix factorization. Technical report, Tampere University of Technology, 2007.