Séparation voix / musique à partir d'enregistrements mono : quelques remarques sur le choix et l'adaptation des modèles

Alexey OZEROV[1], Rémi GRIBONVAL[2], Pierrick PHILIPPE[1], Frédéric BIMBOT[2]

[1]France Télécom R&D
4, rue du Clos Courtel, BP 91226, 35512 Cesson Sévigné cedex, France

[2]IRISA (CNRS & INRIA) - projet METISS
Campus de Beaulieu, 35042 Rennes Cedex, France

alexey.ozerov@francetelecom.com, remi.gribonval@irisa.fr,
pierrick.philippe@francetelecom.com, frederic.bimbot@irisa.fr

Résumé – Le problème de l'extraction de voix chantée dans des enregistrements musicaux monophoniques, c'est-à-dire la séparation voix / musique avec un seul capteur, est étudié. Les approches utilisées sont basées sur les modèles statistiques *a priori* de deux sources. Une étude comparative des différentes modèles et estimateurs est effectuée, ainsi qu'une étude d'impact de l'hétérogénéité entre les données d'apprentissage et les données à séparer. On montre que l'adaptation du modèle de la musique sur les parties non vocales permet d'obtenir des bonnes performances dans un cadre relativement réaliste.

Abstract – The problem of singing voice extraction from mono audio recordings, i.e., one microphone separation of voice and music, is studied. The approaches are based on *a priori* probabilistic models for two sources. A comparative study of different models and estimators is done together with a study of the impact of heterogeneity between training data and data to be separated. We show that the adaptation of music model from the non vocal parts allows to obtain the good results in realistic conditions.

1 Introduction

Nous nous intéressons à l'extraction de voix chantée dans des enregistrements musicaux, c'est-à-dire à la séparation de la voix par rapport à l'accompagnement musical. Nous considérons des enregistrements mono (par opposition à stéréo). Il s'agit donc de séparation de sources avec un seul capteur. Nous supposons que chaque enregistrement est une simple somme $x(k) = v(k) + m(k)$ du signal de voix $v(k)$ et du signal de musique $m(k)$. Etant donné le signal observé $x(k)$, notre problème consiste à estimer la contribution de la voix $\hat{v}(k)$ pour l'indexer, la reconnaître, etc.

Notre approche est basée sur des modèles statistiques *a priori* des deux sources et un apprentissage sur des données d'entraînement. Notre première contribution est une étude de l'effet d'hétérogénéité des données d'apprentissage sur la performance de séparation. Nous montrons qu'en adaptant le modèle de musique à partir des parties non vocales de la chanson, il est possible d'obtenir de bonnes performances tout en restant dans un contexte réaliste. La seconde contribution de cet article est une étude des interactions entre le critère de performance, le domaine de modélisation des algorithmes et la fonction de coût optimisée.

Le reste de cet article est organisé comme suit. Les méthodes de séparation et les mesures de performance sont décrites dans les sections 2 et 3. Les questions d'adaptation des modèles et du choix de la méthode de séparation sont abordées dans la section 4. La section 5 contient la description des données expérimentales et la définition de l'estimateur oracle. Les résultats des expérimentations sont résumés dans la section 6.

2 Méthodes de séparation

Dans cette section, trois méthodes de séparation de sources avec un seul capteur [1, 2, 3] sont présentées. Les deux dernières méthodes [2, 3] ont été utilisées à l'origine pour le débruitage de la parole. Cette tâche peut être présentée sous la forme du problème de séparation de deux sources (la parole et le bruit). Le bruit est souvent considéré stationnaire et donc modélisé par un modèle simple. Ici, nous considérons les extensions de ces deux méthodes pour un problème plus complexe, c'est-à-dire la séparation de deux sources non stationnaires (la voix et la musique).

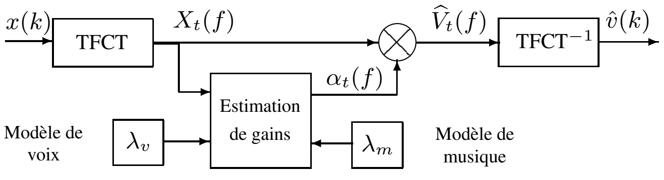

FIG. 1 – Le principe des méthodes de séparation de sources basées sur les modèles MMG

La figure 1 illustre le principe de ces trois méthodes de séparation de sources [1, 2, 3]. Soit $X_t(f)$ la Transformée de Fourier à Court Terme (TFCT) du mélange $x(k)$ pour la trame numéro t et la fréquence f. A partir de $X_t(f)$ et de paramètres de modèles de voix λ_v et de musique λ_m (estimés lors d'une phase d'apprentissage), un gain $\alpha_t(f) \geq 0$ est calculé en minimisant un critère d'erreur donné. Ensuite, l'estimation de la TFCT de la voix $\widehat{V}_t(f)$ est obtenue en multipliant $X_t(f)$ par ce gain et la voix estimée $\hat{v}(k)$ est obtenue en appliquant la TFCT inverse à l'estimation $\widehat{V}_t(f)$.

Comme critère d'erreur, on considère souvent l'Erreur Quadratique Moyenne (EQM) spectrale :

$$\|\widehat{V} - V\|^2 = \sum_{t,f} |\widehat{V}_t(f) - V_t(f)|^2, \qquad (1)$$

On peut aussi être amené à utiliser l'EQM log spectrale : $\| \log|\widehat{V}| - \log|V| \|^2$.

Benaroya [1] considère les Modèles de Mélange de Gaussiennes (MMG) spectraux $\lambda_v^{\mathrm{spec}}$ et $\lambda_m^{\mathrm{spec}}$. Les spectres à court terme de la voix V_t sont modélisés comme des variables aléatoires de densité MMG avec des vecteurs moyens nuls et des matrices de covariance $\Sigma_{vi} = \mathrm{diag}\left[\{\sigma_{vi}^2(f)\}_f\right]$, c'est-à-dire

$$p(V_t|\lambda_v^{\mathrm{spec}}) = \sum_i \omega_{vi} N(V_t; 0, \Sigma_{vi}), \qquad (2)$$

où $N(V_t; \mu, \Sigma)$ est la densité d'un vecteur gaussien V_t avec le vecteur moyen μ et la matrice de covariance Σ. La diagonale de chaque matrice de covariance $\{\sigma_{vi}^2(f)\}_f$ représente une Densité Spectrale de Puissance (DSP) locale. Ce MMG de voix est noté $\lambda_v^{\mathrm{spec}} = \{\omega_{vi}, \Sigma_{vi}\}_i$. De la même façon, la musique est modélisée par le modèle MMG $\lambda_m^{\mathrm{spec}} = \{\omega_{mj}, \Sigma_{mj}\}_j$. Le gain $\alpha_t(f)$ minimisant le critère d'EQM spectrale est calculé comme suit :

$$\alpha_t(f) = \sum_{i,j} \gamma_{i,j}(t) \frac{\sigma_{vi}^2(f)}{\sigma_{vi}^2(f) + \sigma_{mj}^2(f)}, \qquad (3)$$

où $\gamma_{i,j}(t)$ est la probabilité de choisir la paire d'états (i,j) pour l'observation X_t [1]. Ce gain satisfait $\alpha_t(f) \in [0,1]$ et l'estimation des sources revient à effectuer un *filtrage de Wiener pondéré*.

Pour les mêmes modèles, Ephraim et Malah [2] développent l'estimateur minimisant le critère d'EQM log spectrale avec le gain

$$\alpha_t(f) = \sum_{i,j} \gamma_{i,j}(t) \frac{\sigma_{vi}^2(f)}{\sigma_{vi}^2(f) + \sigma_{mj}^2(f)} \exp\left[\frac{E_1(\theta_{ij})}{2}\right], \qquad (4)$$

où $\theta_{ij} = \frac{\sigma_{vi}^2(f)|X_t(f)|^2}{[\sigma_{vi}^2(f)+\sigma_{mj}^2(f)]\sigma_{mj}^2(f)}$ et $E_1(\theta) = \int_\theta^\infty \frac{e^{-t}}{t} dt$ est connue comme l'intégrale exponentielle.

Burshtein et Gannot [3] proposent d'utiliser les modèles MMG log spectraux $\lambda_v^{\log}$ et $\lambda_m^{\log}$. Les logarithmes des spectres de la voix $\log|V_t|$ (ici l'opération $\log|\cdot|$ pour un vecteur s'applique élément par élément) sont modélisés par un MMG avec des vecteurs moyens μ_{vi} et des matrices de covariance diagonales Σ_{vi} :

$$p\left(\log|V_t| | \lambda_v^{\log}\right) = \sum_i \omega_{vi} N(\log|V_t|; \mu_{vi}, \Sigma_{vi}), \qquad (5)$$

Ce MMG est paramétrisé comme $\lambda_v^{\log} = \{\omega_{vi}, \mu_{vi}, \Sigma_{vi}\}_i$. La musique est modélisée par le MMG $\lambda_m^{\log}$. Avec cette modélisation, les DSP locales sont plutôt représentées par les vecteurs moyens μ_{vi}. Les sources sont estimées en utilisant le critère d'EQM log spectrale et en faisant l'approximation MIXMAX (Mixture Maximum) [4] (voir [3] pour les détails).

L'estimation du modèle de chaque source à partir de données d'apprentissage est basée sur le critère du Maximum de Vraisemblance (MV). En pratique, l'apprentissage utilise l'algorithme EM (Expectation-Maximisation) [5].

3 Mesures de performance

De nombreuses mesures de performance peuvent être utilisées pour évaluer la qualité de séparation. Nous en considérons ici deux : le RSDN et la DLSN.

Pour un enregistrement (une chanson) donné, le RSD Normalisé (RSDN) mesure l'amélioration du Rapport Source à Distorsion (RSD) [6]

$$\mathrm{RSD}(\hat{v}, v) = 10 \log_{10} \left[\frac{\langle \hat{v}, v\rangle^2}{\|\hat{v}\|^2\|v\|^2 - \langle \hat{v}, v\rangle^2}\right] \qquad (6)$$

entre le signal non traité x et la voix estimée $\hat{v}$: $\mathrm{RSDN} = \mathrm{RSD}(\hat{v}, v) - \mathrm{RSD}(x, v)$.

La DLS Normalisée (DLSN) est l'amélioration de la Distorsion du Log Spectre (DLS) [7]

$$\mathrm{DLS}(\hat{v}, v) = \frac{1}{T} \sum_{t=0}^{T-1} \left[\frac{1}{F} \sum_{f=0}^{F-1} \left(10 \log_{10} \frac{|V_t(f)|^2 + \epsilon}{|\widehat{\mathcal{V}}_t(f)|^2 + \epsilon}\right)^2\right]^{\frac{1}{2}} \qquad (7)$$

entre x et $\hat{v}$: $\mathrm{DLSN} = \mathrm{DLS}(x, v) - \mathrm{DLS}(\hat{v}, v)$, où $\widehat{\mathcal{V}}_t(f)$ dans l'équation (7) est la TFCT de la voix estimée $\hat{v}(k)$. Remarquons que, puisque la TFCT est une transformée redondante, *la TFCT d'estimation* $\widehat{\mathcal{V}}_t(f)$ et *l'estimation de la TFCT* $\widehat{V}_t(f)$ ne coïncident pas en général : $\widehat{\mathcal{V}}_t(f) \neq \widehat{V}_t(f)$.

Le choix de la mesure de performance dépend fortement de l'application pour laquelle la séparation est effectuée. Par exemple, si la séparation est faite pour la Reconnaissance Automatique de la Parole (RAP) effectuée sur le signal de la parole séparé, il semble plus judicieux de choisir la DLSN. En effet, les coefficients cepstraux utilisés par la plupart des systèmes du RAP étant obtenus à partir des log spectres par une transformation linéaire, il faut vraisemblablement mieux minimiser la distorsion log spectrale.

4 Adaptation des modèles et choix de la méthode de séparation

Les méthodes de séparation de sources existantes donnent des performances satisfaisantes quand les données d'apprentissage sont homogènes aux données à séparer, par exemple dans le cas où les sources d'apprentissage sont extraites de l'enregistrement musical que l'on souhaite séparer [1]. Les modèles appris dans ces conditions sont appelés *modèles adaptés* aux sources. Dans ce travail, nous considérons un cas plus réaliste où les données d'apprentissage et les données à séparer sont hétérogènes. Nous y étudions d'abord les performances de *modèles généraux* de voix et de musique, c'est-à-dire de modèles appris sur des extraits issus d'autres chansons que celles à séparer. Enfin, nous nous intéressons à un cas intermédiaire (*modèles semi-adaptés*) qui peut correspondre à un cadre d'utilisation réaliste : le modèle général de voix est combiné à un modèle de musique adapté par apprentissage sur des segments non vocaux (sans voix chantée) de l'enregistrement à séparer. Pour cela chaque enregistrement est au préalable segmenté à la main en parties vocales / non vocales.

Comme précisé section 3, le choix de la mesure de performance dépend de l'application visée. Par ailleurs, la mesure influence le choix d'une méthode particulière. Pour comprendre cette dépendance, nous testons les méthodes présentées dans la section 2 avec les mesures RSDN et DLSN. Ces méthodes sont caractérisées par le domaine de modélisation (MMG spectral / log spectral) et le critère d'erreur minimisé (EQM spectrale / log spectrale). Nous nous attendons *a priori* à ce qu'en basculant progressivement du domaine spectral dans le domaine log spectral, le RSDN se dégrade et la DLSN s'améliore.

5 Cadre expérimental

5.1 Description des données.

La base d'apprentissage du modèle général de voix contient 34 extraits de voix chantée masculine issus de chansons populaires. Chaque extrait dure approximativement une minute. Le modèle général de musique est appris sur 30 extraits de musique populaire sans voix. Chaque extrait dure également environ une minute et tous les extraits proviennent d'auteurs différents. La base d'évaluation contient cinq chansons du même genre pour lesquelles les pistes de voix et de musique sont disponibles séparément, ce qui permet d'évaluer la performance de la séparation en comparant l'estimation à l'original. Tous les enregistrements sont en mono et échantillonnés à 11025 Hz.

5.2 Estimateur oracle et limites de performance

Remarquons que l'estimation de la voix $\hat{v}$ ne dépends que du mélange x et de l'ensemble des gains $A = \{\alpha_t(f)\}_{t,f}$ (fig. 1). Il est donc possible d'exprimer cette estimation comme $\hat{v} = g(x, A)$. L'ensemble des gains A appartient à un ensemble des gains admissibles $\mathcal{A}$ dépendant de la méthode de séparation. Ici, on considère $\mathcal{A}_{[0,1]} = \{A|\alpha_t(f) \in [0,1]\}$ pour le filtrage de Wiener pondéré [1] et $\mathcal{A}_+ = \{A|\alpha_t(f) \geq 0\}$ pour les autres méthodes [2, 3]. Etant donné une mesure de performance $h(\hat{v}, v, x)$, *l'estimateur oracle* consiste à trouver l'ensemble des gains $A \in \mathcal{A}$ qui donne la meilleur performance [8] :

$$\tilde{v} = g(x, \tilde{A}), \quad \tilde{A} = \arg\max_{A \in \mathcal{A}} h(g(x, A), v, x) \qquad (8)$$

Cet estimateur nous permet donc de calculer pour un jeu de données la limite de performance qui ne peut pas être dépassée avec la méthode correspondante.

Comme, il est difficile de calculer l'oracle (8) pour les mesures de performance RSDN et DLSN, on calcule à la place les estimateurs oracles pour le Rapport Signal à Bruit (RSB) spectral défini comme

$$\text{RSB spec.} = 10 \log_{10} \left[\|V\|^2 / \|\widehat{V} - V\|^2 \right] \qquad (9)$$

Les valeurs de la RSDN et du DLSN (voir (6), (7)) calculés pour ces estimateurs oracles nous indiquent des performances que l'on pourrait atteindre en améliorant l'estimation des gains $\alpha_t(f)$.

6 Expérimentations et résultats

Les problèmes qui se posent, et que nous allons étudier dans cette section, sont les suivants :

1. choix des paramètres de la TFCT (taille et type de la fenêtre d'analyse)
2. effet du corpus d'apprentissage (hétérogénéité ou homogénéité) et dimensionnement des modèles MMG (nombre de Gaussiennes)
3. choix du domaine de modélisation et du critère d'erreur (spectre / log spectre)

6.1 Choix de la fenêtre d'analyse

En utilisant l'estimateur oracle calculé pour $A \in \mathcal{A}_{[0,1]}$, nous avons testé quelle taille et quel type de fenêtre d'analyse dans la TFCT donnent la meilleur performance (fig. 2). Le meilleur résultat est obtenu avec une fenêtre de Hamming de taille 1024 échantillons (soit 93 ms), qui a été utilisée pour le reste des expériences.

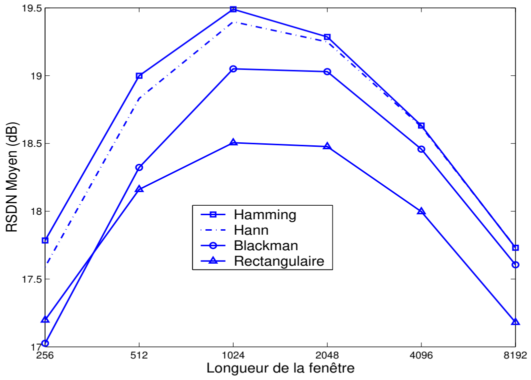

FIG. 2 – Le RSDN pour l'estimateur oracle en fonction de la taille et du type de fenêtre d'analyse.

6.2 Effet de l'hétérogénéité des données

Avec les MMG spectraux et l'estimateur minimisant l'EQM spectrale (3), nous avons testé l'effet sur le RSDN du nombre de Gaussiennes des MMG de voix n_v et de musique n_m dans les trois configurations suivantes :

1. modèles généraux de voix et de musique,
2. modèle général de voix et modèle adapté de musique (appris sur les parties non vocales), c'est-à-dire modèles semi-adaptés,
3. modèle adapté de voix (appris sur la voix séparée) et modèle adapté de musique.

Les résultats sont résumés sur la figure 3. Dans le cas du filtrage de Wiener ($n_v = n_m = 1$), l'adaptation du modèle de musique augmente le RSDN de 2 dB par rapport aux modèles généraux. L'adaptation supplémentaire du modèle de voix augmente encore la performance de 1.5 dB. Avec deux modèles généraux, l'augmentation du nombre total de Gaussiennes $n_v n_m$ n'améliore pas sensiblement la performance par rapport à $n_v = n_m = 1$, voire la fait légèrement décroître. Dès que l'on adapte le modèle de musique sur les parties non vocales, l'augmentation du nombre $n_v n_m$ de Gaussiennes permet d'améliorer la performance. Avec $n_v = n_m = 128$, le RSDN est de 2.5 dB plus grand que pour le filtrage de Wiener semi-adapté et de 4 dB plus grand que le meilleur résultat avec les deux modèles généraux. Les résultats obtenus avec deux modèles adaptés (sans doute irréalistes en pratique) nous montrent qu'il reste encore une marge de 3 à 4 dB pour améliorer l'apprentissage du modèle de voix.

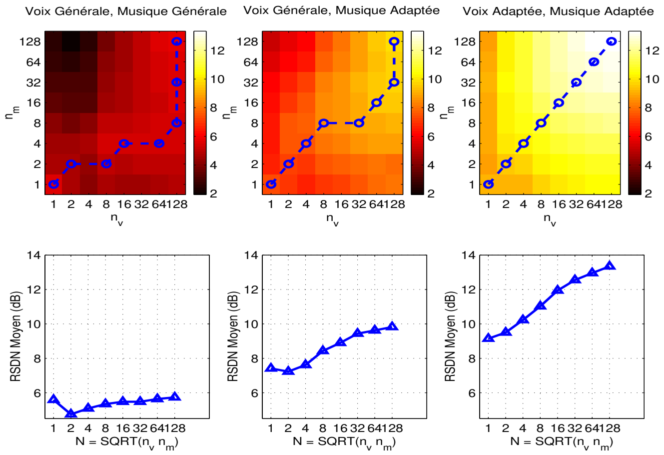

FIG. 3 – En haut : Les RSDN en fonction de (n_v, n_m). La ligne pointillée représente pour un N donné ($N = 1, 2, 4, \ldots, 128$) la paire (n_v^*, n_m^*) qui donne la meilleur performance parmi toutes les paires (n_v, n_m), telles que $n_v n_m = N^2$. En bas : Le RSDN le long de la ligne pointillée. Une référence de performance obtenue avec l'estimateur oracle est de l'ordre 19.5 dB (fig. 2).

6.3 Effets du critère optimisé et du domaine de modélisation

Enfin, nous avons comparé les performances des algorithmes en fonction du domaine de modélisation (MMG spectral / log spectral) et du critère d'erreur minimisé (EQM spectrale / log spectrale), avec des modèles semi-adaptés et $n_v = n_m = 64$. Pour chaque paire (modèle, critère d'erreur) les deux mesures de performance (le RSDN et le DLSN) sont calculées. Les résultats, accompagnés par des références de performance obtenues avec des estimateurs oracles, sont résumés dans le tableau 1. Comme on s'y attendait (sec. 4), en passant progressivement du domaine spectral dans le domaine log spectral, le RSDN se dégrade. Par contre, la DLSN ne s'améliore pas de façon monotone. En effet, la DLSN est plus mauvaise pour la deuxième méthode (spectral / log spectrale) que pour la première (spectral / spectrale). Nous avons rajouté dans le tableau 1 les valeurs de la mesure DLSN', calculée en remplaçant la TFCT d'estimation $\widehat{\mathcal{V}}_t(f)$ par l'estimation de la TFCT $\widehat{V}_t(f)$ dans l'equation (7). Cette mesure est ajoutée à titre informatif, puisque elle ne peut pas être calculée quand la séparation est terminée, car $\widehat{V}_t(f)$ n'est plus accessible (voir la fig. 1). Pour la DLSN', cette amélioration monotone est vérifiée, vraisemblablement parce que la DLSN' est plus cohérente avec le critère d'EQM log spectrale que la DLSN. Le meilleur RSDN est toujours obtenu pour la première méthode (spectral / spectrale) et la meilleure DLSN pour la troisième (log spec. / log spec.).

7 Conclusions et perspectives

Nous proposons une procédure d'adaptation du modèle de musique permettant de rester dans un cadre d'utilisation réaliste,

TAB. 1 – Les performances des méthodes (modèle MMG / le critère d'EQM minimisé). Les références de performance obtenues à l'aide des oracles sont indiquées dans les parenthèses. La DLSN' est ajoutée à titre informatif.

MMG / EQM	RSDN	DLSN	DLSN'
spectral / spectrale [1]	9.1 (19.5)	4.5 (13.0)	*1.5*
spectral / log spec [2]	8.4 (20.1)	3.5 (13.0)	*3.3*
log spec / log spec [3]	6.4 (20.1)	5.4 (13.0)	*4.9*

puisque la segmentation manuelle en parties vocales / non vocales peut être effectuée par un utilisateur. Dans le cadre de notre étude, nous avons montré que cette adaptation permet d'améliorer considérablement la performance de séparation par rapport au cas de deux modèles généraux. Dans l'avenir, nous comptons remplacer la segmentation manuelle en parties vocales / non vocales par un module de segmentation automatique [9].

Les effets du critère optimisé et du domaine de modélisation ont été étudiés en utilisant deux mesures de performance différentes. Cet étude donne des indication sur le choix d'une méthode de séparation en fonction de la mesure de performance, qui dépend en même temps de la tâche pour laquelle la séparation est effectuée.

Références

[1] L. Benaroya, "Séparation de plusieurs sources sonores avec un seul microphone," Ph.D. dissertation, Université de Rennes 1, 2003.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," in *IEEE Trans. on Acoust., Speech, and Sig. Proc.*, vol. ASSP-33, Apr 1985, pp. 443–445.

[3] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," in *EUROSPEECH*, vol. 6, Budapest, Hungary, Sep 1999, pp. 2591–2594.

[4] A. Nádas, D. Nahamoo and M. A. Picheny, "Speech recognition using noise-adaptive prototype," in *IEEE Trans. on Speech and Audio Proc.*, 1989, pp. 1495–1505.

[5] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, Feb 1989.

[6] R. Gribonval, L. Benaroya, E. Vincent and C. Févotte, "Proposals for performance measurement in source separation," in *ICA*, Apr 2003, pp. 763–768.

[7] Valin J.-M., Rouat J. and Michaud F., "Microphone array post-filter for separation of simultaneous non-stationary sources," in *ICASSP*, 2004.

[8] E. Vincent et R. Gribonval, "Construction d'estimateurs oracles pour la séparation de sources," in *GRETSI*, 2005, (à paraître).

[9] Wei-Ho Tsai, Dwight Rogers and Hsin-Min Wang, "Blind clustering of popular music recordings based on singer voice characteristics," *Computer Music Journal*, 2004.