Choix et adaptation de modèles statistiques pour la séparation de voix chantée à partir d'un seul microphone

Choice and adaptation of statistical models for single channel singing voice separation

Alexey Ozerov [1,2], Pierrick Philippe [1], Rémi Gribonval [2] et Frédéric Bimbot [2]

[1] Orange Labs, 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
loha_ozerov@mail.ru, pierrick.philippe@orange-ftgroup.com, remi.gribonval@irisa.fr, frederic.bimbot@irisa.fr

Manuscrit reçu le 12 janvier 2006

Résumé et mots clés

Le problème de l'extraction de la 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 des modèles statistiques *a priori* des deux sources (musique et voix), notamment sur des Modèles de Mélange de Gaussiennes (MMG). Une méthode d'adaptation des modèles aux caractéristiques des sources mélangées est proposée, et une étude comparative des différents modèles et estimateurs est effectuée. Les résultats montrent que l'adaptation du modèle de musique sur les parties non-vocales des chansons permet d'obtenir de bonnes performances dans un cadre réaliste.

Séparation de sources avec un seul capteur, voix chantée, modèles statistiques, modèles de mélange de gaussiennes, filtrage de Wiener adaptatif, adaptation de modèles.

Abstract and key words

The problem of singing voice extraction from mono audio recordings, *i.e.*, one microphone separation of voice and music, is studied. The approach is based on *a priori* probabilistic models for two sources, more precisely on Gaussian Mixture Models (GMM). A method for model adaptation to the characteristics of the mixed sources is developed and a comparative study of different models and estimators is performed. We show that the adaptation of the model of music from the non-vocal parts of songs yields good results in realistic conditions.

Single channel source separation, singing voice, statistical models, Gaussian mixture models, adaptive Wiener filtering, models adaptation.

1. Introduction

Cet article s'intéresse à l'extraction de la 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 [1]. Nous supposons que chaque enregistrement est composé d'une simple somme $x(\tau) = v(\tau) + m(\tau)$ du signal de voix $v(\tau)$ et du signal de musique $m(\tau)$. Etant donné le signal observé $x(\tau)$, le problème consiste à estimer la contribution de la voix $\hat{v}(\tau)$ pour pouvoir ensuite l'analyser, la reconnaître, etc.

L'approche proposée est basée sur des modèles statistiques *a priori* des deux sources, notamment sur des Modèles de Mélange de Gaussiennes (MMG), appris sur des données d'entraînement. Une première contribution de ce travail est le développement de *modèles adaptés* qui se situent entre des *modèles de référence* (appris sur des données de test et non-utilisables dans un cadre réaliste) et des *modèles généraux* (appris sur des données différentes de celles de test). L'intérêt de ces modèles est qu'ils peuvent être utilisés dans un contexte réaliste comme les modèles généraux, tout en apportant des performances de séparation qui se rapprochent des bonnes performances des modèles de référence. Nous montrons qu'en adaptant le modèle de musique à partir des parties non-vocales de la chanson, il est possible d'obtenir des performances satisfaisantes 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 et la mesure de distorsion minimisée pour la séparation.

Après avoir présenté de façon générale les méthodes de séparation basées sur des modèles statistiques *a priori* (section 2), les méthodes de séparation basées sur des MMG et des mesures de performance sont décrites (sections 3 et 4). Les techniques d'adaptation des modèles et le choix de la méthode de séparation sont abordés dans la section 5. La section 6 concerne la description des données expérimentales et la définition de l'estimateur « oracle ». Les résultats des expérimentations sont résumés dans la section 7.

2. Présentation générale des méthodes de séparation

Dans le cadre assez général des méthodes de séparation de sources avec un seul capteur, basées sur des modèles statistiques *a priori* (Fig. 1), nous introduisons brièvement les trois transformées associées aux différentes méthodes traitées dans cet article :

• La transformée $\mathcal{F}$ transforme le signal temporel dans un domaine particulier, dans lequel le traitement est ensuite effectué (souvent, $\mathcal{F}$ est une TFCT (c'est-à-dire une Transformée de Fourier à Court Terme) [2, 3, 4]).

• La transformée $\mathcal{L}$ définit (dans le domaine transformé par $\mathcal{F}$) le domaine de modélisation des sources (par exemple, $\mathcal{L} = \mathrm{Id}$ dans [2, 3] et $\mathcal{L} = \log|\cdot|$ dans [4], voir section 3).

• La transformée $\mathcal{D}$ définit (dans le domaine transformé par $\mathcal{F}$) le domaine dans lequel sera calculée l'Erreur Quadratique Moyenne (EQM) minimisée pour l'estimation des sources (par exemple, $\mathcal{D} = \mathrm{Id}$ dans [2] et $\mathcal{D} = \log|\cdot|$ dans [3, 4], voir section 3).

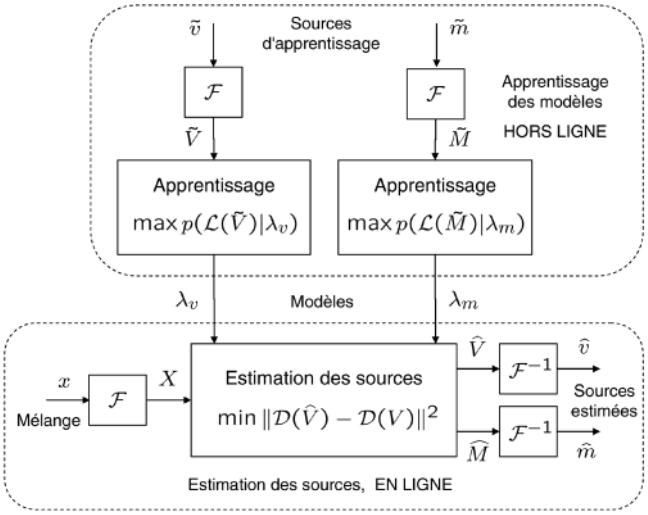


Figure 1. Schéma général de séparation.

Dans le cas général, le traitement est effectué dans un domaine autre que le domaine temporel. Nous supposons que la représentation des signaux est soumise à une transformée $\mathcal{F}$. Cette transformée peut être redondante, mais il est préférable qu'elle soit inversible pour que les signaux puissent être reconstruits dans le domaine temporel à l'aide d'une transformée de reconstruction $\mathcal{F}^{-1}$. Les signaux dans le domaine transformé sont notés par des lettres majuscules (par ex. $V = \mathcal{F}(v)$).

La procédure de séparation dans le domaine transformé consiste en deux phases : l'apprentissage des modèles des sources (hors ligne) et l'estimation des sources (en ligne).

2.1. Apprentissage

L'apprentissage des modèles *a priori* λ_v et λ_m des deux sources se fait indépendamment à partir des sources d'apprentissage $\tilde{v}$ et $\tilde{m}$. En général, le critère du Maximum de Vraisemblance (MV) est utilisé :

$$\lambda_v = \arg\max_{\lambda'_v} p(\mathcal{L}(\widetilde{V})|\lambda'_v), \qquad (1)$$

où $\mathcal{L}(\widetilde{V})$ est le processus modélisé par le modèle λ_v et $\mathcal{L}$ est une transformée. Le critère du MV pour le modèle de musique λ_m s'écrit de façon similaire.

2.2. Estimation de sources

Pour estimer les sources, on cherche à minimiser la mesure de distorsion $d(\widehat{V},V)$ entre la source estimée et la source réelle. Puisque la source réelle n'est pas observée, la valeur de la mesure de distorsion est remplacée par son espérance, qui est ensuite minimisée. Les sources sont donc estimées comme suit :

$$\widehat{V} = \arg\min_{V'} E\left[d(V',V)|X,\lambda_v,\lambda_m\right] \quad (2)$$

Par la suite, les formules ne seront données que pour l'estimation de la source de voix. L'estimation de la source de musique peut être calculée en utilisant des expressions analogues.
Supposons que $d(\widehat{V},V) = \|\mathcal{D}(\widehat{V}) - \mathcal{D}(V)\|^2$ est l'Erreur Quadratique Moyenne (EQM) de $\mathcal{D}(V)$, où $\mathcal{D}$ est une transformée inversible. En utilisant l'expression pour l'estimateur minimisant l'EQM [5], nous avons :

$$\widehat{V} = \mathcal{D}^{-1}\left(E\left[\mathcal{D}(V)|X,\lambda_v,\lambda_m\right]\right) \quad (3)$$

3. Méthodes de séparation à base de MMG

Cette section présente trois méthodes de séparation de sources avec un seul capteur basées sur des MMG [2, 3, 4]. Les deux dernières méthodes [3, 4] ont été utilisées à l'origine pour le débruitage de la parole. Cette tâche peut être présentée sous la forme d'un 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 à un problème plus complexe, c'est-à-dire la séparation de deux sources non-stationnaires (la voix et la musique).

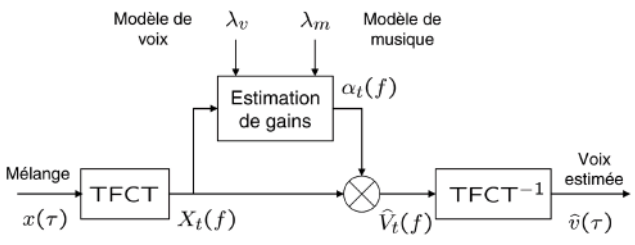


Figure 2. Estimation des sources avec des méthodes basées sur les MMG.

Le bloc « estimation des sources » (voir Fig. 1) pour ces trois méthodes [2, 3, 4] est décrit par la figure 2. Une Transformée de Fourier à Court Terme (TFCT) avec un recouvrement des fenêtres d'analyse à 50 % est utilisée pour représenter les signaux. Ceux-ci sont reconstruits dans le domaine temporel en utilisant la méthode OLA (OverLap – Add, voir par exemple [6]).
Soit $X_t(f)$, la TFCT du mélange $x(\tau)$ pour la trame numéro t et la fréquence f. À partir de $X_t(f)$ et des paramètres des modèles de voix λ_v et de musique λ_m, un gain réel $\alpha_t(f) \geq 0$ est calculé, en minimisant une mesure de distorsion donnée (2). Ensuite, l'estimation de la TFCT de la voix $\widehat{V}_t(f)$ est obtenue en multipliant $X_t(f)$ par ce gain. Autrement dit, dans le domaine de la TFCT, l'estimation de la voix $\widehat{V}_t(f)$ est obtenue en filtrant le mélange $X_t(f)$ avec un filtre dont la fonction de transfert $\alpha_t(f)$ est réestimée pour chaque trame t à partir de X_t, λ_v et λ_m. La voix estimée $\hat{v}(\tau)$ est enfin reconstruite dans le domaine temporel.
La transformée $\mathcal{F}$ étant définie, il reste à spécifier pour chaque méthode les transformées $\mathcal{L}$ et $\mathcal{D}$ (Fig. 1) définissant respectivement le *domaine de modélisation* et le *domaine de minimisation de l'EQM*. Nous allons considérer les possibilités suivantes :
1. Domaine de modélisation : $\mathcal{L} = \text{Id}$ ou $\mathcal{L} = \log|\cdot|$. La méthode correspondante est appelée *MMG spectral* ou *MMG log spectral*.
2. Domaine de minimisation de l'EQM : $\mathcal{D} = \text{Id}$ ou $\mathcal{D} = \log|\cdot|$. La méthode correspondante est appelée *EQM spectrale* ou *EQM log spectrale*.
Chaque méthode est ainsi définie par un duo MMG / EQM et les trois méthodes que nous allons décrire sont :
1. **MMG spectral / EQM spectrale** (Sec. 3.1) [2],
2. **MMG spectral / EQM log spectrale** (Sec. 3.1) [3],
3. **MMG log spectral / EQM log spectrale** (Sec. 3.2) [4].
Expliquons d'abord de manière informelle les principales différences entre les modèles et les estimateurs étudiés.
Chaque MMG (spectral ou log spectral) représente la source modélisée par un ensemble de formes spectrales caractéristiques. La différence principale entre les MMG spectraux et les MMG log spectraux est que ces formes sont construites à partir des spectres dans le premier cas et des log spectres dans le deuxième. Ainsi, avec un MMG spectral, davantage d'effort est porté sur la modélisation des parties énergétiques des spectres. Tandis qu'avec un MMG log spectral, l'effort de la modélisation est distribué plus uniformément sur tous les niveaux énergétiques (en échelle logarithmique). Un MMG spectral et un MMG log spectral appris sur les mêmes données d'entraînement sont représentés figure 10 (A) et (B). On voit qu'il y a plus de formes caractéristiques représentant les sons harmoniques pour le MMG spectral que pour le MMG log spectral. Vraisemblablement, ceci provient de la différence dans l'effort de modélisation par rapport aux différents niveaux énergétiques. La différence principale entre les estimateurs minimisant l'EQM spectrale et l'EQM log spectrale est de même nature. L'estimateur minimisant l'EQM spectrale privilégie les parties énergétiques des spectres.

3.1. Modélisation des spectres par des MMG

Benaroya [2] propose de modéliser la distribution statistique des spectres à court terme par des MMG. Ces modèles seront dénommés *MMG spectraux* et notés λ_v^{spec} et λ_m^{spec}. Les spectres à court terme de la voix V_t et de la musique M_t sont modélisés comme des vecteurs aléatoires complexes circulaires de densité MMG, avec des vecteurs moyens nuls et des matrices de covariance diagonales $\Sigma_{vi} = \text{diag}\left[\{\sigma_{vi}^2(f)\}_f\right]$ et $\Sigma_{mj} = \text{diag}\left[\{\sigma_{mj}^2(f)\}_f\right]$, c'est-à-dire :

$$p(V_t|\lambda_v^{\text{spec}}) = \sum_i \omega_{vi} N_C(V_t; \bar{0}, \Sigma_{vi}), \tag{4}$$

$$p(M_t|\lambda_m^{\text{spec}}) = \sum_j \omega_{mj} N_C(M_t; \bar{0}, \Sigma_{mj}), \tag{5}$$

où

$$N_C(V_t; \mu, \Sigma) = \prod_f \frac{1}{\pi\sigma^2(f)} \exp\left[-\frac{|V_t(f) - \mu(f)|^2}{\sigma^2(f)}\right] \tag{6}$$

est la densité d'un vecteur gaussien complexe circulaire V_t de vecteur moyen $\mu = \{\mu(f)\}_f$ et de matrice de covariance diagonale $\Sigma = \text{diag}\left[\{\sigma^2(f)\}_f\right]$ (voir 3.1.1 ci-dessous).

Les MMG spectraux sont paramétrisés comme suit : $\lambda_v^{\text{spec}} = \{\omega_{vi}, \Sigma_{vi}\}_i$ et $\lambda_m^{\text{spec}} = \{\omega_{mj}, \Sigma_{mj}\}_j$.

L'apprentissage des modèles est basé sur le critère du MV (1) avec $\mathcal{L} = \text{Id}$. En pratique, l'apprentissage utilise l'algorithme EM (Expectation-Maximisation) [7]. Les formules de réestimation des paramètres peuvent être trouvées dans [8]. L'algorithme des K-moyennes [9] est utilisé pour l'initialisation de EM.

La diagonale de chaque matrice de covariance $\{\sigma_{vi}^2(f)\}_f$ représente une Densité Spectrale de Puissance (DSP) locale. Ainsi chaque modèle explique la source modélisée par un nombre fini de formes spectrales caractéristiques (voir par exemple la figure 10 (A), (C) et (D)).

Dans l'article [10] une extension de cette modélisation est proposée. Un facteur multiplicatif appelé *facteur de gain* est associé à chaque DSP (chaque état) du MMG. Ces facteurs modélisent l'énergie locale du signal et ils sont réestimés *a posteriori* pour chaque trame pendant la séparation. Cela rend la modélisation invariante par rapport à l'énergie locale des sources. Cette extension est assez intéressante et prometteuse, mais nous ne la traitons pas dans cet article.

3.1.1. Rappel sur la densité d'un vecteur gaussien complexe circulaire

Les moments centrés d'ordre 2 d'un vecteur aléatoire complexe V_t sont définis par une matrice de covariance $\Sigma = E[(V_t - \mu)(V_t - \mu)^H]$ et une matrice de *pseudo-covariance* $\widetilde{\Sigma} = E[(V_t - \mu)(V_t - \mu)^T]$, où $\mu = E[V_t]$ est le vecteur moyen et où H et T en exposant d'un vecteur complexe représentent respectivement sa transposée-conjuguée et sa transposée. Ce vecteur aléatoire complexe est appelé *circulaire* (*proper* en anglais) si la matrice de pseudo-covariance est nulle [11, 12]. Si, en plus de la circularité, ce vecteur complexe est gaussien et la matrice de covariance Σ est diagonale, on peut montrer que sa densité s'exprime selon la formule (6) [11,12].

La différence entre la formule (6) et celle de la densité d'un vecteur gaussien réel (16), (17) peut être expliquée en imaginant que chaque vecteur $[\text{Re}V_t(f), \text{Im}V_t(f)]^T \in \mathbb{R}^2$ est un vecteur aléatoire gaussien réel bidimensionnel ayant comme vecteur moyen $[\text{Re}\mu(f), \text{Im}\mu(f)]^T$ et comme matrice de covariance diagonale $\text{diag}[0.5\sigma^2(f), 0.5\sigma^2(f)]$ (Fig. 3).

3.1.2. Estimateur minimisant l'EQM spectrale

Considérons la mesure de distorsion suivante :

$$d_{\text{spec}}(\widehat{V}, V) = \|\widehat{V} - V\|^2 = \sum_{t,f} |\widehat{V}_t(f) - V_t(f)|^2, \tag{7}$$

Si l'on minimise cette mesure de distorsion, appelée par la suite *EQM spectrale*, en utilisant l'expression (3), on arrive à la formule suivante pour le gain $\alpha_t(f)$ [2] :

$$\alpha_t^{\text{wien\_ada}}(f) = \sum_{i,j} \gamma_{i,j}(t) \frac{\sigma_{vi}^2(f)}{\sigma_{vi}^2(f) + \sigma_{mj}^2(f)}, \tag{8}$$

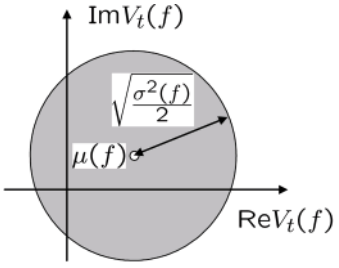
où $\gamma_{i,j}(t)$ est la probabilité de la paire d'états (i,j) pour l'observation X_t, satisfaisant $\sum_{i,j} \gamma_{i,j}(t) = 1$ et :

$$\gamma_{i,j}(t) \propto \omega_{vi}\omega_{mj} N_C(X_t; \bar{0}, \Sigma_{vi} + \Sigma_{mj}) \tag{9}$$

Ce gain satisfait $\alpha_t^{\text{wien\_ada}}(f) \in [0,1]$ et l'estimation des sources revient à effectuer un *filtrage de Wiener pondéré*.

3.1.3. Estimateur minimisant l'EQM log spectrale

Considérons maintenant une autre mesure de distorsion, que l'on appellera *EQM log spectrale* :



$$V_t(f) \sim \mathcal{N}(\mu(f), \sigma^2(f)) \qquad V_t(f), \mu(f) \in \mathbb{C}, \ \sigma^2(f) \in \mathbb{R}$$

$$\begin{bmatrix} \text{Re}V_t(f) \\ \text{Im}V_t(f) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \text{Re}\mu(f) \\ \text{Im}\mu(f) \end{bmatrix}, \begin{bmatrix} \frac{\sigma^2(f)}{2} & 0 \\ 0 & \frac{\sigma^2(f)}{2} \end{bmatrix}\right)$$

Figure 3. Distribution d'une variable aléatoire gaussienne complexe circulaire $V_t(f) \in \mathbb{C}$.

$$d_{\log}(\widehat{V},V)=\sum_{t,f}\left|\log|\widehat{V}_t(f)|-\log|V_t(f)|\right|^2 \quad (10)$$

En utilisant de nouveau l'équation (3), avec $\mathcal{D}(V)=\log|V|$ [1], il est possible d'obtenir le gain pour l'EQM log spectrale [3]:

$$\log\alpha_t^{\text{spec\_log}}(f)=\sum_{i,j}\gamma_{i,j}(t)\left[\log\frac{\sigma_{vi}^2(f)}{\sigma_{vi}^2(f)+\sigma_{mj}^2(f)}+\frac{E_1(\theta_{ij})}{2}\right], \quad (11)$$

où $\theta_{ij}=\dfrac{\sigma_{vi}^2(f)|X_t(f)|^2}{[\sigma_{vi}^2(f)+\sigma_{mj}^2(f)]\sigma_{mj}^2(f)}$ et la fonction $E_1(\theta)=\displaystyle\int_\theta^\infty\frac{e^{-t}}{t}\,dt$, connue sous le nom de *l'intégrale exponentielle*, peut être calculée numériquement de façon efficace (voir par ex. [13], pages 222-226). Les probabilités $\gamma_{i,j}(t)$ sont calculées en utilisant l'équation (9).

3.1.4. Une remarque sur la phase

Dans l'expression de l'EQM log spectrale (10), la phase de la TFCT $\angle V_t(f)=\arg V_t(f)$ n'est pas prise en compte, c'est-à-dire que, contrairement aux exigences de la section 2.2, la transformée $\mathcal{D}(V)=\log|V|$ n'est pas inversible.
Formellement, il faut ajouter l'EQM de la phase

$$d_{\text{phase}}(\widehat{V},V)=\sum_{t,f}\left|e^{j\angle\widehat{V}_t(f)}-e^{j\angle V_t(f)}\right|^2 \quad (12)$$

à l'EQM log spectrale (10). En utilisant la définition du MMG spectral (4), (5) et l'expression d'un vecteur gaussien complexe circulaire [11, 12] (Fig. 3), on peut en déduire que la phase est distribuée uniformément entre 0 et 2π:

$$\angle V_t(f)\sim\mathcal{U}(0,2\pi) \quad (13)$$

Cette propriété signifie qu'il n'y a aucune connaissance *a priori* sur la phase. Dans ce cas, Ephraim [14] montre que l'estimation minimisant l'EQM de la phase (12) est la phase du mélange $\angle X_t(f)$. Cela reste vrai pour toutes les méthodes considérées ici et transparaît sur la figure 2 par le fait que la TFCT complexe du mélange $X_t(f)$ est multipliée par un gain réel $\alpha_t(f)$.
Cet estimateur de la phase peut se comprendre intuitivement: si l'on n'a pas de connaissances *a priori* sur la phase, mieux vaut ne rien modifier et conserver la phase du mélange.

3.2. Modélisation des log spectres par des MMG

À l'instar de Burshtein et Gannot [4], il est possible de modéliser les logarithmes des spectres par des MMG. Ces modèles, appelés *MMG log spectraux*, sont notés $\lambda_v^{\log}$ et $\lambda_m^{\log}$. Les logarithmes des spectres de la voix $\mathbf{V}_t=\log|V_t|$ et de la musique $\mathbf{M}_t=\log|M_t|$ sont modélisés par des MMG avec des vecteurs moyens μ_{vi} et μ_{mj} et des matrices de covariance diagonales Σ_{vi} et Σ_{mj}:

$$p\left(\mathbf{V}_t|\,\lambda_v^{\log}\right)=\sum_i\omega_{vi}N(\mathbf{V}_t;\mu_{vi},\Sigma_{vi}), \quad (14)$$

$$p\left(\mathbf{M}_t|\,\lambda_m^{\log}\right)=\sum_j\omega_{mj}N(\mathbf{M}_t;\mu_{mj},\Sigma_{mj}), \quad (15)$$

où

$$N(\mathbf{V}_t;\mu,\Sigma)=\prod_f N\left(\mathbf{V}_t(f);\mu(f),\sigma^2(f)\right) \quad (16)$$

est la densité d'un vecteur gaussien réel $\mathbf{V}_t$ avec comme vecteur moyen $\mu=\{\mu(f)\}_f$ et comme matrice de covariance diagonale $\Sigma=\text{diag}\left[\{\sigma^2(f)\}_f\right]$. La densité d'une variable aléatoire réelle gaussienne $\mathbf{V}_t(f)$ de moyenne $\mu(f)$ et de variance $\sigma^2(f)$ est définie comme suit:

$$N\left(\mathbf{V}_t(f);\mu(f),\sigma^2(f)\right)=\frac{1}{\sqrt{2\pi\sigma^2(f)}}\exp\left[-\frac{1}{2}\frac{(\mathbf{V}_t(f)-\mu(f))^2}{\sigma^2(f)}\right] \quad (17)$$

De plus, il est supposé que, comme pour les MMG spectraux (4) et (5), la phase de la TFCT est distribuée uniformément entre 0 et 2π. Ces MMG sont paramétrisés comme suit: $\lambda_v^{\log}=\{\omega_{vi},\mu_{vi},\Sigma_{vi}\}_i$ et $\lambda_m^{\log}=\{\omega_{mj},\mu_{mj},\Sigma_{mj}\}_j$.
Avec cette modélisation, les DSP locales sont représentées par les vecteurs moyens μ_{vi} plutôt que par les diagonales des matrices de covariance $\{\sigma_{vi}^2(f)\}_f$, comme dans le cas des MMG spectraux. En effet, selon l'équation (14) les moyennes $\{\mu_{vi}(f)\}_f$ déterminent la forme spectrale typique et les variances $\{\sigma_{vi}^2(f)\}_f$ déterminent la variation de cette forme. Pour donner un exemple, un MMG log spectral à 16 états est représenté figure 10 (B). Ce modèle est appris sur les même données que celui représenté figure 10 (A).
Comme pour les MMG spectraux (Sec. 3.1), l'apprentissage est basé sur le critère du MV (1) avec $\mathcal{L}=\log|\cdot|$ et est réalisé en pratique à l'aide de l'algorithme EM [7], selon les formules de réestimation des paramètres décrites dans [15]. L'initialisation est effectuée à l'aide de l'algorithme des K-moyennes [9].

3.2.1. Distribution approchée du log spectre de mélange

Pour calculer l'estimateur minimisant l'EQM log spectrale, il faut pouvoir calculer la distribution de log spectre du mélange $\mathbf{X}_t=\log|X_t|$. En faisant l'approximation $|X_t(f)|^2\approx|V_t(f)|^2+|M_t(f)|^2$ [4], on obtient:

$$\mathbf{X}_t(f)\approx G[\mathbf{V}_t(f),\mathbf{M}_t(f)]\triangleq\frac{1}{2}\log\left[\exp\{2\mathbf{V}_t(f)\}+\exp\{2\mathbf{M}_t(f)\}\right] \quad (18)$$

1. Partout dans cet article l'opération $\log|\cdot|$ pour un vecteur ou une matrice s'applique élément par élément.

Les distributions de $\mathbf{V}_t$ et $\mathbf{M}_t$ sont supposées connues ((14) et (15)). Pour simplifier le calcul de la distribution de $\mathbf{X}_t$, la fonction non-linéaire G est généralement approchée [16, 17, 1]. Par exemple :

• **MIXMAX** (Mixture Maximum) [16]: la fonction G est approchée par le maximum de $\mathbf{V}_t(f)$ et $\mathbf{M}_t(f)$:

$$G[\mathbf{V}_t(f),\mathbf{M}_t(f)] \approx \max[\mathbf{V}_t(f),\mathbf{M}_t(f)] \tag{19}$$

• **VTS** (Vector Taylor Series) [17] : conditionnellement à la paire d'états (i,j), la fonction G est approchée par son développement en série de Taylor d'ordre 1, calculé au point $(\mu_{vi}(f),\mu_{mj}(f))$:

$$G[\mathbf{V}_t(f),\mathbf{M}_t(f)] \approx G[\mu_{vi}(f),\mu_{mj}(f)] + \nabla G[\mu_{vi}(f),\mu_{mj}(f)] \begin{bmatrix} \mathbf{V}_t(f) - \mu_{vi}(f) \\ \mathbf{M}_t(f) - \mu_{mj}(f) \end{bmatrix} \tag{20}$$

Toujours conditionnellement à (i,j), la distribution de $\mathbf{X}_t$ est gaussienne car l'approximation est linéaire et les distributions de $\mathbf{V}_t$ et $\mathbf{M}_t$ sont gaussiennes.

• **MeanMAX** (Mean Maximum) [1]: conditionnellement à (i,j), le mélange $\mathbf{X}_t$ est supposé gaussien avec les paramètres suivants :

$$\begin{cases} \mathbf{X}_t(f) \sim \mathcal{N}(\mu_{vi}(f),\sigma^2_{vi}(f)), & \mu_{vi}(f) \geq \mu_{mj}(f) \\ \mathbf{X}_t(f) \sim \mathcal{N}(\mu_{mj}(f),\sigma^2_{mj}(f)), & \mu_{vi}(f) \leq \mu_{mj}(f) \end{cases}$$

Ces approximations sont comparées sur la figure 4. Premièrement, on remarque qu'en éloignant les moyennes de $\mathbf{V}_t(f)$ et $\mathbf{M}_t(f)$, toutes les approximations se rapprochent de la distribution exacte. Deuxièmement, la précision des approximations décroît dans l'ordre de leur présentation. L'approximation MIXMAX est donc la plus précise et nous l'utiliserons par la suite.

3.2.2. Estimateur minimisant l'EQM log spectrale

On peut obtenir l'estimateur minimisant l'EQM log spectrale (10) en utilisant l'approximation MIXMAX (19). Le gain $\alpha_t(f)$ de cet estimateur se calcule comme suit [4] :

$$\log\, \alpha_t^{\log\text{-}\log}(f) = \sum_{i,j} \tilde{\gamma}_{i,j}(t) \frac{\left[\mu_{vi}(f) - \sigma^2_{vi}(f) R_{vit}(f) - \mathbf{X}_t(f)\right] R_{mjt}(f)}{R_{vit}(f) + R_{mjt}(f)}, \tag{21}$$

Les probabilités $\tilde{\gamma}_{i,j}(t)$ satisfont $\sum_{i,j} \tilde{\gamma}_{i,j}(t) = 1$ et sont calculées par (27). Les grandeurs $R_{vit}(f)$ et $R_{mjt}(f)$ sont définies par (28) et (29). Dans les formules (27), (28) et (29), les grandeurs suivantes sont utilisées :

$$\phi_{vit}(f) \triangleq N\left(\mathbf{X}_t(f); \mu_{vi}(f),\sigma^2_{vi}(f)\right) \tag{22}$$

$$\phi_{mjt}(f) \triangleq N\left(\mathbf{X}_t(f); \mu_{mj}(f),\sigma^2_{mj}(f)\right) \tag{23}$$

$$\Phi_{vit}(f) \triangleq \Phi\left[\frac{\mathbf{X}_t(f) - \mu_{vi}(f)}{\sigma_{vi}(f)}\right] \tag{24}$$

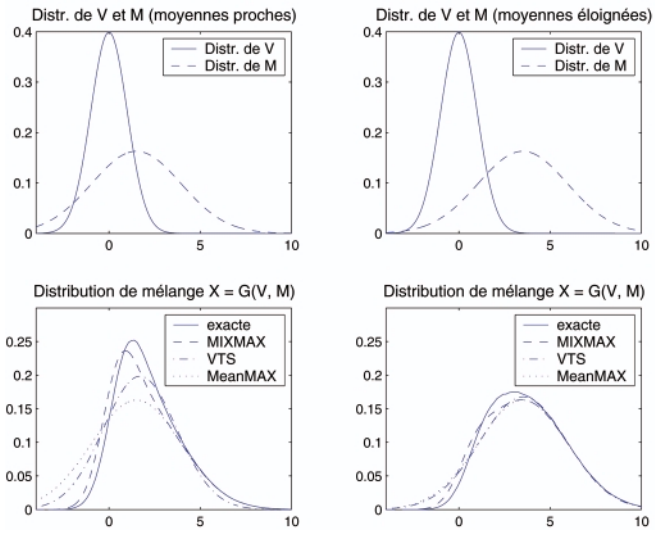


Figure 4. En haut : les distribution gaussiennes de $\mathbf{V}_t(f)$ et de $\mathbf{M}_t(f)$ conditionnellement à la paire d'états (i,j). Deux scénarios sont représentés : des moyennes proches (à gauche) et des moyennes éloignées (à droite). En bas : la distribution exacte du mélange $\mathbf{X}_t(f)$ et plusieurs approximations.

$$\Phi_{mjt}(f) \triangleq \Phi\left[\frac{\mathbf{X}_t(f) - \mu_{mj}(f)}{\sigma_{mj}(f)}\right] \tag{25}$$

où la densité $N(\cdot)$ est définie par (17) et la fonction Φ (la fonction de répartition de la loi normale centrée de variance unitaire) est définie comme

$$\Phi(\tau) = \int_{-\infty}^{\tau} \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}} dw = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\tau}{\sqrt{2}}\right)\right] \tag{26}$$

où `erf` s'appelle la fonction d'erreur et peut être calculée efficacement en utilisant une tabulation.

$$\tilde{\gamma}_{i,j}(t) \propto \omega_{vi}\omega_{mj} \prod_f \left[\phi_{vit}(f)\Phi_{mjt}(f) + \phi_{mjt}(f)\Phi_{vit}(f)\right] \tag{27}$$

$$R_{vit}(f) \triangleq \phi_{vit}(f)/\Phi_{vit}(f) \tag{28}$$

$$R_{mjt}(f) \triangleq \phi_{mjt}(f)/\Phi_{mjt}(f) \tag{29}$$

En ce qui concerne l'estimation de la phase de la TFCT, la même remarque que dans la section 3.1.4 peut être faite.

4. Mesures de performance

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

Pour un enregistrement (une chanson) donné(e), le RSD Normalisé (RSDN) mesure l'amélioration du Rapport Source à Distorsion (RSD) [18]

$$\mathrm{RSD}(\hat{v},v) = 10\ \log_{10}\left[\frac{\langle \hat{v},v\rangle^2}{\|\hat{v}\|^2\|v\|^2 - \langle \hat{v},v\rangle^2}\right] \tag{30}$$

entre le signal non traité x et la voix estimée $\hat{v}$:

$$\mathrm{RSDN}(\hat{v},v,x) = \mathrm{RSD}(\hat{v},v) - \mathrm{RSD}(x,v) \tag{31}$$

L'idée de cette normalisation est de combiner la mesure absolue $\mathrm{RSD}(\hat{v},v)$ avec la difficulté de la séparation pour la chanson traitée $\mathrm{RSD}(x,v)$. Cette difficulté est mesurée en tant que performance de la « séparation passive », c'est-à-dire quand le mélange x lui-même est pris à la place de l'estimation $\hat{v}$. Cette normalisation est représentée sur la figure 5.

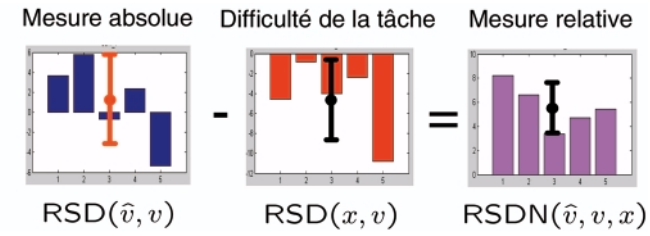


Figure 5. Interprétation du RSDN pour la séparation de 5 chansons différentes.

La DLS Normalisée (DLSN) est l'amélioration de la Distorsion du Log Spectre (DLS) [19]

$$\mathrm{DLS}(\hat{v},v) = \frac{1}{T}\sum_{t=0}^{T-1}\left[\frac{1}{F}\sum_{f=0}^{F-1}\left(10\ \log_{10}\frac{|V_t(f)|^2+\epsilon}{|\widehat{\mathcal{V}}_t(f)|^2+\epsilon}\right)^2\right]^{\frac{1}{2}} \tag{32}$$

entre le mélange x et la voix estimée $\hat{v}$:

$$\mathrm{DLSN}(\hat{v},v,x) = \mathrm{DLS}(x,v) - \mathrm{DLS}(\hat{v},v) \tag{33}$$

Dans l'équation (32) $V_t(f)$ et $\widehat{\mathcal{V}}_t(f)$ sont des TFCT de la voix $v(\tau)$ et de son estimée $\hat{v}(\tau)$ [2]. Pour calculer cette dernière TFCT, nous utilisons la même fenêtre d'analyse que celle utilisée pour la séparation. La constante ϵ est ajoutée pour éviter que ce critère ne devienne égal à $-\infty$ quand $V_t(f) = 0$. Elle correspond à un bruit dont l'énergie est typiquement 100 dB plus petite que celle de la source, autrement dit, elle est calculée comme suit : $\epsilon = 10^{-\mathbf{100}/10}\|V\|^2$.

Le choix de la mesure de performance dépend fortement de l'application pour laquelle la séparation est effectuée. Par exemple, si la séparation a pour but la Reconnaissance Automatique de la Parole (RAP) [20] sur le signal de voix obtenu, il semble plus judicieux de choisir la DLSN. En effet, les coefficients cepstraux [21] utilisés par la plupart des systèmes du RAP étant obtenus à partir des log spectres par une transformation linéaire, il est vraisemblablement préférable de minimiser la distorsion log spectrale. Cependant, nous ne donnons pas de préférence au RSDN ou à la DLSN pour des *applications de création*, c'est-à-dire des applications dont le but est de créer de nouveaux enregistrements à partir des sources séparées (par exemple, en remixant les sources). Ces enregistrements sont destinés à être écoutés et il est difficile de dire laquelle de ces deux mesures sera la plus appropriée pour l'écoute. Les deux possèdent des avantages et des inconvénients.

5. Adaptation des modèles et choix de la méthode de séparation

5.1. Adaptation des modèles

Dans la section 2, les modèles ont été présentés comme ayant été appris à partir des sources d'apprentissage $\tilde{v}$ et $\tilde{m}$ (Fig. 1), sans préciser exactement quelles données utiliser pour l'apprentissage.

Dans un premier temps, on peut utiliser comme données d'apprentissage, les signaux des sources que l'on veut séparer, c'est-à-dire les sources v et m avant le mélange. Cette approche n'est évidemment pas réaliste, mais elle nous permettra d'évaluer des bornes empiriques de performances qu'on peut espérer obtenir au mieux en adaptant les modèles. Nous allons appeler les modèles ainsi obtenus *modèles de référence*.

Une autre approche consiste à utiliser des *modèles généraux*, c'est-à-dire des modèles appris sur des sources issues d'autres enregistrements que ceux à séparer [1, 22]. Cette utilisation est plus réaliste. Par contre, la complexité calculatoire reste très élevée si les modèles sont de complexité importante. En effet, la complexité de l'approche utilisée est de l'ordre $n_1 \times n_2$ (voir par exemple (8) et (9)), où n_1 et n_2 sont les nombres de Gaussiennes de deux modèle *a priori*. Dans [1], $n_1 = n_2 = 8192$ et dans [22] $n_1 = n_2 = 512$, ce qui mène à des complexités d'ordre $6.7 \cdot 10^7$ et $2.6 \cdot 10^5$, respectivement[3]. Notons que la tâche considérée

2. Dans l'équation (32), nous avons volontairement noté la *TFCT de l'estimation* $\widehat{\mathcal{V}} = \mathrm{TFCT}[\mathrm{TFCT}^{-1}(\widehat{V})]$ par une lettre calligraphique pour la différencier de l'*estimation de la TFCT* $\widehat{V}$, qui en général ne coïncide pas avec $\widehat{\mathcal{V}}$. En effet, si $\widehat{V}$ était la TFCT d'un signal temporel, on aurait $\widehat{V} = \widehat{\mathcal{V}}$. Cependant, $\widehat{V}$ n'est pas généralement la TFCT d'un signal temporel car la TFCT est une transformée redondante, et l'image des signaux temporels dans l'espace de la TFCT ne couvre pas entièrement cet espace. De plus, l'estimation de la TFCT $\widehat{V}$ peut sortir de cette image, car elle est obtenue par un traitement non-linéaire à partir de X (Fig. 2). Ainsi, $\widehat{V}$ n'est pas obligatoirement la TFCT d'un signal temporel alors que $\widehat{\mathcal{V}}$ est la TFCT d'un signal temporel par définition.

3. Dans [1], une astuce pour réduire la complexité calculatoire est utilisée, mais cette astuce dépend de la particularité des modèles choisis et ne peut pas être appliquée directement ici.

dans les articles [1] et [22] concerne la séparation de parole homme / femme. Ces deux classes sonores exigent déjà une complexité calculatoire considérable. Dans notre cas, la classe de la musique est beaucoup plus vaste. En effet, à la variabilité des différentes combinaisons d'instruments s'ajoutent la variabilité des notes et la variabilité des accords pour chaque instrument. Il ne faut donc pas espérer décrire toute la musique par un MMG qui consisterait en un nombre de DSP limité. Même une augmentation importante de la taille de modèle ne permettra pas d'améliorer la modélisation. Elle ne conduira qu'à un sur-apprentissage sur la base d'entraînement.

L'autre approche proposée consiste à adapter les modèles généraux aux caractéristiques des sources mélangées, mais en n'utilisant que les informations fournies par le mélange x (Fig. 6). Les *modèles adaptés* obtenus ainsi peuvent être utilisés dans un cadre réaliste et nous nous attendons à ce qu'ils donnent de meilleures performances que les modèles généraux.

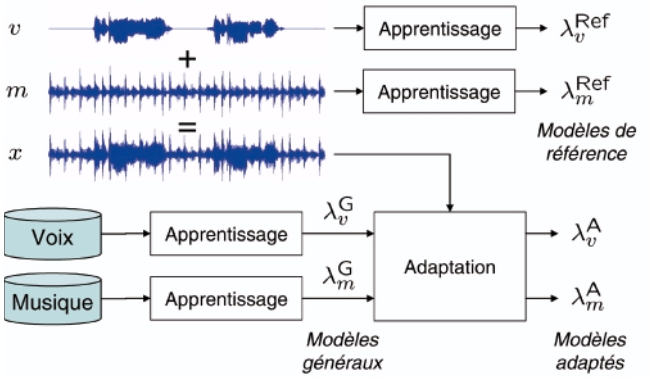


Figure 6. Modèles de référence, généraux et adaptés.

Pour effectuer cette adaptation, nous supposons dans un premier temps que chaque enregistrement est au préalable segmenté (manuellement) en parties vocales (contenant de la voix chantée) et non-vocales (sans voix chantée). Ainsi, le modèle de musique peut être adapté par apprentissage sur des segments non-vocaux.

Le principe de cette adaptation est très simple. Il consiste à prendre les trames non-vocales du mélange et à apprendre le modèle de musique sur ces nouvelles données d'entraînement.

Supposons que l'enregistrement $X = \{X_t\}_t$ (dans le domaine de la TFCT) est segmenté en parties vocales $\{X_t\}_{t\in\mathbf{voc}}$ et non-vocales $\{X_t\}_{t\notin\mathbf{voc}}$, où **voc** dénote l'ensemble des indices des trames vocales. Le modèle de musique λ_m^{A} est appris sur l'ensemble des trames non-vocales $\{X_t\}_{t\notin\mathbf{voc}}$ en utilisant le critère du MV :

$$\lambda_m^{\mathrm{A}} = \arg\max_{\lambda'_m} p(\mathcal{L}(\{X_t\}_{t\notin\mathbf{voc}})|\lambda'_m), \tag{34}$$

optimisé à l'aide de l'algorithme EM [7]. Par exemple, pour les MMG spectraux, cela mène à la formule suivante pour la réestimation des variances $\sigma_{mj}^2(f)$ [8] :

$$\sigma_{mj}^{2,(l+1)}(f) = \frac{\sum_{t\notin\mathbf{voc}} |X_t(f)|^2 \gamma_j^{(l)}(t)}{\sum_{t\notin\mathbf{voc}} \gamma_j^{(l)}(t)}, \tag{35}$$

où l'exposant (l) désigne les paramètres estimés à la l-ème itération de EM. Les poids $\gamma_j^{(l)}(t)$ satisfaisant $\sum_j \gamma_j^{(l)}(t) = 1$ sont calculés par :

$$\gamma_j^{(l)}(t) \propto \omega_{mj}^{(l)} N_C\left(X_t; \bar{0}, \Sigma_{mj}^{(l)}\right), \tag{36}$$

où la densité d'un vecteur gaussien complexe circulaire $N_C(\cdot)$ est définie selon l'équation (6). Une formule analogue à (35) permet de réestimer les poids ω_{mj}.

Malheureusement, on ne peut pas procéder de la même façon pour le modèle de voix puisque dans les chansons pop / rock, il n'y a en général pas de segments de voix chantée pure (sans musique d'accompagnement). Dans l'article [23], nous proposons des solutions pour l'adaptation du modèle de voix sur des segments vocaux en utilisant le modèle de musique adapté comme un *a priori*. Cette solution n'est pas traitée ici.

5.2. Choix de la méthode de séparation

Comme précisé dans la section 4, 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 à base des MMG présentées dans la section 3 avec les mesures RSDN et DLSN. Ces méthodes sont caractérisées par le domaine de modélisation (MMG spectral (Sec. 3.1) / log spectral (Sec. 3.2)) et le critère d'erreur minimisé (EQM spectrale (7) / log spectrale (10)).

Notons que pour un meilleur RSDN, il semble plus approprié d'utiliser des MMG spectraux et de minimiser l'EQM spectrale. En effet, la mesure RSD (30) est une fonction de l'EQM dans le domaine temporel à un gain multiplicatif près. Par conséquent, pour obtenir un meilleur RSDN, il faut minimiser l'EQM temporelle. Ainsi, sachant que la TFCT est une transformée linéaire, il semble plus judicieux de modéliser les spectres et de minimiser l'EQM spectrale, plutôt que de faire la même chose dans le domaine log spectral.

De même, il semble plus approprié d'utiliser les MMG log spectraux et de minimiser l'EQM log spectrale pour optimiser la DLSN.

Ainsi, nous nous attendons *a priori* à ce qu'en basculant progressivement du domaine spectral au domaine log spectral, le RSDN se dégrade et la DLSN s'améliore.

6. Cadre expérimental

6.1. 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 six 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. Ces chansons sont segmentées à la main en parties vocales et non-vocales.
Tous les enregistrements sont en mono et échantillonnés à 11025 Hz. Nous avons choisi cette fréquence d'échantillonnage car elle nous semble être un bon compromis entre la qualité et la complexité calculatoire. En particulier, ce choix est basé sur le fait qu'à l'heure actuelle, la qualité audio des signaux que l'on peut obtenir à l'aide des techniques existantes de séparation de sources avec un seul capteur est assez basse.

6.2. Estimateur oracle et limites de performance

Dans cette section, nous introduisons la notion d'estimateur oracle [24], qui permet de calculer les limites de performance qui ne peuvent pas être dépassées avec la méthode de séparation choisie.
Remarquons que l'estimation de la voix $\hat{v}$ ne dépend que du mélange x et de l'ensemble des gains $A = \{\alpha_t(f)\}_{t,f}$ (fig. 2). 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 (8) et $\mathcal{A}_+ = \{A|\alpha_t(f) \geq 0\}$ pour les méthodes (11) et (21). Etant donnée 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 meilleure performance [24] :

$$\tilde{v} = g(x,\tilde{A}), \quad \tilde{A} = \arg\max_{A\in\mathcal{A}} h(g(x,A),v,x) \tag{37}$$

Cet estimateur permet, pour un jeu de données, de calculer la limite de performance qui ne peut être dépassée avec la méthode correspondante.
Comme, il est difficile de calculer l'oracle (37) 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 par :

$$\text{RSB spec.} = 10\ \log_{10}\left[\frac{\|V\|^2}{\|\widehat{V}-V\|^2}\right] \tag{38}$$

Les valeurs du RSDN et de la DLSN (voir (31), (33)) calculées pour ces estimateurs oracles indiquent les performances qui pourraient être atteintes en améliorant l'estimation des gains $\alpha_t(f)$.
Selon l'ensemble des gains admissibles ($\mathcal{A}_{[0,1]}$ ou $\mathcal{A}_+$), les gains de ces estimateurs oracles se calculent comme suit

$$\tilde{\alpha}_t^{[0,1]}(f) = \min(\max[0,\tilde{\alpha}_t(f)],1) \quad \text{ou} \tag{39}$$
$$\tilde{\alpha}_t^{+}(f) = \max[0,\tilde{\alpha}_t(f)] \tag{40}$$

avec :

$$\tilde{\alpha}_t(f) = \frac{1}{|X_t(f)|^2}\text{Re}\left(X_t(f)\overline{V_t(f)}\right) \tag{41}$$

où $\bar{Y}$ représente le conjugué d'un nombre complexe Y.

7. Expérimentations et résultats

Les problèmes qui se posent et que nous allons étudier dans cette section, sont les suivants [4] :
1. choix des paramètres de la TFCT (taille et type de la fenêtre d'analyse)
2. adaptation des modèles (étude de performance des modèles généraux, adaptés et de référence) et leurs dimensionnement (nombre de gaussiennes)
3. choix du domaine de modélisation et de la mesure de distorsion minimisée (spectre / log spectre)

7.1. Choix de la fenêtre d'analyse

En utilisant l'estimateur oracle avec le gain $\tilde{\alpha}_t^{[0,1]}(f)$ (39), nous avons fait varier la taille et le type de fenêtre d'analyse dans la TFCT (Fig. 7). Le meilleur résultat est obtenu avec une fenêtre de Hamming de taille 1024 échantillons (soit 93 ms). Cette fenêtre a été utilisée pour le reste des expériences.

7.2. Adaptation et dimensionnement des modèles

Avec les MMG spectraux et l'estimateur minimisant l'EQM spectrale (8), nous avons testé l'effet sur le RSDN du nombre de gaussiennes des MMG de voix $n_v = 1,2,4,\ldots,128$ et de

4. Par rapport aux données expérimentales utilisées dans l'article [25], la base de test a été agrandie et une modification légère des segmentations manuelles a été effectuée. C'est pour cela que les valeurs de mesure sont légèrement différentes de celles de [25], mais toutes les conclusions restent les mêmes.

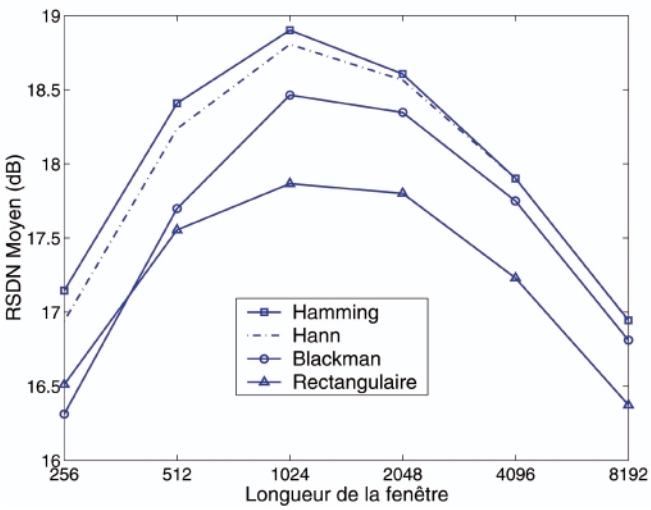


Figure 7. RSDN pour l'estimateur oracle en fonction de la taille et du type de fenêtre d'analyse.

musique $n_m = 1,2,4,\ldots,128$ dans les quatre configurations suivantes :

1. modèles généraux de voix λ_v^{G} et de musique λ_m^{G},
2. modèle général de voix λ_v^{G} et modèle adapté de musique λ_m^{A} (appris sur les parties non-vocales),
3. modèle de référence de voix λ_v^{Ref} (appris sur la voix séparée) et modèle général de musique λ_m^{G}.
4. modèle de référence de voix λ_v^{Ref} et modèle adapté de musique λ_m^{A}.

Les résultats sont résumés sur la figure 8.

Avec deux modèles généraux (Fig. 8 (A)), 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. Remarquons que dans le cas $n_v = n_m = 1$, le RSDN moyen de 5.5 dB est obtenu par un filtrage linéaire simple avec un filtre passe-haut dont la fréquence de coupure est située vers 300 Hz (Fig. 9). En observant les modèles généraux à 16 états (Fig. 10 (A) et (C)) il semble que le modèle de voix est plus structuré que celui de musique.

Dans le cas du filtrage de Wiener ($n_v = n_m = 1$), l'adaptation du modèle de musique (Fig. 8 (B)) augmente le RSDN de 1.5 dB par rapport aux modèles généraux (Fig. 8 (A)). De plus, 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 3 dB supérieur à celui du filtrage de Wiener ($n_v = n_m = 1$) (Fig. 8 (B)) et de 4 dB supérieur au meilleur résultat avec les deux modèles généraux (Fig. 8 (A)). Un modèle de musique adapté est représenté sur la figure 10 (D).

Pour le modèle de voix de référence et le modèle de musique général (cette configuration est irréaliste en pratique) (Fig. 8 (C)) la courbe représentant l'évolution de performance ressemble à celle obtenue avec le modèle de voix général et le modèle de musique adapté (Fig. 8 (B)). Finalement, les résultats obtenus avec le modèle de voix de référence et le modèle de musique adapté (Fig. 8 (D)) (également irréalistes en pratique) montrent qu'il reste encore une marge de 3 dB pour l'adaptation du modèle de voix. Cette adaptation est étudiée dans l'article [23].

Contrairement aux travaux présentés dans [1] et [22], nous n'avons testé des modèles qu'avec au plus 128 gaussiennes parce que nous étions limités par les ressources calculatoires disponibles. Notons que, vu le comportement de la courbe sur la figure 8 (A), l'utilisation de 512 ou même 8192 gaussiennes par modèle n'est probablement pas de nature à améliorer significativement les performances de séparation. En revanche, nous visons et parvenons à améliorer ces performances en adaptant le modèle de musique, tout en gardant une complexité calculatoire raisonnable.

7.3. Effets du domaine de modélisation et de la mesure de distorsion

Nous avons finalement comparé les performances des algorithmes en fonction du domaine de modélisation (MMG spectral / log spectral) et de la mesure de distorsion (EQM spectrale / log spectrale), dans la configuration (modèle de voix général, modèle de musique adapté) et $n_v = n_m = 64$. Pour chaque paire (modèle, mesure de distorsion), les deux mesures de performance (le RSDN et la DLSN) sont calculées. Les résultats accompagnés des références de performance obtenues avec des estimateurs oracles sont résumés dans le tableau 1. Comme attendu (Sec. 5.2), 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 moins favorable 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'équation (32). Cette mesure est ajoutée à titre informatif, puisque elle ne peut pas être calculée quand la séparation est terminée du fait que $\widehat{V}_t(f)$ n'est plus accessible [5]. Remarquons que la mesure DLSN' peut avoir un sens dans le cas où l'on n'est pas intéressé par la reconstruction du signal dans le domaine temporel, et que l'on peut utiliser directement l'estimation de la TFCT $\widehat{V}_t(f)$. Par exemple, pour la reconnaissance automatique de la parole [20], il est possible de calculer les coefficients cepstraux [21] directement à partir de $\widehat{V}_t(f)$.

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.).

5. Comme on le voit sur la figure 2, on n'a plus accès à l'estimation de la TFCT $\widehat{V}_t(f)$ après la reconstruction du signal dans le domaine temporel $\hat{v}(\tau)$ en utilisant la méthode OLA.

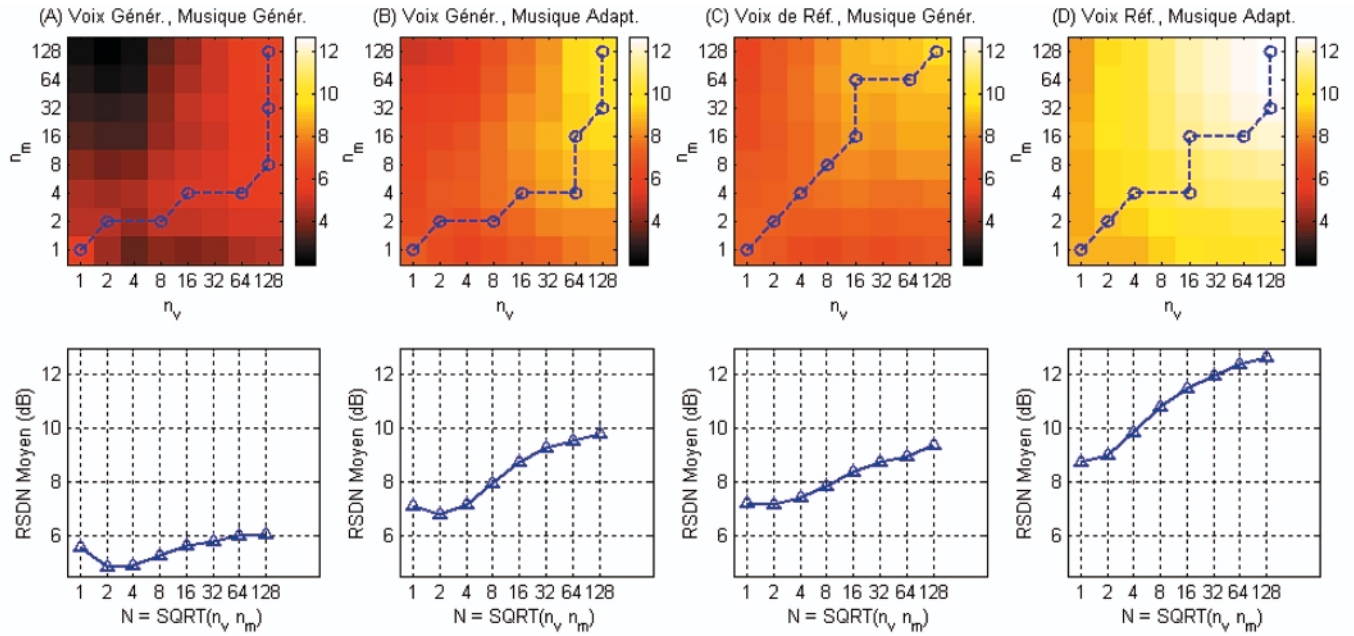


Figure 8. En haut : Les RSDN en fonction de (n_v,n_m)*. La ligne pointillée représente, pour un N donné (*$N = 1,2,4,\ldots,128$*), la paire* (n_v^*,n_m^*) *qui fournit la meilleure performance parmi toutes les paires* (n_v,n_m)*, telles que* $n_v n_m = N^2$*. En bas : Le RSDN le long de la ligne pointillée. De gauche à droite :* **(A)** *:* $(\lambda_v^{\mathrm{G}},\lambda_m^{\mathrm{G}})$*,* **(B)** *:* $(\lambda_v^{\mathrm{G}},\lambda_m^{\mathrm{A}})$*,* **(C)** *:* $(\lambda_v^{\mathrm{Ref}},\lambda_m^{\mathrm{G}})$*,* **(D)** *:* $(\lambda_v^{\mathrm{Ref}},\lambda_m^{\mathrm{A}})$*. Une référence de performance obtenue avec l'estimateur oracle est légèrement inferieure à 19 dB (Fig. 7).*

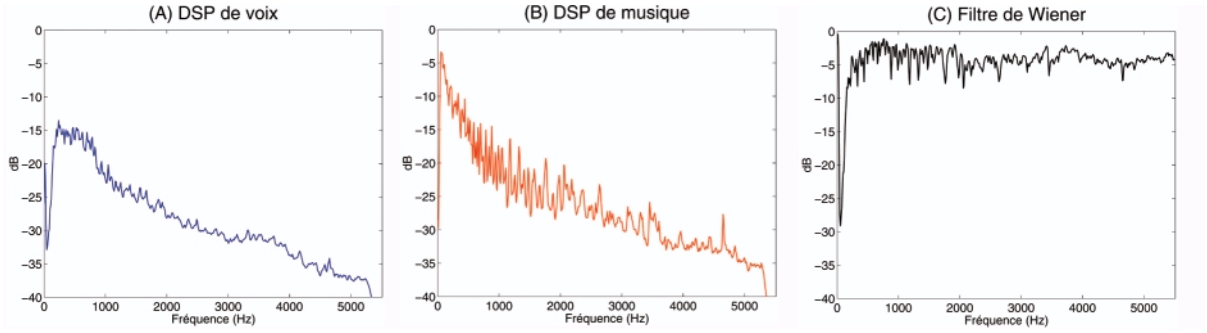


Figure 9. **(A)** *: MMG de voix à 1 état, c'est-à-dire la DSP de voix.* **(B)** *: DSP de musique.* **(C)** *: Filtre de Wiener pour l'estimation de voix.*

Tableau 1. Performances des méthodes (modèle MMG / EQM minimisée). Les références de performance obtenues à l'aide des oracles sont indiquées entre parenthèses. La DLSN' est ajoutée à titre informatif.

MMG / EQM	RSDN	DLSN	DLSN'
spectral / spec [2]	9.4 (18.9)	4.3 (9.5)	*1.2*
spectral / log spec [3]	8.7 (19.6)	3.5 (9.7)	*3.3*
log spec / log spec [4]	6.8 (19.6)	4.8 (9.7)	*3.8*

7.3.1. Tests d'écoute informels

Pour rendre compte *subjectivement* des résultats de séparation obtenus en général ainsi qu'en utilisant différents modèles et estimateurs, nous rapportons ici quelques tests d'écoute informels. Le protocole expérimental est très simple : le premier auteur de cet article écoute les résultats de séparation (voix estimée et musique estimée) pour lesquels les valeurs des mesures objectives sont résumées dans le tableau 1. Ensuite, il décrit ce qu'il a perçu. Ses remarques sont résumées ci-dessous :

- **Estimation de voix** (remarques générales sur toutes les méthodes) : La voix estimée reste compréhensible. La musique est bien supprimée en général, mais il y a des périodes où certains instruments s'entendent toujours. Nous pensons que ce problème est surtout lié à la différence entre la musique issue des parties non-vocales et celle issue des parties vocales. De plus, on entend fréquemment des artefacts qui ressemblent à un bruit de frottement rythmique. Ce sont vraisemblablement des traces de la suppression de la musique qui sont en général assez corrélées au rythme.

- **MMG spectral / EQM spectrale** : Parmi les trois méthodes testées, l'estimation de voix obtenue avec cette méthode semble être la plus satisfaisante à l'écoute.

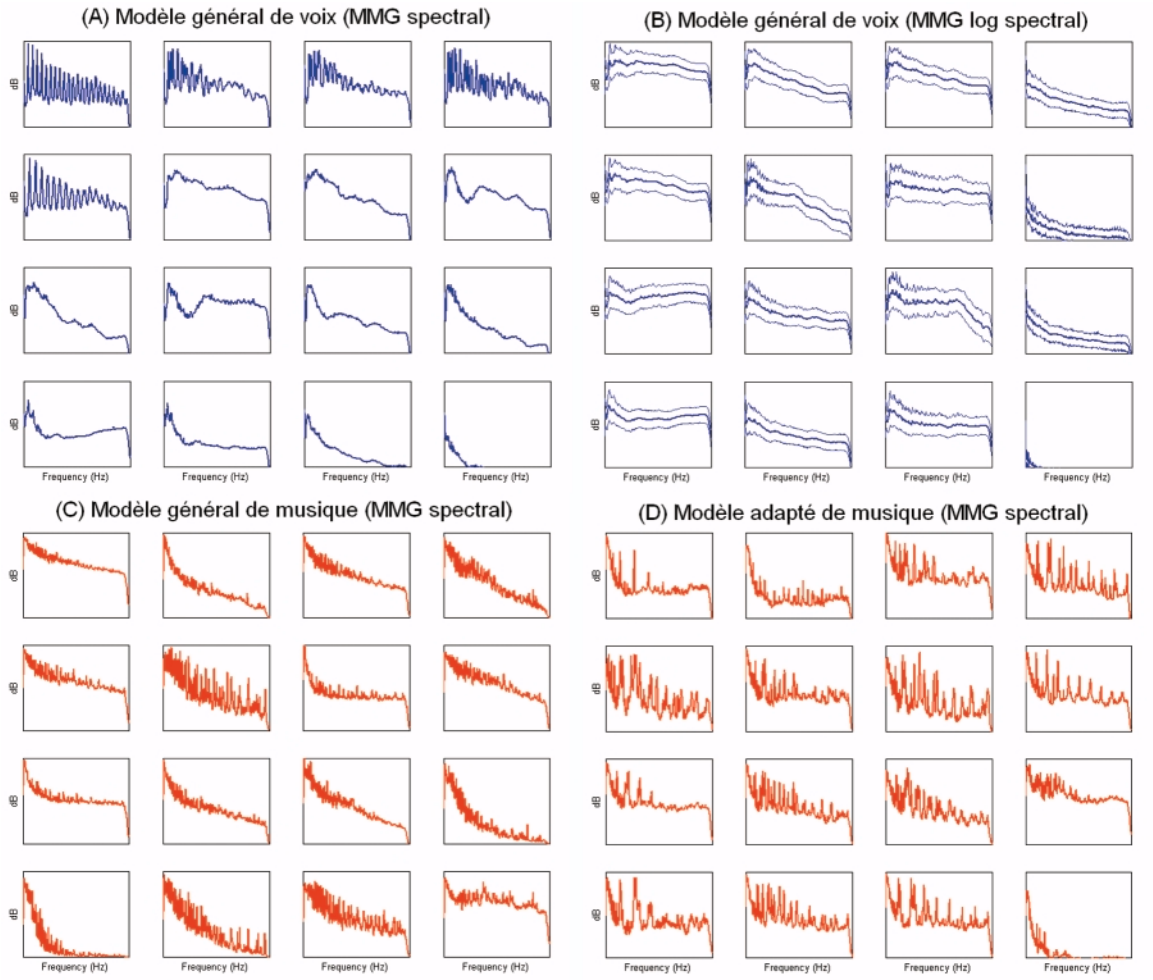


Figure 10. MMG à 16 états. Pour les MMG spectraux, chaque état i est représenté par sa DSP : $\log \sigma_i^2(f)$. Pour le MMG log spectral, chaque état i est représenté par sa DSP : $\mu_i(f)$ et DSP ± l'écart-type : $\mu_i(f) \pm \sigma_i(f)$. **(A)** *: Modèle général de voix (spectral).* **(B)** *: Modèle général de voix (log spectral).* **(C)** *: Modèle général de musique (spectral).* **(D)** *: Modèle adapté de musique (spectral).*

- **MMG spectral / EQM log spectrale** : Pour cette méthode le bruit de frottement mentionné ci-dessus est le plus perceptible par rapport aux deux autres méthodes.
- **MMG log spectral / EQM log spectrale** : La coloration du son change par rapport aux deux autres méthodes. Il est difficile de décrire la différence. De plus, le « bruit musical » (un artefact agaçant, connu surtout en débruitage de la parole) est perceptible par endroits.

- **Estimation de musique** (remarques générales sur toutes les méthodes) : Dans la plupart des cas, la voix s'entend plus dans l'estimation de musique que la musique dans l'estimation de voix. Ainsi, on peut dire que, perceptivement, la séparation marche mieux pour extraire la voix d'une chanson plutôt que pour l'éliminer.

Nous ne décrivons pas ici de résultats détaillés pour l'estimation de musique obtenue à l'aide des trois méthodes testées puisque dans cet article nous nous focalisons surtout sur l'estimation de voix.

Les descriptions que nous venons de présenter donnent une idée sur les résultats de séparation que l'on peut obtenir à l'aide des méthodes étudiées. Cependant, en général, les tests d'écoute (même les plus formels) ne sont pas nécessairement la manière la plus sûre d'évaluer les algorithmes de séparation. Tout dépend de l'application. En particulier, pour des *applications d'extraction des informations sémantiques* (reconnaissance de la parole, transcription de la mélodie, reconnaissance du locuteur / chanteur, etc.) les tests d'écoute peuvent être trompeurs.

8. Conclusions et perspectives

Dans cet article, il a été montré expérimentalement que les modèles généraux ne sont pas applicables pour la séparation

voix / musique à cause de faibles performances de séparation qui, de plus, ne croissent pratiquement pas avec l'augmentation du dimensionnement des modèles au-delà d'une certaine complexité. Ainsi, l'utilisation de modèles adaptés aux sources mélangées est proposée. Nous introduisons une procédure d'adaptation du modèle de musique permettant de rester dans un cadre d'utilisation réaliste, puisque la segmentation manuelle en parties vocales et non-vocales peut être effectuée par un utilisateur. Dans le cadre de notre étude, nous avons montré que cette adaptation permet d'améliorer de 4 dB la performance de séparation par rapport au cas de deux modèles généraux [6]. Des travaux en cours consistent à remplacer la segmentation manuelle en parties vocales et non-vocales par un module de segmentation automatique [26].

Les effets de la mesure de distorsion minimisée et du domaine de modélisation ont été étudiés en utilisant deux mesures de performance différentes. Cette étude donne des indications sur le choix d'une méthode de séparation en fonction de la mesure de performance, qui dépend également de la tâche pour laquelle la séparation est effectuée.

Références

[1] S.T. ROWEIS. "One microphone source separation", in *Advances in Neural information Processing Systems*, vol. 13, MIT Press, 2001, pp. 793-799.

[2] L. BENAROYA. « Séparation de plusieurs sources sonores avec un seul microphone », Ph.D. dissertation, Université de Rennes 1, 2003.

[3] 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.

[4] D. BURSHTEIN and S. GANNOT. "Speech enhancement using a mixture-maximum model", in *European Conf. on Speech Communication and Technology (EuroSpeech'99)*, ol. 6, Budapest, Hungary, Sep 1999, pp. 2591-2594.

[5] S. M. KAY. *Fundamentals of Statistical Signal Processing, Estimation Theory*. Prentice Hamm, 1993.

[6] G. PEETERS and X. RODET. "SINOLA : A new analysis/synthesis method using spectrum peak shape distortion, phase and reassigned spectrum", in *International Computer Music Conference (ICMC'99)*, Oct. 1999, pp. 153-156.

[7] A. P. DEMPSTER, N. M. LAIRD, and D. B. RUBIN. "Maximum likehood from incomplete data via the EM algorithm," J*ournal of the Royal Statistical Society*, vol. 39, pp. 1-38, 1977.

[8] L. BENAROYA and F. BIMBOT. "Wiener based source separation with HMM/GMM using a single sensor," in *Intl. Conf. on Indep. Component Analysis and Blind Source Separation (ICA'03)*, Nara, Japan, Apr. 2003, pp. 957-961.

[9] J. MCQUEEN. "Some methods for classification and analysis of multivariate obervations," in *5th Berkeley Symposium on mathematics, Statistics and Probability*, 1967, pp. 281-298.

[10] L. BENAROYA, F. BIMBOT, and R. GRIBONVAL. "Audio source separation with a single sensor," *IEEE Trans. Audio, Speech and Language Proc.*, vol. 14, no. 1, pp. 191-199, january 2006.

[11] F. D. NEESER AND J. L. MASSEY. " Proper complex random processes with applications to information theory, *IEEE Trans. inform. Theory*, vol. 39, no. 4, pp. 1293-1302, July 1993.

[12] B. PICINBONO. "Second-order complex random vectors and normal distributions," *IEEE Trans. Signal Processing*, vol. 44, no. 10, pp. 2637-2640, October 1996.

[13] W. H. PRESS, B. P. FLANNERY, S. A. TEUKOLSKY, and W. T. VETTERLING. *Numerical Recipes in C. : The Art of Scientific Computing*, 2nd ed. Cambridge University Press, October 1992. [Online]. Available : http://www.library.cornell.edu/nr/bookcpdf.htlm

[14] Y. EPHRAIM. "A bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Processing*, vol. SP-40, pp. 725-735, April 1992.

[15] L.R. RABINER. "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286, 1989.

[16] 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.

[17] P.J. MORENO, B. RAJ, and R. M. STERN. "A vector taylor series approach for environment-independent speech recognition," in *IEEE intl. Conf. on Acoustics, Speech and Signal Proc. (ICASSP'96)*, vol. 2, 1996.

[18] R. GRIBONVAL, L. BENAROYA, E. VINCENT, and C. FÉVOTTE. "Proposals for performance measurement in source separation," in *Intl. Conf. Indep. Component Analysis and Blind Source Separation (ICA'03)*, April 2003, pp. 763-768.

[19] J.-M. VALIN, J. ROUAT, and F. MICHAUD. "Microphone array post-filter for separation of simultaneous non-stationary sources," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Proc. (ICASSP'04)*, 2004.

[20] L. RABINER and B.-H. JUANG. *Fundamentals of speech recognition*. Englewood Cliffs, n.J. : Prentice Hall, 1993.

[21] R. VERGIN, D. O'SHAUGHNESSY, and A. FARHAT. "Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independant continuous-speech recognition," *IEEE Trans. on Speech and Audio Proc.*, vol. no. 5, pp. 525-532, Sep 1999.

[22] T. KRISTJANSSON, H. ATTIAS, and J. HERSHEY, "Single microphone source separation using high resolution signal recontruction," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Proc. (ICASSP'04)*, vol. 2, 2004, pp. 817-820.

[23] A. OZEROV, P. PHILIPPE, R. GRIBONVAL, and F. BIMBOT. "One microphone singing voice separation using source-adapted models," in *IEEE Worksh. on Apps. of Signal Processing to Audio and Acoustics (WASPAA'05),* Mohonk, NY, Oct. 2005, pp. 90-93.

[24] E. VINCENT and R. GRIBONVAL. « Construction d'estimateurs oracles pour la séparation de sources », in *GRETSI'05 Syposium on Signal and Image Processing*, Louvain-la-Neuve, Belgium, 2005.

[25] A. OZEROV, R. GRIBONVAL, P. PHILIPPE, and F. BIMBOT. « Séparation voix/musique à partir d'enregistrements mono quelques remarques sur le choix et l'adaptation des modèles », in *GRETSI'05 Symposium on Signal and Image Processing*, Louvain-la-Neuve, Belgique, Sept. 2005.

[26] W.-H. TSAI, D. ROGERS, and H.-M. WANG. "Blind clustering of popular music recordings based on singer voice characteristics," *Computer Music Journal*, vol. 28, no. 3, pp. 68-78, 2004.

6. Quelques exemples de séparation se trouvent sur `www.irisa.fr/metiss/ozerov/demos.html`

Alexey **Ozerov**

Alexey Ozerov a obtenu en 2006 un doctorat de l'Université Rennes 1, à l'issue des travaux qu'il a mené à Orange Labs et à l'IRISA. Il est également titulaire d'un master en mathématiques de l'Université d'Etat de Saint-Pétersbourg (1999) et d'un master en mathématiques appliquées de l'Université Bordeaux 1 (2003). Entre 1999 et 2002, il a travaillé à Terayon Communicational Systems comme ingénieur logiciel R&D, d'abord à Saint-Pétersbourg, puis à Prague. Maintenant, il est en post-doc au laboratoire du Traitement du Signal et d'Image (SIP) au KTH, Stockholm. Ses intérêts scientifiques contiennent la reconnaissance de la parole, la séparation de sources audio et le codage de source.

Pierrick **Philippe**

Pierrick Philippe est titulaire d'un Doctorat de l'Université de Paris Orsay (1995). Avant de rejoindre Orange Labs (France Telecom R&D), il a travaillé comme expert audio à Innova Son, TDF et Envivio. Ses domaines d'intérêt sont axés sur le traitement du signal audio en général, et concernent tout particulièrement le codage à réduction de débit. Il contribue activement à ISO/MPEG audio depuis 1998.

Rémi **Gribonval**

Ancien élève de l'Ecole Normale Supérieure de Paris, Rémi Gribonval obtient en 1999 un doctorat en Mathématiques Appliquées de l'Université Paris IX Dauphine. Il effectue alors un séjour post-doctoral à l'Institut de Mathématiques Industrielles de l'Université de Caroline du Sud avant de rejoindre l'INRIA en tant que chargé de recherche en 2000. Ses travaux de recherche, au sein du projet de recherche METISS de l'IRISA à Rennes, sont dédiés à la modélisation de scènes sonores et la séparation de sources, avec un accent particulier sur les aspects théoriques et algorithmiques des décompositions parcimonieuses de signaux avec des dictionnaires redondants.

Frédéric **Bimbot**

Frédéric Bimbot est Ingénieur Télécom-Paris (1985) et a obtenu un Doctorat de l'ENST (spécialité « Signal et Image ») sur la décomposition temporelle du signal de parole (1988). Il détient également une Maîtrise de Linguistique de Paris III (1987). Il est Chargé de Recherche au CNRS depuis 1990, affecté à l'ENST, puis à l'IRISA (depuis 1997). Entre 1990 et 1999, il a également effectué plusieurs séjours chez AT&T (Bell) Labs. Ses travaux de recherche portent sur le traitement de la parole et du signal sonore, avec une focalisation particulière sur la caractérisation du locuteur, la séparation de sources audio et les méthodologies d'évaluation. Depuis 2002, il est responsable scientifique de l'équipe METISS à l'IRISA.