Classification pondérée pour la séparation de sources sous-déterminée

ZAHER EL CHAMI[1], ANTOINE PHAM[2], CHRISTINE SERVIERE[3], ALEXANDRE GUERIN[1]

[1] Orange Labs,
2 avenue Pierre Marzin, 22307 Lannion, France
[2] Laboratoire Jean Kuntzmann
51 rue des Mathématiques, 38400 Saint Martin d'Hères, France
[3] Gipsa-lab, département Image-Signal
961 rue de la Houille Blanche, 38402 Grenoble, France

[1]zaher.elchami@orange-ftgroup.com, [1]alexandre.guerin@orange-ftgroup.com,
[2]dinh-tuan.pham@imag.fr, [3]christine.serviere@gipsa-lab.inpg.fr

Résumé – Dans ce papier, on s'intéresse à la séparation de sources aveugle dite sous-déterminée, en se focalisant sur l'influence de certains critères sur les performances des algorithmes de séparation par classification de type Kmeans. On montre notamment que la classification sur la base du couple [log(ILD), IPD] donne de meilleurs résultats, par rapport au couple traditionnel [ILD, IPD], ou dérivés comme ceux utilisés dans [1]. D'autre part, on montre aussi que l'ajout de fonctions de pondération permet de légèrement améliorer les performances de séparation en pondérant les caractéristiques utilisées par les algorithmes de classification.

Abstract – In this paper, the feature and weight choice impact on the source separation performance is studied. We show that using the feature couple [log(ILD), IPD] instead of the traditional one [ILD, IPD] gives better separation performance. Also we show that introducing a weight in the clustering algorithm enhances the separation performance in terms of objective criteria.

1 Introduction

Le problème de séparation de source sous déterminé (nombre de source *N* plus grand que celui de mélange *M*) a connu un grand essor ces dernières années. Suite à l'hypothèse de parcimonie, introduite par Van Hulle [1] pour un mélange instantané et étendue au cas anéchoïque par Rickard [3], séparer un nombre de sources supérieur au nombre de capteurs est devenu possible. Plus précisément, c'est l'hypothèse de support disjoint, généralement associée à la parcimonie, qui est largement utilisée. Sous cette hypothèse, à chaque instant temps-fréquence, une source au maximum est dominante et donc une information (caractéristique) de cette source peut être extraite. En utilisant un algorithme de catégorisation, ces caractéristiques sont classifiées dans des groupes, pour en déduire des masques de séparation binaire.

Les performances de la séparation sont fortement liées au choix des caractéristiques à classifier et à l'algorithme de catégorisation utilisé. Plusieurs ensembles de caractéristiques, dont le traditionnel couple *Interchanel Level/Phase Difference* (ILD/IPD), ont été testés et classifiés dans [2] par le K-moyen (ou Kmeans). Dans cet article, on propose de réaliser la classification à l'aide un nouveau couple de caractéristiques (log(ILD) / IPD). Ce couple a déjà donné de bonnes performances dans [4] où il était modélisé. Au contraire de [2], on se propose de classifier ce couple à l'aide d'un Kmeans pondéré. En termes de performance, le nouveau couple montre sa supériorité par rapport aux autres caractéristiques présentées dans [2]. On montre aussi l'apport de la pondération sur les performances de la séparation, où deux types de pondérations sont testés (rapport des valeurs propres et énergie)

2 Formulation du problème

Le problème de la séparation de sources consiste à estimer N sources, $S(t)=\left[s_1(t),\ldots,s_N(t)\right]$, à partir de M mélanges observés, $X(t)=\left[x_1(t),\ldots,x_M(t)\right]$, qui, pour un modèle convolutif, s'écrivent :

$$x_j(t)=\sum_{i=1}^{N}x_j^{(i)}(t)=\sum_{i=1}^{N}\sum_{k}a_{ji}(k)s_i(t-k)$$

où $x_j^{(i)}(t)$ est la contribution de la $i^{\text{ème}}$ source à la $j^{\text{ème}}$ observation. Les valeurs $a_{ji}(k)$ sont les coefficients du filtre mélangeant cette source *i* sur l'observation *j*. Pour des sources s_i non-stationnaires, le modèle de mélange convolutif est ramené à un modèle de mélange instantané par la transformation de Fourier à court terme (TFCT) associé à une fenêtre apodisante *w*(*k*) (celle de Hanning par ex.). Les observations temps-fréquence (TF) sont alors données par :

$$\begin{aligned}X_j(t,\omega)&=\sum_{k=0}^{L-1}w(k)x_j(t+k)e^{-j\omega k}\\&=\sum_{i=1}^{N}X_j^{(i)}(t,\omega)=\sum_{i=1}^{N}A_{ji}(\omega)S_i(t,\omega)\end{aligned}\qquad(1.1)$$

où $X_j^{(i)}(t,\omega)$ et $S_i(t,\omega)$ sont respectivement les TFCT de $x_j^{(i)}(t)$ et $s_i(t)$. $A_{ji}(\omega)$ est la transformée de Fourier de $a_{ji}(k)$. On s'intéresse au cas sous-déterminé: *N*>*M*. Par suite, même après identification de la matrice de

mélange, le système n'est pas inversible et donc les sources ne peuvent pas être séparées linéairement. Pour résoudre ce problème, il est nécessaire d'avoir des informations *a priori* sur les sources, qui nous sont données par l'hypothèse de parcimonie.

3 Séparation de sources basée parcimonie

Une source est dite parcimonieuse dans un espace donné (ici : temps-fréquence) quand elle est représentée par un nombre réduit de composantes. Ainsi, on peut supposer que, dans cet espace, les supports des différentes sources parcimonieuses sont disjoints. Sous cette hypothèse, une seule source au maximum est dominante à chaque instant TF et l'observation (1.1) se réduit à :

$$X_j(t,\omega) = \sum_{i=1}^{N} X_j^{(i)}(t,\omega) \approx X_j^{(q)}(t,\omega) \approx A_{jq}(\omega) S_q(t,\omega)$$

où q représente l'indice de la source dominante. Il est alors possible d'extraire chacune des sources s_q par application d'un masque binaire B_q. Pour estimer ces masques, nous choisissons d'utiliser les algorithmes de catégorisation, de part leur simplicité et leur rapidité de calcul. Ainsi, l'algorithme K-Means sera mis en œuvre comme cela est fait dans [2] mais avec deux différences majeures : les caractéristiques spatiales permettant de discriminer les sources et l'introduction d'une pondération permettant d'atténuer l'influence des *outliers*

3.1 Caractéristiques (*Features*)

Les caractéristiques (*Features* en anglais) $\Theta(t,\omega)$ à classifier jouent un rôle important dans les performances de la séparation. Pour un mélange stéréo, le couple *Interchannel Level/Phase Difference* [ILD, IPD] est largement utilisé. Sans perte de généralité, si on définit $A_{1i}=1$ et $A_{2i}=a_i(\omega)e^{j\delta q(\omega)}$, le rapport entre les observations s'écrit alors :

$$\frac{X_2(t,\omega)}{X_1(t,\omega)} \approx \frac{X_2^{(q)}(t,\omega)}{X_1^{(q)}(t,\omega)} \approx \frac{A_{2q}(\omega)}{A_{1q}(\omega)} = a_q(\omega) e^{j\delta_q(\omega)}.$$

Le module et la phase de ce rapport à une fréquence donnée ω, qui donnent respectivement l'ILD et l'IPD, ne dépendent que de l'indice q de la source dominante. Ainsi, ils caractérisent les sources et peuvent être utilisés comme paramètres dans les algorithmes de catégorisation pour grouper les ensembles des points TF appartenant à la même source. Dans [2], l'auteur exploite ce couple de caractéristiques [ILD, IPD] ainsi que d'autres, dont celui ayant les meilleures performances de séparation [$|X_1|/U$, $|X_2|/U$, IPD] avec $U = \sqrt{|X_1|^2 + |X_2|^2}$.

La figure 1 trace la dispersion du couple (ILD/IPD) dans la bande de fréquence 2.7 kHz pour un mélange de 3 sources : les points où chaque source est prépondérante sont repérés par des marqueurs différents. Il apparaît que les caractéristiques des sources sont regroupées en 3 nuages parfaitement distincts de l'espace (ILD/IPD). Par suite, un simple algorithme de catégorisation sera apte à identifier les

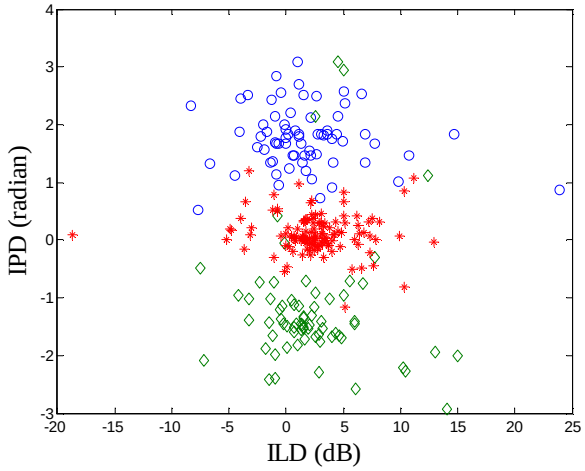

Fig.1 Dispersion du couple (ILD, IPD) pour un mélange de 3 sources à la fréquence 2.7 kHz

centroïdes de ces trois groupes ainsi que les masques binaires de séparation par une simple comparaison des distances aux centroïdes. Dans cet article, nous proposons d'utiliser le couple [log(ILD), IPD] : l'emploi de l'indice de localisation log(ILD) étant motivé par sa variance finie, contrairement à celle infinie de la variable ILD [4], rendant de fait sa modélisation impossible

4 K-Means

L'objectif du K-Means est de classifier les données en minimisant la distance euclidienne entre les membres d'un cluster C_i et son centre $\bar{c}_i$. La fonction à minimiser est donc donnée par :

$$\zeta = \sum_{i=1}^{N} \zeta_i \;;\; \zeta_i = \sum_{\Theta(t,\omega)\in C_i} \left\| \Theta(t,\omega) - \bar{c}_i \right\|^2 \quad .(1.2)$$

avec ζ_i la somme des distances euclidienne entre le centre d'un cluster et les points formant ce cluster. Après une initialisation appropriée des centres $\bar{c}_i$, la fonction de coût ζ est minimisée par les itérations suivantes :

$$C_i = \left\{ \Theta(t,\omega) \;\middle|\; i = \arg\min_i \left\| \Theta(t,\omega) - \bar{c}_i \right\|^2 \right\}$$

$$\bar{c}_i = \sum_{\Theta(t,\omega)\in C_i} \left\| \Theta(t,\omega) \right\|^2$$

Ces itérations sont répétées jusqu'à ce que tous les centres $\bar{c}_i$ ne changent pas. Après convergence, dans une bande de fréquence ω, l'extraction de chacune des sources s_i est réalisée par le masque binaire $B_i(t,\omega)$ défini par :

$$B_i(t,\omega) = \begin{cases} 1 \text{ si } \Theta(t,\omega) \in C_i \\ 0 \text{ ailleurs} \end{cases}$$

Les estimées des images des sources originales, issues de chacune des observations, sont alors données par :

$$\hat{X}_j^{(i)}(t,\omega) = B_i(t,\omega) X_j(t,\omega)$$

Noter que la séparation des sources est réalisée dans chaque bande de fréquence indépendamment des autres, d'où la nécessité de résoudre l'ambiguïté de permutation entre les bandes de fréquences. S'intéressant juste aux

performances de la séparation, cette ambiguïté est résolue en utilisant les enveloppes des sources originales à estimer. Le vecteur de permutation est donné par :

$$\Pi_\omega = \arg\min_{\Pi_\omega} \sum_{j=1}^{M}\sum_{i=1}^{N} \left\| S_i(t,\omega) - X_j^{(\Pi_\omega(i))}(t,\omega) \right\|$$

Finalement, une estimation temporelle des sources originales est obtenue par la transformée inverse TFCT (avec overlap-add) de $X_j^{(\Pi_\omega(i))}(t,\omega)$.

4.1 Pondération

Pour limiter l'effet des *outliers* sur la performance de la séparation, on se propose d'évaluer l'apport potentiel d'une fonction de pondération $p(t,\omega)$ appliquée à la fonction de coût utilisée par le Kmeans, ici la distance euclidienne entre un point appartenant à un cluster et son centre. La fonction de coût est alors donnée par :

$$\zeta_i = \sum_{\Theta(t,\omega)\in C_i} p(t,\omega) \left\| \Theta(t,\omega) - \bar{c}_i \right\|^2 .$$

Les itérations minimisant la fonction de coût ζ seront donnée par :

$$C_i = \left\{ \Theta(t,\omega) \;\middle|\; i = \arg\min_i \; p(t,\omega) \left\| \Theta(t,\omega) - \bar{c}_i \right\|^2 \right\}$$

$$\bar{c}_i = \sum_{\Theta(t,\omega)\in C_i} p(t,\omega) \left\| \Theta(t,\omega) \right\|^2$$

Deux types de poids p_1 et p_2 sont utilisés et testés. Le premier est énergétique, inspiré du [3], et donné par :

$$p_1(t,\omega) = (\log_{10} U(t,\omega) + ct)$$

avec *ct* la plus petite constante assurant la positivité du poids. Sous l'hypothèse de support disjoint, ce poids permet de favoriser les points TF moins bruités. Le deuxième poids que nous allons tester, déjà utilisé dans [5], représente le degré de confiance qu'une seule source soit présente à un instant TF donné. Ce poids p_2 est défini par le rapport $[\lambda_1-\lambda_2]/[\lambda_1+\lambda_2]$ avec λ_1 , λ_2 les deux valeurs propres ($\lambda_1 \geq \lambda_2$) de la matrice d'auto-corrélation du vecteurs de deux observations:

$$R_{XX}(t,\omega) = \sum_{k=-K}^{k=K} X(t+k,\omega) X^*(t+k,\omega)$$

où *K* représente une fenêtre temporelle. En fait, si dans une bande de fréquence ω, une seule source est présente dans l'intervalle de temps [t-*K*, t+*K*] alors $\lambda_2(t,\omega) \approx 0$ et $p_2(t,\omega) \approx 1$.

5 Résultats et performance

Pour évaluer les performances de la méthode proposée, nous avons utilisé la base de données *Stereo audio source separation evaluation campaign* [6]. Quatre différents types de mélange synthétique (syn) et réel (liv) sont séparés : 4 voix d'hommes (M), 4 voix de femmes (W), 3 corpus de musique sans batteries (NoD) et 3 corpus de musique dont un est fait exclusivement de batterie (WiD). Pour tous ces mélanges la distance entre micro est de 5cm. Les sources ont une durée de 10s et sont échantillonnées à 16kHz. Une fenêtre de Hanning de 2048 échantillons est utilisée avec un recouvrement temporel de 75%. Les différentes positions des sources par rapport aux deux microphones sont données dans le tableau 1.

Type de source	Parole				Music		
	S1	S2	S3	S4	M1	M2	M3
Distance (m)	1,2	1,1	1	0,8	1,1	0,9	1
Angle (deg)	50	-15	-45	15	45	-30	5

Tableau 1. Configuration des mélanges : position des sources (distance, angle) par rapport au couple de microphones

Pour plus d'information sur l'expérimentation, voir [6]. Les performances de la séparation sont estimées en calculant le *Signal to Distortion, Interference* et *Artefact Ratio* (respectivement SDR, SIR et SAR) et le ISR *Image to Signal Ratio*. Pour plus d'information sur l'utilisation et le calcul de telles distorsions, voir [6].

L'utilisation du couple [log(ILD), IPD] comme caractéristique a été validée par comparaison avec [ILD, IPD] et $[|X_1|/U, |X_2|/U,$ IPD] déjà utilisé en [2]. Pour comparer les performances entre les différentes caractéristiques, nous avons calculé la moyenne des performances sur l'ensemble des pondérations en fonction du mélange[1]. Ces performances sont tracées sur les deux figures 2 et 3 représentant respectivement ceux pour le cas d'un mélange de parole et de musique. La figure 2 montre que, pour un mélange de parole, les performances du couple [log(ILD), IPD] sont les meilleures par rapport à celles des deux autres couples pour tout type de mélange. Ces performances sont les meilleures pour tout type de distorsion : SDR, ISR, SIR

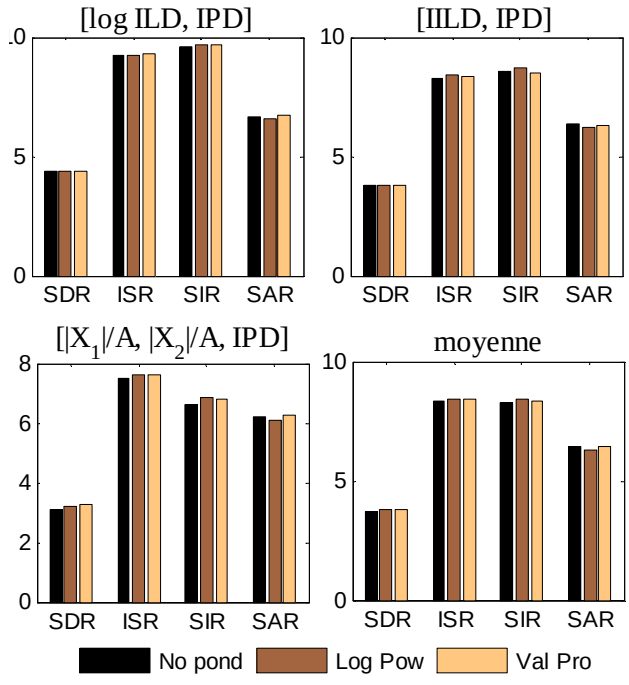

Fig.4 moyenne sur l'ensemble des mélanges pour chaque caractéristique et la moyenne réalisée sur l'ensemble des caractéristiques.

[1] Un mélange est représenté par deux symbole : type de source et type de mélange, exemple : **W_syn = mélange synthétique de 4 voix de femmes, NoD_liv = mélange réel de 3 morceaux de music sans contenant de batterie, …**

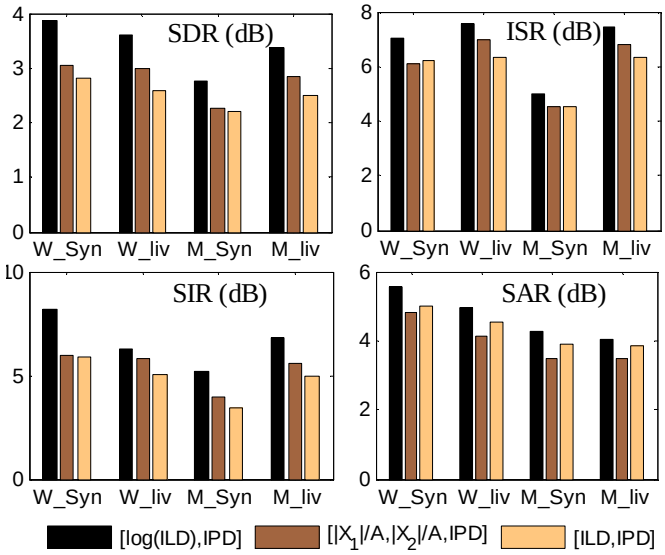

Fig.2 SDR, ISR, SIR et SAR de la moyenne sur l'ensemble de pondération en fonction des caractéristiques; cas des mélanges de parole

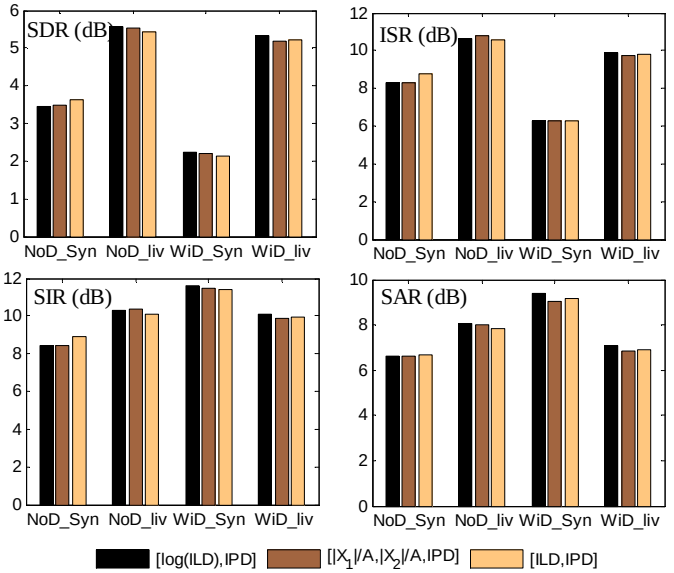

Fig.3 SDR, ISR, SIR et SAR de la moyenne sur l'ensemble de pondération en fonction des caractéristiques; cas des mélanges de musique

et SAR, alors que pour un mélange de musique (figure 2) le couple [log ILD, IPD] admet des performances similaires et parfois supérieures par rapport aux autres types de caractéristiques. En fait, dans le cas des mélanges d'instruments, les directions d'arrivée des mélanges d'instruments, les directions d'arrivée des trois sources sont très différentes. Donc, indépendamment du choix des caractéristiques, les centres du cluster sont correctement estimés et les performances de la séparation sont les mêmes. Au contraire, dans le cas de la parole où quatre sources sont mélangées, l'estimation des centres de cluster va dépendre du choix des caractéristiques puisque les sources vont être plus rapprochées.

D'autre part, pour montrer l'effet de la pondération nous avons testé les deux différents poids cités dans 2.2 ainsi que le cas non pondéré. Ensuite, nous avons calculé la moyenne du SDR dans trois cas : sans pondération (No pond) et avec les poids p_1 (log Pow) et p_2 (Val Pro). La moyenne montre une supériorité négligeable du poids p_2 (3.82 dB) sur p_1 (3.81 dB). Dans tous les cas, la pondération par ces poids reste supérieure au cas non pondéré (3.6 dB). Ces résultats sont tracés dans la figure 4, où est tracé la moyenne des performances sur l'ensemble des mélanges pour chaque type de caractéristique. Comme on peut le voir, indépendamment du choix des caractéristiques, la distorsion moyenne (SDR) pour le cas pondéré est légèrement supérieure au cas non pondéré. Cela est du à une légère amélioration de la distorsion provenant de l'interférence en gardant des performances similaires pour la distorsion provenant de l'artefact.

6 Résumé

Dans ce travail on a proposé d'utiliser un nouveau couple de caractéristiques [log(ILD), IPD] au lieu du couple traditionnel [ILD, IPD] dans la séparation de sources basée sur la classification. Le nouveau couple a montré sa supériorité en termes de distorsion par rapport au couple traditionnel. En plus, l'apport d'un poids dans un tel algorithme a été étudié. On a montré que les algorithmes de classification pondérée améliorent légèrement les performances de séparation. Par contre, les mesures des performances sont réalisées sur des indices objectifs : une étude subjective est nécessaire pour mesurer l'effet psychoacoustique de la pondération et du choix des caractéristiques sur la séparation.

Références

[1] Van Hulle, M., "Clustering approach to square and non-square blind source separation," *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop* , pp.315-323, Aug 1999

[2] S. Araki, H. Sawada and S. Makino (2007). K-means Based Underdetermined Blind Speech Separation. In *Blind Speech Separation*: 243--270. S. Makino, Te-Won Lee and H. Sawada Editors, Springer: New-York.

[3] O. Yilmaz and S. Rickard (2004). Blind separation of speech mixtures via time-frequency masking. IEEE Transactions on Signal Processing 52(7): 1830—1847

[4] D.T Pham, Z. El-Chami, A. Guérin, and C. Servière Modeling the Short Time Fourier Transform Ratio and Application to Underdetermined Audio Source Separation. ICA 2009, Paraty

[5] Simon Arberet, Rémi Gribonval, Frédéric Bimbot, A Robust Method to Count and Locate Audio Sources in a Stereophonic Linear Anechoic Mixture, Proc. IEEE Intl. Conf. Acoust. Speech Signal Process (ICASSP'07)

[6] http://sassec.gforge.inria.fr/