Comment évaluer les algorithmes de séparation de sources audio ?

Emmanuel VINCENT[1], Cédric FÉVOTTE[2], Rémi GRIBONVAL[3],
Laurent BENAROYA[3], Xavier RODET[1], Axel RÖBEL[1], Éric LE CARPENTIER[2], Frédéric BIMBOT[3]

[1]IRCAM, équipe Analyse-Synthèse
1, place Igor Stravinsky, F-75004 PARIS

[2]IRCCyN, équipe ADTS
1, rue de la Noë – BP 92 101, F-44321 NANTES CEDEX 03

[3]IRISA, projet METISS
Campus de Beaulieu, F-35042 RENNES CEDEX

prenom.nom@ircam.fr, prenom.nom@irccyn.ec-nantes.fr, prenom.nom@irisa.fr

Résumé – Dans cet article, nous présentons des applications de la séparation de sources audio et nous proposons quelques idées en vue de constituer des ressources communes pour l'évaluation des algorithmes de séparation. Notre démarche se décompose en trois parties : identifier les tâches typiques à résoudre par les algorithmes, construire des critères de mesure de performance, et collecter des jeux de données appropriés à l'évaluation.

Abstract – In this article, we describe a few applications of audio source separation and we propose some ideas towards the construction of an agreed-upon evaluation framework for audio source separation algorithms. Our work is composed of three steps : identifying the typical tasks to be addressed by the algorithms, designing numerical performance criteria, and collecting relevant datasets.

1 Introduction

La séparation de sources audio (SSA) est un domaine en plein essor que l'on sait maintenant comment aborder grâce à plusieurs modèles, comme l'Analyse en Composantes Indépendantes (ACI) [1] ou les Décompositions Parcimonieuses (DP). Cependant, la SSA ne consiste pas seulement à résoudre un modèle simple, mais à obtenir des résultats pertinents pour l'application visée. Outre la séparation et le débruitage de sources musicales visant une restitution haute qualité, la SSA recouvre des applications aussi diverses que la séparation à des fins d'indexation dans le domaine du multimédia, la reconnaissance de la parole en "cocktail party" ou la localisation de sources pour l'analyse de scènes auditives. La difficulté du problème est influencée par des facteurs qui dépendent de l'application, et il en est de même des critères et signaux tests à utiliser pour évaluer la performance d'un algorithme de SSA. On conçoit donc qu'il est difficile de comparer plusieurs méthodes si la tâche à résoudre n'est pas explicitement précisée.

C'est pourquoi nous pensons qu'il est aujourd'hui opportun de constituer des ressources communes pour l'évaluation des méthodes de SSA. Nous proposons pour cela une démarche en trois parties [2] : identifier les tâches typiques, construire des critères de mesure de performance adaptés à chaque tâche, et collecter des jeux de données pertinents pour l'évaluation. Une réflexion avait été initiée sur le sujet par les auteurs de [3, 4] lors du congrès ICA'99, mais volontairement limitée au cas de la restitution haute qualité des sources.

Dans le cadre de cette démarche, après une présentation des applications de la SSA en Section 2, nous dressons en Section 3 la liste de quelques tâches à résoudre par les algorithmes de SSA et en proposons une typologie. En Section 4, nous nous intéressons aux tâches d'extraction des signaux sources pour lesquelles nous proposons des critères numériques de mesure de performance adaptés. Enfin, nous concluons sur la présentation d'une structure de base de données que nous avons adoptée pour la collecte de données tests et discutons la validité de notre démarche pour d'autres domaines d'applications de la séparation de sources.

Le but de cet article n'est donc pas de fournir un panorama des applications de la SSA, ni d'énumérer les algorithmes les plus performants. Les lecteurs intéressées par ces questions pourront se référer à [5] pour une liste plus complète d'applications et d'algorithmes de SSA assortie de nombreuses références, et consulter les résultats de quelques algorithmes sur des signaux test au sein d'une base de données mise à disposition sur internet [2].

2 Applications de la SSA

Une distinction importante peut être faite entre les applications de la SSA selon que la sortie de l'algorithme est ou non un ensemble de sources extraites destinées à être écoutées. Nous appelons ces deux catégories séparation Orientée Qualité Audio (OQA) et Orientée Extraction de Caractéristiques (OEC).

2.1 Séparation Orientée Qualité Audio

La séparation OQA vise à extraire les sources d'un mélange en vue de les écouter. Elle se divise en deux familles : extraction "un contre tous" et modification de scènes audio.
Quelques notations permettent de clarifier ces termes. Le problème général (éventuellement convolutif) de la SSA $x_i(t) = \sum_{j=1}^{N}(a_{ij} \star s_j)(t) + n_i(t)$ est exprimé grâce au formalisme des matrices de filtres comme $\mathbf{x} = \mathbf{A} \star \mathbf{s} + \mathbf{n}$, où $\mathbf{s}$ est le vecteur des N signaux sources $(s_j)_{j=1}^{N}$, $\mathbf{x}$ est le vecteur des M signaux d'observation $(x_i)_{i=i}^{M}$, $\mathbf{A}$ est la matrice des filtres de mélange, et $\mathbf{n}$ est le vecteur des M signaux de bruit additif $(n_i)_{i=i}^{M}$.

L'extraction "un contre tous" consiste à extraire d'un mélange une seule sorte de son (la source d'intérêt s_j). Parfois il n'est pas nécessaire d'extraire s_j proprement dit : l'estimation de l'image $\mathbf{s^j_{img}} = A \star [0, \ldots, 0, s_j, 0, \ldots, 0]^T$ de s_j sur les capteurs suffit.
Parmi les exemples, on peut citer la restauration de vieux enregistrements musicaux [6], le débruitage et la déréverbération de la voix pour les prothèses auditives ou les communications téléphoniques [7], et l'extraction de sons particuliers dans des extraits musicaux pour la création de musique électronique. La difficulté du problème varie en fonction du nombre de sources et de capteurs, de l'information *a priori* disponible, du niveau de bruit, de la dépendance entre les sources, du type de mélange, etc. Une bonne séparation dans ce cas se mesure par des critères dérivés du Rapport Signal-à-Bruit (RSB), nous discutons ce point en Section 4.

La modification de scènes audio vise à obtenir un nouveau mélange $\mathbf{x_{remix}} = \mathbf{B} \star [f_1(s_1), \ldots, f_n(s_n)]^T$ correspondant à l'application d'un traitement audio adapté f_j (compression de dynamique, ...) à chaque source suivi du mélange des nouvelles pistes à l'aide d'une nouvelle matrice $\mathbf{B}$.
Le *remastering* d'un CD [8], la diffusion sur plusieurs canaux d'enregistrements stéréo [9] et le "karaoké automatique" (suppression de la voix dans une chanson) en sont des exemples. La difficulté dépend des mêmes facteurs que précédemment, ainsi que du niveau de modification introduit par f_j et $\mathbf{B}$. Le problème est généralement plus simple car au sein de $\mathbf{x_{remix}}$ les erreurs d'estimation d'une source peuvent être masquées par la présence des autres sources. L'évaluation du résultat peut se faire en calculant un critère dérivé du RSB entre la scène *remixée* à partir des sources estimées et celle à partir des vraies sources.

2.2 Séparation Orientée Extraction de Caractéristiques

Le but de la séparation OEC est d'extraire d'un mélange complexe des informations d'ordre perceptif et cognitif sur les sources et/ou les paramètres de mélange. Cela n'implique pas forcément deux processus successifs de séparation de sources et de description abstraite : la reconnaissance peut aider la séparation en fournissant des informations contextuelles.
Les appplications OEC concernent principalement l'indexation de bases de données [10] et la création de systèmes d'écoute intelligents [11]. Des exemples de descripteurs utilisés dans ce cadre sont le nom et la partition de chaque instrument dans un extrait musical, le texte prononcé par un locuteur et les caractéristiques de ce locuteur dans un enregistrement de parole, la position spatiale des sources et le lien avec des objets visuels dans un film.
La difficulté du problème varie selon le nombre de sources et de capteurs, la quantité de réverbération, la vitesse de déplacement des sources, le nombre de catégories pour la classification et la robustesse des algorithmes d'extraction de caractéristiques. La qualité d'une description globale prenant en compte plusieurs paramètres peut se mesurer en combinant les qualités d'estimation de chaque paramètre (taux d'erreur, distances,etc.) ou en effectuant une série de tests d'écoute.

3 Typologie proposée des tâches typiques en SSA

Les applications de la SSA sont donc nombreuses et très variées et, pour chacune, des critères appropriés sont nécessaires pour évaluer la performance d'un algorithme. Nous proposons donc de regrouper ces applications en un plus petit nombre de "tâches", afin de permettre la comparaison des diverses méthodes de séparation en identifiant les critères qualitatifs appropriés à chaque tâche.
Le Tableau 1 propose un certain nombre de tâches typiques à accomplir, définies par la nature des entrées et sorties des algorithmes (les observations $\mathbf{x}$ étant une entrée implicite dans tous les cas). Les noms des tâches sont proches de ceux utilisés dans la littérature, et les indéterminations du modèle de SSA (filtrage et permutation des sources) sont prises en compte par une matrice de permutation $\mathbf{P}$ et une matrice diagonale $\mathbf{D}$ arbitraires.

Nous séparons les tâches en deux familles, selon qu'un modèle des sources est disponible ou non. En effet, la définition d'une tâche dépendant de la nature des données en entrée, la différence entre les tâches aveugles et les tâches non aveugles correspondantes apparaît non négligeable.
Par contre, contrairement à [3, 4], nous regroupons dans chaque tâche les mélanges instantanés et convolutifs. Nous pensons que la structure du mélange $\mathbf{A}$ (nombre de sources, taille des filtres), parfois fournie en entrée à l'algorithme, devrait être considérée comme un critère de difficulté (ou une sous-tâche) plutôt qu'une tâche à part entière.
Nous regroupons de même dans certaines tâches non aveugles plusieurs types d'information *a priori* donnant lieu à divers niveaux de difficulté. L'information *a priori* la plus simple est un modèle général des sources (famille de distributions de probabilité, modèle physique, etc), dont les paramètres peuvent éventuellement être appris sur un ensemble d'enregistrements. Parfois, une description des observations est également disponible, où nous entendons par description n'importe quel type d'information (segmentation temporelle, partition musicale, etc) qui précise les paramètres du modèle général en fonction des observations proprement dit.

TAB. 1 – Quelques tâches de SSA, définies par les données en entrée et en sortie

Tâche	Entrée	Sortie
Comptage		$\hat{n}$
Identification aveugle de mélange	structure de $\mathbf{A}$ (*pas toujours*)	$\mathbf{\hat{A}PD}$
Extraction aveugle de sources	structure de $\mathbf{A}$	$\mathbf{PD\hat{s}}$ ou $\{\mathbf{\hat{s}_{img}^j}\}_{j=1}^N$ non ordonnées
Modification aveugle de scène	structure de $\mathbf{A}$, $\mathbf{B}$ et $(f_j)_{j=1}^N$	$\mathbf{\hat{x}_{remix}}$
Détection	modèles des sources $(\mathcal{M}_k)_{k=1}^K$	nombre $\hat{\epsilon}_k$ de sources suivant $\mathcal{M}_k$
Identification/ Représentation	modèle de $\mathbf{s}$	description de $\mathbf{s}$ et $\mathbf{A}$
Extraction de sources	modèle et description de $\mathbf{s}$ et $\mathbf{A}$	$\mathbf{\hat{s}}$ ou $(\mathbf{\hat{s}_{img}^j})_{j=1}^N$ ordonnées
Modification de scène	modèle et description de $\mathbf{s}$ et $\mathbf{A}$, $\mathbf{B}$ et $(f_j)_{j=1}^N$	$\mathbf{\hat{x}_{remix}}$

4 Critères de performance pour l'extraction de sources

Chaque tâche décrite en Section 3 sous-entend un moyen qualitatif de mesurer la performance des algorithmes qui s'y attaquent. Cependant, un cadre d'évaluation rigoureux nécessite la définition de critères objectifs pertinents et partagés [3, 4]. Pour les tâches d'extraction de sources, aveugle ou non, il est possible de mesurer la qualité de chaque source estimée $\hat{s}_m$ en fonction de la vraie source s_m par des critères de type RSB. La difficulté tient à la prise en compte des indéterminations du modèle et au choix de critères adaptés aux applications.

Dans le cas d'un mélange instantané, une possibilité consiste à écrire la source estimée sous la forme $\hat{s}_m = \alpha_m s_m + e_{total}$, où l'erreur totale e_{total} est orthogonale à la vraie source s_m, et à définir un Rapport Signal-à-Distorsion (RSD) $RSD_m = \|\alpha_m s_m\|^2 / \|e_{total}\|^2$ [12].
Lorsque $\mathbf{A}$ est inversible et $\mathbf{n} = \mathbf{0}$, ce RSD peut aussi se calculer sur la matrice de mixage estimée, mais dans les cas plus complexes l'estimation correcte de la matrice de mixage ne suffit plus à retrouver les sources. C'est pourquoi il est important de distinguer les mesures de performance pour les tâches d'extraction de sources et d'identification de mélange.

Pour certaines applications, il peut être utile de séparer dans la qualité du résultat les erreurs d'estimation dues aux interférences des autres sources, à des résidus de bruit additif et à d'éventuels artefacts dus à l'algorithme. Il est particulièrement important de mesurer la quantité d'artefacts introduite dans le cas des mélanges sous-déterminés, car les algorithmes de séparation donnent souvent dans ce cas une teinte artificielle aux sources estimées, les rendant inutilisables pour une application musicale par exemple [12].
Une mesure possible de ces différentes contributions consiste à décomposer l'erreur totale sous la forme de termes orthogonaux $e_{total} = \sum_{l=1}^N \alpha_l s_l + \sum_{k=1}^M \beta_k n_k + e_{artef}$ et à calculer pour chaque terme un rapport d'énergies de type RSD.
Une boîte à outils MATLAB implémentant ces critères est disponible sur internet [2].

Enfin, puisque les sources sont destinées à être écoutées, il est possible de modifier de façon mineure ces critères pour tenir compte des spécificités de l'audition, telles que les lois de masquage auditif spectral et temporel.

5 Conclusion et perspectives

Ce travail donne un aperçu des applications de la SSA et propose une démarche à suivre pour la construction d'un cadre d'évaluation commun des algorithmes de SSA. Celle-ci consiste en trois points : dresser un tableau des tâches typiques à résoudre ; pour chaque tâche, construire des critères de mesure de performance ; rassembler des données test structurées pour l'évaluation.

Nous avons abordé les deux premiers points dans les Sections 3 et 4, en proposant une typologie de quelques tâches identifiées dans la littérature et des critères adaptés aux tâches d'extraction de sources. Nous encourageons les chercheurs à engager une discussion à partir de nos propositions sur notre liste de discussion [2], de façon à aboutir à un cadre d'évaluation partagé par la communauté.
En ce qui concerne le troisième point, une base de données structurée et plusieurs signaux test ont été mis à disposition sur internet [2]. Ordonnée par tâches et par niveaux de difficulté (ou sous-tâches), cette base de données permet de consulter indépendamment les informations sur un jeu de données, sur un algorithme, et sur ses performances. Le succès de cette initiative dépend des contributions de la communauté SSA, aussi nous encourageons nos collègues à soumettre leurs jeux de données et les résultats de leurs algorithmes.
Enfin, en vue de fournir une liste objective des algorithmes les plus performants pour certaines applications, nous projetons d'organiser un concours de séparation de sources audio dès que possible.

Nous pensons que la démarche que nous avons entreprise en vue de l'évaluation des méthodes de séparation de sources appliquées aux signaux audio pourrait être appliquée avec succès à d'autres domaines comme les applications biomédicales, l'imagerie hyperspectrale, la classification multimédia, la compression, etc. L'approche consistant à identifier des tâches ty-

piques à résoudre préablablement à la collecte de jeux de données permet de structurer ces jeux de données de façon adaptée au problème et d'évaluer tous les algorithmes par les mêmes critères.
D'autre part, une typologie des tâches peut permettre d'identifier certaines tâches moins étudiées que les autres, fournissant ainsi de nouveaux buts de recherche. C'est le cas en audio pour les tâches de modification de scènes qui, malgré leurs critères de performance moins restrictifs, semblent moins étudiées que les tâches d'extraction de sources par exemple.

6 Remerciements

Ce travail a été réalisé dans le cadre de l'Action Jeunes Chercheurs du GdR ISIS "Ressources pour la séparation de signaux audiophoniques". La démarche décrite dans cet article constitue l'objectif de l'Action, et ses résultats, y compris les articles [5, 12], sont disponibles sur le site internet [2].

Références

[1] J.-F. Cardoso, "Blind source separation : statistical principles," in *IEEE Proc.*, 1998, vol. 90, pp. 2009–2026.

[2] Action Jeunes Chercheurs du GDR ISIS (CNRS), "Ressources pour la séparation de signaux audiophoniques," http ://www.ircam.fr/anasyn/ISIS/.

[3] D. Schobben, K. Torkkola, and P. Smaragdis, "Evaluation of blind signal separation methods," in *Proc. Int. Workshop on ICA and BSS (ICA'99)*, 1999, pp. 261–266.

[4] R.H. Lambert, "Difficulty measures and figures of merit for source separation," in *Proc. Int. Workshop on ICA and BSS (ICA'99)*, 1999, pp. 133–138.

[5] E. Vincent, X. Rodet, A. Röbel, C. Févotte, R. Gribonval, L. Benaroya, and F. Bimbot, "A tentative typology of audio source separation tasks," in *Proc. Int. Worshop on ICA and BSS (ICA'03)*, 2003, pp. 715–720.

[6] Olivier Cappé, *Techniques de réduction de bruit pour la restauration d'enregistrements musicaux*, Ph.D. thesis, Télécom Paris, 1993.

[7] H. Attias, J.C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," in *Proc. Int. Workshop on Neural Information Processing Systems (NIPS'01)*, 2001.

[8] R. Radke and S. Rickard, "Audio interpolation," in *Proc. Int. Conf. on Virtual, Synth. and Entertainment Audio*, 2002.

[9] R. Dressler, "Dolby Surround Pro Logic II decoder : Principles of operation," Dolby Laboratories Information, 2000.

[10] M. Casey, "Generalized sound classification and similarity in MPEG-7," *Organized Sound*, vol. 6, no. 2, 2002.

[11] D.P.W. Ellis, *Prediction-driven computational auditory scene analysis*, Ph.D. thesis, MIT, 1996.

[12] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte, "Proposals for performance measurement in source separation," in *Proc. Int. Workshop on ICA and BSS (ICA'03)*, 2003, pp. 763–768.