Outils d'évaluation des algorithmes de séparation de sources

Introduction

L'évaluation ou la comparaison des performances des algorithmes de séparation de sources a pour longtemps été une tâche difficile. En effet, il y avait un manque de méthodes de mesure des performances et celles qui existaient présentaient de nombreuses limitations [88]. Prenons l'exemple de la différence inter-symbole entre la source estimée et la vraie source ou la comparaison directe de la source estimée et la source originale : ces deux méthodes ne considèrent que le problème de permutation et d'échelle et ne prennent pas en compte les distorsions autorisées par l'application, de plus ces méthodes ne présentent qu'un seul type de critère de performance englobant toutes les erreurs.

En 2006, Vincent *et al.* [88] proposent une méthode *objective* d'évaluation des performances des algorithmes de séparation de sources basée sur quatre types de distorsions autorisées, allant de la distorsion la plus simple du gain invariant temporellement à la plus complexe des filtres variables dans le temps. Dans chaque cas, ils décomposent la source estimée en la vraie source plus des termes d'erreurs correspondant aux interférences, au bruit additif et aux artéfacts introduits par l'algorithme de séparation. Ensuite, ils dérivent des mesures de performance globale en utilisant le rapport d'énergie et une mesure de performance séparée pour chaque terme d'erreur. Ces mesures de performance sont implémentées dans une boite à outils MATLAB sous le nom de BSS_EVAL distribuée sous la licence publique générale GNU.

Cependant, les algorithmes de décomposition de la distorsion proposés jusqu'à maintenant ne donnent pas toujours les composantes attendues et on peut mettre

en doute la capacité des rapports des énergies à correspondre aux scores *subjectifs* puisque des phénomènes audio tels que la perception de la "loudness" et le masquage spectral ne sont pas pris en compte. Emiya *et al.* [30] proposent d'évaluer *la qualité perçue* des signaux sources estimées dans le contexte de la séparation de sources audio. Ces signaux peuvent impliquer un ou plusieurs types de distorsions, y compris les distorsions de la source cible, les interférences provenant des autres sources ou les artéfacts du bruit musical. Ils proposent un protocole de test subjectif pour évaluer la qualité perçue par rapport à chaque type de distorsion, et une famille de mesures objectives afin de prédire ces scores subjectifs. Cette série de mesures objectives est basée sur la décomposition de l'erreur estimée en différentes composantes de distorsions et l'utilisation de l'outil de mesure perceptuelle PEMO-Q. Cette méthode d'évaluation réalisée avec MATLAB est disponible en téléchargement libre sous le nom de PEASS (Perceptual Evaluation methods for Audio Source Separation).

Pour l'évaluation de nos algorithmes de séparation de sources, nous utilisons les deux méthodes BSS_EVAL et PEASS que nous détaillerons dans la suite, elles ont été utilisées aussi dans la campagne d'évaluation de séparation des signaux SISEC (Signal Separation Evaluation Campaign). Pour ces méthodes d'évaluation, la connaissance des sources originales est nécessaire. De plus, ces mesures ne tiennent pas compte du problème de permutation de la séparation aveugle de sources audio : si nécessaire, la source estimée peut être comparée à toutes les sources originales et la "vraie source" peut être sélectionnée comme celle donnant les meilleurs résultats.

8.1 Évaluation objective des performances de séparation (BSS_EVAL)

Les mesures de performance sont calculées pour chaque source estimée en la comparant avec la source originale. Le calcul du critère se fait en deux étapes : la décomposition de la source estimée et le calcul des différents rapports d'énergies qui donnent les mesures de performances globales. On note le vecteur temporel relatif à une source j et de longueur T : $\mathbf{x}_j = [x_j(1), \ldots, x_j(T)]^T$.

8.1.1 Décomposition de la source estimée

La première étape consiste à décomposer la source estimée $\mathbf{y}_j$ comme suit :

$$\mathbf{y}_j = \mathbf{s}_j^{\text{target}} + \mathbf{e}_j^{\text{interf}} + \mathbf{e}_j^{\text{noise}} + \mathbf{e}_j^{\text{artif}} \tag{8.1}$$

où $\mathbf{s}_j^{\text{target}} = f(\mathbf{s}_j)$ est une version de la vraie source $\mathbf{s}_j$ modifiée par une distorsion autorisée $f \in \mathcal{F}$, dans notre cas la distorsion consiste en un filtrage temporel invariant, et $\mathbf{e}_j^{\text{interf}}$, $\mathbf{e}_j^{\text{noise}}$ et $\mathbf{e}_j^{\text{artif}}$ sont respectivement les termes d'erreur relatifs aux *interférences*, au *bruit* et aux *artéfacts*. Ces quatre termes doivent représenter la partie de $\mathbf{y}_j$ perçue comme venant de la source désirée $\mathbf{s}_j$, des sources non désirées $(\mathbf{s}_{j'})_{j' \neq j}$, du bruit des capteurs $(\mathbf{n}_i)_{1 \leq i \leq M}$ et d'autres causes comme les distorsions non autorisés et/ou le bruit musical). Quand une distorsion par filtrage temporel invariant est autorisée, $\mathbf{s}_j^{\text{target}}$ est une version filtrée de $\mathbf{s}_j$ telle que $s_j^{\text{target}}(t) = \sum_{l=0}^{L-1} h(l) s_j(t-l)$. Par conséquent, $\mathbf{s}_j^{\text{target}}$ appartient au sous-espace engendré par des versions décalées de $\mathbf{s}_j$, ce qui implique que $\mathbf{s}_j^{\text{target}}$ peut être définie en projetant $\mathbf{y}_j$ sur ce sous-espace.

Soit $\Pi\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ le projecteur orthogonal sur le sous-espace engendré par les vecteurs $\mathbf{x}_1, \ldots, \mathbf{x}_k$. Ce projecteur est une matrice $T \times T$ où T est la longueur de ces vecteurs. On note $\mathbf{s}_j^l$ et $\mathbf{n}_i^l$ le signal source $\mathbf{s}_j$ et le signal bruit $\mathbf{n}_i$ décalés par l échantillons, ce qui donne : $s_j^l(t) = s_j(t-l)$ et $n_i^l(t) = n_i(t-l)$. Soient les trois projecteurs suivant :

$$P_{\mathbf{s}_j} = \Pi\left\{\left(\mathbf{s}_j^l\right)_{0 \leq l \leq L-1}\right\} \tag{8.2}$$

$$P_{\mathbf{s}} = \Pi\left\{\left(\mathbf{s}_{j'}^l\right)_{0 \leq j' \leq N, 0 \leq l \leq L-1}\right\} \tag{8.3}$$

$$P_{\mathbf{s},\mathbf{n}} = \Pi\left\{\left\{\left(\mathbf{s}_{j'}^l\right)_{0 \leq j' \leq N}, \left(\mathbf{n}_i^l\right)_{0 \leq i' \leq M}\right\}_{0 \leq l \leq L-1}\right\} \tag{8.4}$$

La source estimée y_j se décompose alors comme la somme des quatre termes :

$$\mathbf{s}_j^{\text{target}} = P_{\mathbf{s}_j}\mathbf{y}_j \tag{8.5}$$

$$\mathbf{e}_j^{\text{interf}} = P_{\mathbf{s}}\mathbf{y}_j - P_{\mathbf{s}_j}\mathbf{y}_j \tag{8.6}$$

$$\mathbf{e}_j^{\text{noise}} = P_{\mathbf{s},\mathbf{n}}\mathbf{y}_j - P_{\mathbf{s}}\mathbf{y}_j \tag{8.7}$$

$$\mathbf{e}_j^{\text{artif}} = \mathbf{y}_j - P_{\boldsymbol{s},\mathbf{n}}\mathbf{y}_j \tag{8.8}$$

8.1.2 Mesures des performances globales

Dans la deuxième étape, les rapports des énergies en décibel sont calculés afin d'évaluer le taux de participation de chacun de ces quatre termes à la source estimée $\mathbf{y}_j$, soit dans toute la durée du signal soit sur des trames locales :

– le rapport source-à-distorsion SDR (Signal-to-Distorsion Ratio) :

$$\mathrm{SDR} = 10\log_{10}\frac{\left\|\mathbf{s}_j^{\mathrm{target}}\right\|^2}{\left\|\mathbf{e}_j^{\mathrm{interf}}+\mathbf{e}_j^{\mathrm{noise}}+\mathbf{e}_j^{\mathrm{artif}}\right\|^2} \tag{8.9}$$

– le rapport source-à-interférences SIR (Sources-to-Interferences Ratio) :

$$\mathrm{SIR} = 10\log_{10}\frac{\left\|\mathbf{s}_j^{\mathrm{target}}\right\|^2}{\left\|\mathbf{e}_j^{\mathrm{interf}}\right\|^2} \tag{8.10}$$

– le rapport sources-sur-bruit SNR (Sources-to-Noise Ratio) :

$$\mathrm{SNR} = 10\log_{10}\frac{\left\|\mathbf{s}_j^{\mathrm{target}}+\mathbf{e}_j^{\mathrm{interf}}\right\|^2}{\left\|\mathbf{e}_j^{\mathrm{noise}}\right\|^2} \tag{8.11}$$

– le rapport sources-à-artéfacts SAR (Sources-to-Artifacts Ratio) :

$$\mathrm{SAR} = 10\log_{10}\frac{\left\|\mathbf{s}_j^{\mathrm{target}}+\mathbf{e}_j^{\mathrm{interf}}+\mathbf{e}_j^{\mathrm{noise}}\right\|^2}{\left\|\mathbf{e}_j^{\mathrm{artif}}\right\|^2} \tag{8.12}$$

Pour l'évaluation de nos algorithmes, nous évaluons le SDR, le SIR et le SAR. Le SNR ne sera pas pris en compte car nous considérons un bruit diffus spatialement décorrélé dont l'énergie est supposée être negligeable par rapport à celle des sources. Si le bruit est ponctuel, il sera considéré comme une source sonore.

8.2 Évaluation perceptuelle des performances de séparation (PEASS)

Emiya *et al.* [30] proposent une famille de mesures objectives basée sur une décomposition de la distorsion différente de celle de BSS_EVAL et valident la capacité de ces mesures objectives à prédire les scores subjectifs multi-critères obtenus lors d'un protocole de test dédié à évaluer la qualité de la séparation de sources audio. Les critères subjectifs considérés dans l'évaluation perceptuelle de la qualité de la

séparation de sources audio sont :

1. score global : la *qualité globale* par rapport à la référence pour chaque signal test (chaque source estimée) ;
2. préservation de la source cible : la qualité en termes de *préservation de la source cible* dans chaque signal test ;
3. suppression des autres sources : la qualité en termes de *suppression des autres sources* dans chaque signal test ;
4. addition de bruit artificiel : la qualité en termes d'*absence de bruit artificiel additionnel* dans chaque signal test.

Dans ce qui suit, nous présenterons la famille des quatre mesures objectives dont le but est de prédire les scores subjectifs que nous venons de citer. L'approche proposée par [30] consiste à diviser le signal de distorsion en une somme de composantes liées à la *distorsion cible*, aux *interférences* et aux *artéfacts*, à évaluer leur saillance perceptuelle en utilisant des métriques de nature auditives et à combiner les attributs résultants.

8.2.1 Modélisation et estimation des composantes de distorsions

Le distorsion entre la source estimée $y_j(t)$ et la source cible $s_j(t)$ est décomposée en la somme d'une composante de distorsion cible $e_j^{\text{target}}(t)$, une composante d'interférence $e_j^{\text{interf}}(t)$ et une composante d'artéfact $e_j^{\text{artif}}(t)$ comme suit :

$$y_j(t) - s_j(t) = e_j^{\text{target}}(t) + e_j^{\text{interf}}(t) + e_j^{\text{artif}}(t) \tag{8.13}$$

Pour accomplir cette décomposition, on doit spécifier comment la distorsion cible et les composantes d'interférences sont liées aux sources originales. Cependant, la manière dont le système auditif sépare les flux associés à ces composantes demeure inconnue. Une approche consiste à supposer que ces composantes sont des versions linéairement distordues des sources réelles et cette distorsion est modélisée par un filtre à réponse impulsionnelle finie (FIR) multicanal invariant dans le temps. Cette hypothèse a été prise en compte notamment dans BSS_EVAL. Cependant, ces composantes de distorsion ne correspondent pas toujours à celles perçues par l'oreille humaine. Ceci est dû en particulier au modèle invariant dans le temps qui ne correspond pas à la nature variable dans le temps des distorsions rencontrées et à la résolution fréquentielle constante des filtres RIF qui ne correspond pas à celle de

l'oreille. Une décomposition à temps-variable a été proposée par [88]. Cependant, à cause de son grand coût de calcul, elle est restreinte en pratique aux filtres avec une basse résolution spatiale et temporelle, et par conséquent, elle n'a pas amélioré les résultats. La décomposition proposée par Emiya *et al.* [30] a pour but de résoudre ces problèmes et donne des composantes de distorsion perceptuellement plus pertinentes s'approchant de la résolution temps-fréquence auditive, grâce à l'utilisation de banc de filtre. Ceci se fait en trois étapes :

1. analyse temps-fréquence : la source estimée $y_j(t)$ et les sources originales $s_i(t)$, $1 \leq i, j \leq N$ sont partitionnés en temps et en fréquence par un banc de filtres type gammatone [1] en des signaux $y_{ib}(t)$ et $s_{ib}(t)$ indexés par b. Dans chaque sous bande, après une étape de sous-échantillonnage, ces signaux sont ensuite fenêtrés en des trames recouvrantes indexées par u : $y_{jbu} = w_a(t)y_{jb}(t - uN)$ et $s^{\tau}_{ibu} = w_a(t)s_{ib}(t - uN - \tau)$, où w_a est la fenêtre d'analyse, N est le pas d'avancement et $s_{ib}(t-\tau)$ est la version décalée de la vraie source $s_{ib}(t)$, $-L/2 \leq \tau \leq L/2$;

2. décomposition par moindres carrés jointe : à cause de la large bande passante des filtres gammatone, les composantes de distorsion sont estimées par un filtrage additionnel en chaque sous-bande et trame temporelle ; ces composantes sont définies par un filtrage RIF multicanal invariant dans le temps des sources cibles et des sources interférentes ; les coefficients de ces filtres sont estimés par une projection des moindres carrés de la distorsion $y_{jbu}(t) - s^0_{jbu}(t)$ sur le sous-espace engendré par les versions décalées des signaux sources $s^{\tau}_{ibu}(t)$, $1 \leq i \leq N$ et $-L/2 \leq \tau \leq L/2$; les composantes de distorsion s'écrivent :

$$e^{\text{target}}_{jbu}(t) = \sum_{\tau=-L/2}^{L/2} \alpha_{jbu,j}(\tau)\, s^{\tau}_{jbu}(t) \tag{8.14}$$

$$e^{\text{interf}}_{jbu}(t) = \sum_{i \neq j} \sum_{\tau=-L/2}^{L/2} \alpha_{jbu,l}(\tau)\, s^{\tau}_{ibu}(t) \tag{8.15}$$

$$e^{\text{artif}}_{jbu}(t) = y_{jbu}(t) - s^0_{jbu}(t) - e^{\text{target}}_j(t) - e^{\text{interf}}_j \tag{8.16}$$

3. reconstruction des signaux temporels : les signaux sont reconstruit par overlap-add et inversion du banc de filtre $e^{\text{target}}_j(t)$, $e^{\text{interf}}_j(t)$ et $e^{\text{artif}}_j(t)$.

1. Un banc de filtres gammatone est un banc de filtres qui modélise la non-linéarité et la variance temporelle du système auditif.

8.2.2 Mesures objectives

Étant données des composantes de distorsion comme celles calculées à la sous-section précédente ou comme celles utilisées dans BSS_EVAL, le but est d'évaluer la similarité entre la source estimée et la source originale en s'appuyant sur les quatre critères subjectifs cités dans l'introduction de cette section. Ceci est fait en deux étapes.

La première étape consiste à évaluer perceptuellement chaque composante de distorsion en utilisant la métrique de similarité perceptuelle PSM (Perceptual Similarity Measure) fournit par le modèle auditif PEMO-Q [47].

La métrique de similarité perceptuelle donne les attributs suivant :

$$q_j^{\text{overall}} = \text{PSM}\left(\mathbf{y}_j, \mathbf{s}_j\right) \tag{8.17}$$

$$q_j^{\text{target}} = \text{PSM}\left(\mathbf{y}_j, \mathbf{y}_j - e_j^{\text{target}}\right) \tag{8.18}$$

$$q_j^{\text{interf}} = \text{PSM}\left(\mathbf{y}_j, \mathbf{y}_j - e_j^{\text{interf}}\right) \tag{8.19}$$

$$q_j^{\text{artif}} = \text{PSM}\left(\mathbf{y}_j, \mathbf{y}_j - e_j^{\text{artif}}\right) \tag{8.20}$$

Dans la deuxième étape, une combinaison non linéaire de ces métriques donne les quatre mesures objectives suivantes [30] :

- Le score perceptuel global OPS (Overall Perceptual Score) qui évalue la qualité globale de la source estimée par rapport à la source originale.
- Le score perceptuel relatif à la cible TPS (Target-related Perceptual Score) qui traduit la qualité en termes de préservation de la source cible dans la source estimée.
- Le score perceptuel relatif aux interférences IPS (Interference-related Perceptual Score) qui évalue la qualité en termes de suppression des autres sources dans la source estimée.
- Le score perceptuel relatif aux artéfacts APS (Artifacts-related Perceptual Score) qui estime la qualité en termes d'absence de bruit artificiel additionnel dans chaque source estimée.

Conclusion

Dans ce chapitre, nous avons introduit les outils d'évaluation des performances des algorithmes de séparation de sources que nous avons utilisés pour évaluer nos algorithmes. Dans un premier temps, nous avons présenté la boite à outils BSS_EVAL

[88] contenant des mesures objectives de la qualité des sources estimées : le rapport source-à-interférence (SIR), le rapport source-à-distorsion (SDR), le rapport sources-à-artéfacts (SAR) et le rapport sources-sur-bruit (SNR). Parfois, ces mesures objectives sont insuffisantes pour évaluer la qualité perçue des sources estimées. Par conséquent, nous avons introduit dans un deuxième temps la boite à outils PEASS [30] qui propose une évaluation perceptuelle des performances des algorithmes de séparation de sources selon les scores suivants : le score perceptuel global (OPS), le score perceptuel relatif à la cible (TPS), le score perceptuel relatif aux interférences (IPS) et le score perceptuel relatif aux artéfacts (APS).