Audio Signal Analysis, Indexing and Transformation, Master MVA, 2017/2018

Compte-Rendu de Lecture: Méthode DUET pour la séparation de sources

Raphaël Romero

raphael.romero@telecom-paristech.fr

1 Apports principaux de l'article

En séparation de sources, on cherche à estimer m signaux "sources" $s_1, ..., s_m$ à partir de n mélanges de ces sources $x_1, ..., x_n$, que l'on peut observer (à l'aide d'un microphone par exemple). En général on ne peut pas résoudre ce problème sans faire d'hypothèse sur les signaux s_j. Typiquement des hypothèses statistiques ou algébriques sur les signaux peuvent être utilisées.

Dans ce premier type de méthodes, on suppose que les signaux observés peuvent s'écrire comme des combinaisons (mélanges) linéaires des sources et par conséquent à un instant donné le vecteur des n signaux observés s'écrit comme le produit matriciel d'une matrice (matrice de mélange) avec les m signaux sources. Sous ce modèle estimer les sources revient à estimer l'inverse de la matrice de mélange. Cependant lorsque cette matrice n'est pas inversible de telles méthodes ne peuvent pas être utilisées. C'est notamment le cas dès que le nombre de sources est supérieur au nombre de capteurs.

Certaines méthodes existant avant DUET proposent des solutions valables lorsque le nombre de sources est légèrement supérieur au nombre de capteurs. Au contraire dans cet article, les auteurs présentent une nouvelle méthode se permettant de traiter le cas où l'on a deux capteurs et potentiellement un grand nombre de sources. En choisissant sans perte de généralité de considérer des sources au niveau du premier capteur, les deux signaux perçus s'écrivent comme :

$$\begin{cases} x_1(t) = \sum_{j=1}^{m} s_j(t) \\ x_2(t) = \sum_{j=1}^{m} a_j s_j(t - \delta_j) \end{cases}$$

La méthode DUET est basée sur le fait que les représentations temps-fréquence des signaux émis par différentes sources ont en général des supports disjoints : les signaux sont dits *W-disjoints orthogonaux* pour une certaine fonction de fenêtrage W. En utilisant cette observation la séparation de sources revient à partitioner le plan temps-fréquence en régions caractéristiques de chaque source. Pour définir ces régions on utilise le fait que, dans le

modèle de mixture anéchoïque, chaque sources$_j$ est caractérisée par son atténuation a_j (ou bien son atténuation symétrique α_j) et son retard δ_j. En chaque point (τ, ω) la méthode DUET permet de calculer une estimation $(\tilde{\alpha}(\tau,\omega), \tilde{\delta}(\tau,\omega)$ de l'atténuation et du retard. Une région caractérisant une source donnée j peut alors être définie comme l'ensemble des points (τ, ω) conduisant à une estimation $(\tilde{\alpha}(\tau,\omega), \tilde{\delta}(\tau,\omega))$ "proche" des paramètres a_j, δ_j caractérisant la source j.

L'outil utilisé pour les estimations est la transformée de Fourier à court terme (TFCT). Pour un point temps-fréquence (τ, ω) et une fonction de fenêtrage w cette transformée exprime à partir du signal s comme :

$$\hat{s}(\tau,\omega) = \int w(t-\tau)s(t)e^{-2i\pi\omega t}dt$$

Grâce à l'hypothèse d'orthogonalité on sait qu'en un point donné du plan T-F, les TFCT deux signaux perçus sont égales à un facteur multiplicatif complexe près. Ce facteur peut être estimé par le quotient $\frac{\hat{x_2}}{\hat{x_1}}$. Son module et son argument sont alors des estimations locales de l'atténuation et du retard respectivement. Pour définir les régions caractéristiques pour chaque source, il reste à définir les valeurs $(\tilde{\alpha}_j, \tilde{\delta}_j)$ caractérisant chaque source. Pour cela, en introduisant des variables d'erreurs sur le modèle de mélange et en s'inspirant de la forme des estimateurs de maximum de vraisemblance, on aimerait en théorie calculer une espérance des estimations locales par rapport à une distribution f_j définie sur le plan T-F par $f_j(\tau,\omega) \propto |x_1(\tau,\omega)x_2(\tau,\omega)|^p\omega^q \mathbf{1}_{\Omega_j}(\tau,\omega)|$ où Ω_j désigne le support temps-fréquence du signal émis par la j-ième source. Cependant ce support est justement ce que l'on cherche à déterminer dans le problème, donc on n'a pas accès à ces distributions. Pour palier à ce problème on utilise l'histogramme des valeurs de α, δ, qui lui peut être calculé en se donnant des résolutions Δ_α et Δ_δ par

$$H(\alpha,\delta) = \int |x_1(\tau,\omega)x_2(\tau,\omega)|^p\omega^q \mathbf{1}_{\{|\tilde{\alpha}(\tau,\omega)-\alpha|<\Delta_\alpha, |\tilde{\delta}(\tau,\omega)-\delta)|<\Delta_\delta\}} d\tau d\omega$$

Les atténuations et retards caractéristiques des sources sont alors obtenus en prélevant les valeurs produisant les pics de cet histogramme. Différentes versions de la méthode DUET permettent de traiter des cas pathologiques, par exemple lorsque les retards sont trop grands, ou bien lorsque l'hypothèse d'orthogonalité n'est pas totalement satisfaite. Cependant le principe de la méthode reste le même.

2 Avantages/Inconvénients de la méthode

On voit que la méthode DUET est avantageuse car elle est réalisable à l'aide de seulement deux capteurs alors que d'autres méthodes supposent qu'il y ait au moins autant de capteurs que de sources. En effet ici le choix du nombre de source n'est nécessaire qu'à l'issue du tracé de l'histogramme, lorsque l'on souhaite isoler les sources donc convertir les régions T-F estimées en signaux temporels. Mais la méthode pourrait par exemple être utilisée pour estimer le nombre de sources à partir de l'histogramme, ou bien dans le cas où la source n'est plus ponctuelle mais continue, obtenir des informations sur sa géométrie.

Du point de vue du temps de calcul, la méthode ne met en jeu que des intégrales et des transformation de Fourier, ce qui la rend plus efficace que d'autres méthodes (NMF,ICA,...) qui souvent passent par des algorithmes d'optimisation pouvant être lents.

En revanche l'hypothèse de signaux sources "W-disjoints orthogonaux" est basée sur une observation valable sur les signaux de parole, qui est que deux personnes ont très peu de chance de parler au même moment à la même fréquence. Même en quantifiant cette orthogonalité disjointe, le modèle n'est donc que très peu applicable à des signaux différents comme les signaux musicaux par exemple où les recoupements temps-fréquence de signaux émis par différentes sources ont plus de chances de se produire. Ainsi le modèle est peut-être moins "universel" que l'ICA ou la NMF par exemple, au sens ou il s'applique à une plus faible variété de signaux.

3 Point de vue personnel

La méthode DUET offre une bonne solution du problème "cocktail party", puisque dans ce cas elle produit une visualisation en 3 dimensions des sources grâce à l'histogramme et permet non seulement de les isoler mais aussi de les compter.

Une possibilité d'amélioration aurait pu être une discussion sur la forme de la fenêtre utilisée pour la TFCT, et son impact sur le résultat. Un autre point qui aurait pu être approfondi est l'hypothèse d'anéchoïcité, qui est choisie dans l'article de manière empirique mais dont il aurait pu être utile de voir les limites. Finalement une dernière possibilité aurait été d'étudier l'applicabilité de la méthode à d'autres types de données lorsque l'on sait que la séparation de sources peut être employée dans des domaines variés allant de la médecine à la finance.