**Reading note TSIA 206 Baptiste Moalic, 2022**

Article sélectionné: Ozerov, A. and Févotte, C. “Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation”, *IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 3*, March 2010

Cet article de recherche d’Alexey Ozerov et Cédric Févotte traite de la séparation de sources audio dans le cadre multicanaux et de mélange convolutionnel. Pour cela, ils ont inspiré leurs algorithmes des méthodes utilisées en factorisation par matrices non-négatives (NMF). Cette démarche est intéressante puisque la NMF n’est normalement utilisée que pour des données sur un unique canal. On peut utiliser des tenseurs plutôt que des matrices par exemple, mais cela ne sera fonctionnel que dans le cadre d’un mélange linéaire instantané. L’approche proposée par les deux chercheurs permet de concilier ces deux problèmes en s’inspirant des méthodes NMF, et en utilisant des méthodes statistiques afin de traiter les informations mutuelles entre chaque canal, permettant ainsi de désintriquer les données convoluées et de traiter le cas multicanaux.

Tout d’abord, pour traiter le cas du mélange convolutif, les chercheurs ont travaillé sur comment modéliser au mieux les sources. Les sources sont modélisées en tant que sommes de réalisations de différentes composantes gaussiennes latentes (5). Ainsi, une telle structure permet d’aboutir à un modèle linéaire (10), tout en conservant l’idée de convolution (car des composantes peuvent se retrouver à différents endroits dans la partition). Retrouver un modèle linéaire justifie l’utilisation des méthodes NMF.

Ensuite, le problème des multiples canaux est adressé. La méthode choisie est celle de l’espérance-maximisation (EM). L’algorithme EM est ici justifié car nous avons bien un modèle probabiliste (gaussien). Le modèle probabiliste nous permet de retrouver des outils comme la log-vraisemblance et l’information mutuelle, qui sont des critères dont on sait que la maximisation permet de comparer différents modèles (en fait c’est une approche de Machine Learning). Comme son nom l’indique, l’algorithme EM consiste en l’estimation de différents paramètres statistiques (espérance, covariance), puis en la maximisation des paramètres de notre modèle à partir de ce que l’on a estimé (matrice de mélange, coefficients de W et H de la NMF). Il nous suffit ensuite de calculer le critère probabiliste (12) pour obtenir un score de notre modèle. Le point important à noter ici est qu’il s’agit de la log-vraisemblance des données complètes, et non pas de chaque canal indépendant, ce qui fait que les informations de chaque canal vont aider les autres canaux lors de la séparation. La reconstruction se fait par filtre de Weiner.

Les chercheurs ont utilisé un deuxième algorithme, complémentaire au premier puisqu’il va chercher à maximiser la somme des log-vraisemblances de chaque canal. Ce deuxième algorithme va s’assurer que l’information partagée ne se propage pas trop, et que chaque canal garde son information intacte. On retrouve l’influence des méthodes de NMF puisque ce deuxième algorithme se base sur des mises à jour multiplicatives (MU) des matrices, comme c’est le cas par exemple de l’algorithme MUR en NMF.

Les chercheurs se sont inspirés des méthodes de NMF qui fonctionnent déjà très bien, et qui sont très bien documentées. Ainsi, la complexité de compréhension des algorithmes est plutôt faible. De plus, les algorithmes EM et MU sont majoritairement des algorithmes de multiplication ou de division, il y a très peu d’inversions de matrices, comme expliqué en III.B.4, les inversions sont faites sur des matrices au plus 2x2, ce qui rend les algorithmes peu longs à s’exécuter. Si ces algorithmes sont efficaces, c’est parce que la majorité du travail a été fait en amont : c’est la modélisation des sources qui leur a permis d’employer des méthodes bien connues. Ainsi les algorithmes sont très polyvalents, et peuvent être utilisés aussi bien en apprentissage supervisé qu’en séparation à l’aveugle.

Cependant, la contrepartie d’utiliser des modèles statistiques, c’est de se retrouver avec les problèmes bien connus de Machine Learning : les auteurs font référence à une possibilité d’overfitting. De surcroît, une grande partie de l’article fait en fait référence à des problèmes d’optimisation, sur comment choisir des modèles de bruit permettant d’éviter les extrema locaux (par exemple par du recuit simulé), ce qui vient au final ajouter de la complexité au modèle sensé être simple, et allonge les temps de calcul. En outre, les critères que nous avons maintenant ne prennent pas en compte la phase et les permutations.

On retrouve aussi une grande dépendance à l’initialisation des paramètres (IV.C.2), problème déjà vu en NMF. Les résultats seront bien entendu meilleurs si on connaît en avance le nombre de canaux à identifier, le nombre de composantes par source, et surtout s’il est nécessaire ou non d’exploiter l’information mutuelle entre les canaux. Mais ces informations ne sont pas toujours à disposition, ou bien elles doivent être suggérées avant à l’algorithme. Ici, les auteurs insistent sur le fait que les données permettent la convergence de l’algorithme (“data-driven”), mais en aucun cas on ne parle d’initialiser en fonction de ces données.

Il aurait été intéressant de développer sur comment sont combinés les algorithmes EM et MU dans l’obtention du résultat. En effet, on voit bien que l’algorithme EM obtient de très bonnes performances, mais l’algorithme MU a beaucoup de mal à surpasser ces performances, et même parfois vient réduire le score de séparation final. Bien qu’il y ait eu une allusion à ces mauvais scores, les auteurs n’ont de mon point de vue pas assez proposé de solutions à l’algorithme MU.

Les performances finales sont bonnes pour ces algorithmes, mais les chercheurs ont identifié un nouveau problème : les “nonpoint sources”, sources audio qui sont le résultat de plusieurs enregistrements avec des filtres non linéaires appliqués. Ces situations sont très réalistes dans le paysage de la production musicale actuelle, et les résultats des algorithmes proposés ici ne sont pas très bons par rapport à cette situation. Comme ils le disent eux-mêmes, c’est une nouvelle piste de recherche à explorer (IV.H).

Cet article fait partie des articles les plus cités pour ces deux chercheurs, comportant presque 700 citations. Il est repris dans des publications scientifiques datant d’il y a à peine quelques jours, concernant des généralisations de cette méthode, ou des approches Deep Learning. Alexey Ozerov a quant à lui continué sur l’adaptabilité de ces méthodes NMF à la séparation de sources audio, avec par exemple l’utilisation de tenseurs plutôt que de matrices [1] comme évoqué dans notre introduction. Les deux auteurs ont d’ailleurs à nouveau collaboré pour adresser notamment le problème de la supervision de l’apprentissage (comment apporter les données, comment apprendre les paramètres à l’algorithme) [2]. Tout cela montre que l’article a eu un vrai impact dans ce domaine, et qu’il a soulevé des questions qui ont été par la suite des nouveaux sujets de recherche.

Personnellement, pour faire suite à ma remarque à propos de l’enchaînement EM/MU, je pense qu’il aurait été intéressant d’activer/désactiver l’algorithme MU de façon aléatoire pour observer le réel impact de cet algorithme sur le score final, on observe en effet que le score s’améliore au début des itérations, mais qu’il décroît sur la fin. Peut-être que des appels moins fréquents à l’algorithme pourraient se révéler efficaces pour notre score. En outre, on aurait aussi pu étudier les résultats en utilisant la divergence de Kullback-Leibler plutôt que celle d’Itakura-Saito (bien que cette dernière soit optimale pour le modèle statistique choisi).

J’ai choisi ce papier car tout d’abord j’ai préféré la partie traitement de l’audio/deep learning par rapport à la partie reconnaissance d’émotions/gestes musicaux du cours, puis la notion de NMF me semblait originale (on ne l’a étudié qu’à la toute fin de l’année, mais le nom m’intriguait). Ce papier semblait bien documenté et très précis sur la méthodologie employée (notamment sur le choix du bruit pour obtenir des résultats optimaux). Le traitement du signal de façon générale m’intéresse (audio cette année, images l’année prochaine), et le traitement audio en particulier possède ses challenges propres : domaine temps-fréquence, grande quantité de données, taille adaptative des résultats et des vecteurs/filtres employés pour le traitement; ainsi je trouve pertinent de s’intéresser à ces aspects spécifiques qui seront certainement utiles pour le traitement d’autres types de signaux.

[1] Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation. A Ozerov, C Févotte, R Blouet, JL Durrieu

*2011 IEEE International Conference on Acoustics, Speech and Signal*, 2011

[2] Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation. A Ozerov, C Févotte, R Blouet, JL Durrieu

*Proc. IEEE International Conference on Acoustics, Speech and Signal*, 2011