Fiche descriptive de la thèse
Campagne 2013

Encadrant de la thèse Orange Labs : Alexandre GUERIN – Marc EMERIT
Pôle : OLPS
Direction : COMSERV
Département : SVQ
Projet : TC Audio
Terrain de conquête :
Site : Rennes

Sujet de la thèse :
« Analyse de scène sonore multi-capteurs : un front-end temps-réel pour la manipulation de scène »

Contexte global de l'étude et état de l'art

On assiste depuis quelques années à l'avènement de la prise de son multi-capteurs : longtemps cantonnée à l'utilisation de couples de micros (AB, XY, ORTF, …), la création de contenus, poussée par la démocratisation du home-cinema et de son format audio 5.1, élargit sa palette à des « antennes de microphones » permettant de meilleures représentation et restitution des scènes sonores. Quant au domaine des télécommunications, les prises de son au moins stéréo sont en passe de devenir un standard : on ne compte plus les téléphones 3G/4G qui intègrent 2 microphones (iPhone4 pour ne citer que celui-ci) ou les pieuvres d'audioconférence qui intègrent plusieurs microphones (3 sur les pieuvres Polycom).
Ces prises de son audio multi-canal permettent d'envisager une analyse et une manipulation « avancées » de la scène sonore : atténuation conséquente du bruit de fond et des sources interférentes (c'est l'application principale pour les terminaux de télécommunication), mixage automatique de différentes prises de son d'une même scène sonore, modification de la scène (exemple : repositionnement d'instrument), …
Cette analyse nécessite en premier lieu le comptage et la localisation des sources présentes dans le mélange : différentes approches existent, suivant que, rapporté aux nombres de sources, le nombre de canaux est supérieur [1,2] (on parle de mélange sur-déterminé) ou inférieur [3,4] (on parle de mélange sous-déterminé). Ensuite, la manipulation de la scène s'apparente à de la séparation aveugle de sources (BSS pour Blind Source Separation), dont les multiples approches dépendent encore une fois du caractère sur ou sous-déterminé du mélange. La littérature est particulièrement foisonnante sur le sujet : on trouvera des bases de BSS sur-déterminée dans [5], ainsi que la description d'une méthode sous-déterminée historique utilisant l'hypothèse de parcimonie [6].

[1] Brandstein (M.) et Silverman (H.). « A robust method for speech signal time-delay estimation in reverberant rooms ». *ICASSP '*97.
[2] Yamamoto (K.), Asano (F.), van Rooijen (W.), Ling (E.), Yamada (T.) et Kitawaki (N.). « Estimation of the number of sound sources using support vector machines and its application to sound source separation ». *ICASSP '03.*
[3] Arberet (S.), Gribonval (R.) et Bimbot (F.), « A robust method to count and locate audio sources in a multichannel underdetermined mixture », *IEEE Transactions on Signal Processing*, **58**(1), Jan. 2010, p. 121–133
[4] El Chami (Z.), Pham (D.-T.), Servière (C.) et Guérin (A.). « A phase based method to count and locate audio sources in reverberant environment ». WASPAA '09.
[5] Comon (P.), « Independent component analysis, a new concept ? », Signal Process., 36(3), 1994, p. 287–314.
[6] Yilmaz (O.) et Rickard (S.), « Blind separation of speech mixtures via time-frequency masking », Signal Processing, IEEE Transactions on Acoustics, Speech, and Signal Processing, 52(7), 2004, p. 1830–1847.

Objectifs de la thèse / Résultats attendus / Défis scientifiques et techniques à relever

L'objectif de la thèse est de développer, à partir d'une prise de son multi-capteurs (i.e. une prise de son au moins stéréo et si possible compacte pour viser l'application mobile), des méthodes d'analyse de scène sonore. L'analyseur de scène développé devra fournir, en temps-réel, une description temps-fréquence-espace de la scène sonore. Plus particulièrement, les travaux de recherche chercheront à résoudre les problématiques suivantes :

- détermination du nombre de sources actives dans le mélange
- détermination de la « position » de ces sources, ou du caractère diffus en cas de bruit de fond
- suivi de ces sources dans le plan temps-fréquence-espace

La principale difficulté réside dans la nature du mélange : les scènes sonores réelles se caractérisent tout d'abord par une réverbération qui vient perturber les modèles efficients d'analyse de scène développés pour les mélanges anéchoïques.
D'autre part, en situation réelle, le nombre de sources, outre qu'il est inconnu, peut être supérieur au nombre de capteurs, ce qui ne permet pas d'utiliser des approches de type séparation de sources pour des mélanges « *sur-déterminés* » (nombre de sources inférieur au nombre de capteurs), approches qui sont relativement robustes, tant que la réverbération reste mesurée.
L'autre difficulté réside dans l'aspect temps-réel, nécessaire pour des applications de télécommunications. En effet, les méthodes automatiques d'analyse de scène utilisent généralement des informations long-terme (moyennage sur des fenêtres temporelles relativement longue) afin de limiter la variance des estimateurs, et donc d'augmenter la robustesse. Le suivi de la trajectoire d'une source potentiellement mobile impose une prise de décision *i*) dite à *faible latence*, *ii*) causale (sans connaissance du futur ; cela est *a contrario* autorisé dans le cas des contenus).
Enfin, la compacité des systèmes de captation (c'est le cas de couples de microphones de type XY, ou encore des microphones de type Ambisonic) peuvent fortement dégrader les méthodes de comptage et de séparation de sources qui utilisent avantageusement des informations de phase.

L'objectif, et le défi, sera de proposer une approche efficiente quelle que soit la nature du mélange, y compris réverbérante et potentiellement sous-déterminée (nombre de sources supérieur au nombre de capteurs), en tenant compte des contraintes de causalité et de « faible latence », ainsi que des contraintes physiques liées au système de prise de son.

Approche méthodologique proposée par le responsable technique

Cette thèse s'inscrit dans la continuité de travaux précédents sur la séparation de sources bi-capteurs, travaux qui avaient abouti à des techniques d'analyse de scènes statiques (les techniques développées se basent uniquement sur des informations statistiques moyennées sur une fenêtre temporelle d'analyse), notamment la détermination du nombre de sources et leurs positions.
Une première phase consistera à prendre en main ces technologies et à identifier leurs limites d'un point de vue statique (agrégation de bins temps-fréquence propres à chaque source) et dynamique (sensibilité à la taille de la fenêtre d'analyse et robustesse au mouvement des sources).
Une deuxième phase consistera à améliorer l'agrégation des bins temps-fréquence propres à chaque source : on introduira dans cette phase la « cohérence » dans le plan fréquentiel qui permettra de fiabiliser les méthodes pré-existantes. Si nécessaire, on pourra utiliser des informations *a priori* sur le modèle des sources.
La troisième phase visera à opérer le suivi des sources dans le plan temps-fréquence. Ce suivi pourra utiliser conjointement les trajectoires temporelles de composantes fréquentielles, et la cohérence fréquentielle des sources.

Les compétences requises pour cette thèse sont les suivantes :

- Formation de type Master2 Recherche
- Bonne maîtrise des outils mathématiques pour le traitement du signal : analyse statistique, décomposition temps-fréquences, traitement d'antennes
- Bases en séparation de sources et traitement de la parole (codage, modélisation)
- Maîtrise des langages C/Matlab
- Sensibilité à l'audio

Contact

Alexandre Guérin
alexandre.guerin@orange.com
02 99 12 49 50