Techniques statistiques pour la séparation de sources audio dans un mélange monocapteur

Mémoire de magistère
de mathémathiques fondamentales et
appliquées et d'informatique
– octobre 2002 –

Emmanuel Vincent

École Normale Supérieure
Université Paris XI
IRCAM

Résumé

Ma recherche porte sur la séparation de sources audio musicales dans un enregistrement monocapteur. Après avoir présenté les prérequis nécessaires en traitement du signal, je pose le problème et ses applications. Je décris les spécificités des sources musicales, ainsi que l'état de l'art du problème. Puis j'étudie deux modèles récents : l'analyse en sous-espaces indépendants et les modèles de Markov cachés. Je présente très brièvement quelques résultats personnels et perspectives.

Table des matières

1 Traitement du signal audio

Le traitement du signal est une discipline datant des années 50, à la croisée entre les mathématiques et l'informatique.
Un signal est la forme physique d'une information véhiculée par un système. L'appellation traitement du signal regroupe les techniques destinées à transformer, analyser, classer, segmenter, débruiter, compresser, synthétiser des signaux, etc...

En audio, un signal peut être une onde acoustique, un courant électrique dans un microphone, ou une représentation binaire sur CD.
Le traitement du signal audio a de nombreuses applications telles que la création de logiciels de modification de sons, l'utilisation d'effets sonores temps-réel 3D pendant des concerts, la détection de défauts dans des machines-outils, la recherche par similarité dans des bases de données de musique...

Mon travail se situe dans le cadre du traitement statistique du signal. La plupart des signaux sont en effet soumis à de nombreuses variations intrinsèques au système ou à du bruit de mesure qui contribuent à cacher l'information utile.

2 Notions de base

Spectre : transformée de Fourier complexe d'un signal, pouvant se représenter en termes d'amplitude et de phase.
L'amplitude se représente souvent en échelle logarithmique ou échelle des décibels (dB).
La précision fréquentielle du spectre est proportionnelle à l'étalement temporel du signal analysé : c'est la loi d'incertitude.

Filtrage : convolution d'un signal par un autre signal appelé réponse impulsionnelle du filtre.
La convolution devient un produit dans le domaine spectral : filtrer permet de mettre en valeur l'énergie à certaines fréquences.
Un filtrage peut être stationnaire si le filtre est le même à chaque instant, ou non stationnaire dans le cas contraire.

Bruit : signal pouvant être représenté par un processus aléatoire i.i.d. (bruit blanc), éventuellement filtré (bruit coloré).
Le spectre d'un bruit blanc contient la même énergie à toutes les fréquences.

Par extension, désigne toute partie indésirable d'un signal.

Partiel : partie d'un signal pouvant être représentée par une sinusoïde de fréquence et d'amplitude variant lentement dans le temps.

Temps-fréquence, temps-échelle : représentations inversibles d'un signal, à deux dimensions, montrant l'évolution du spectre au cours du temps.
Une représentation temps-fréquence (spectrogramme par exemple) utilise une échelle linéaire de fréquence et de temps, alors qu'une représentation temps-échelle (transformée en ondelettes par exemple) utilise une échelle logarithmique de fréquence et de temps, avec une précision fréquentielle proportionnelle à la fréquence et une précision temporelle inversement proportionnelle à la fréquence.

Formule de Bayes : formule probabiliste permettant d'analyser des hypothèses θ sur des observations x grâce à un *a priori* sur ces hypothèses $P(\theta)$, appris ou fixé manuellement.
La loi s'exprime comme $P(\theta|x) = P(x|\theta)P(\theta)/P(x)$ où $P(\theta|x)$ est la probabilité *a posteriori* des hypothèses, $P(\theta)$ leur probabilité *a priori* et $P(x|\theta)$ la vraisemblance des observations.

3 Séparation de sources audio

La séparation de sources désigne les techniques visant à retrouver des signaux inconnus appelés sources à partir d'une ou plusieurs observations de leur mélange, dont les caractéristiques ne sont pas données. Ce problème est difficile, surtout en présence de bruit ou lorsque le mélange est de type convolutif ou non stationnaire.
Lorsqu'on possède moins d'observations que de sources (cas sous-déterminé), le problème est plus difficile car le mélange est non inversible : retrouver la manière dont les sources ont été mélangées ne suffit pas à les séparer.
Quand on dispose d'au moins deux observations, il est possible d'utiliser les caractéristiques spatiales du mélange, c'est-à-dire de localiser dans les observations les zones correspondant aux mêmes sources. Lorsqu'on n'a qu'une seule observation (cas monocapteur), ce n'est plus possible.

Parmi les applications de la séparation de sources audio, on peut citer notamment la séparation du signal utile du bruit environnant en téléphonie mobile, la description automatique (indexation) de do-

cuments multimédia contenant plusieurs objets sonores, la reconnaissance de la parole multi-locuteur (*cocktail party problem*) ou encore la localisation de sources en vidéosurveillance.
La séparation de sources est liée à plusieurs questions importantes du traitement du signal audio : la représentation des signaux, le calcul de la fréquence fondamentale, la reconnaissance d'instruments, la synthèse de sons instrumentaux plausibles perceptivement, la segmentation de morceaux de musique...

La plupart des contributions dans le domaine prétendent être applicables dans de nombreux cas, mais elles posent parfois mal le problème ou utilisent une mauvaise méthode d'évaluation de leurs résultats. À ce sujet, j'ai rédigé et co-rédigé deux rapports au sein du groupe de travail Action Jeunes Chercheurs GdR ISIS dont je fais partie [ISIS] pour montrer l'importance de l'application visée dans le choix d'un modèle et d'une méthode d'évaluation [Vin02a], et pour proposer des critères d'évaluation et des bases de données adaptées [Gri02].

Ma recherche se focalise sur le problème suivant : séparer au sein d'un enregistrement audio monocapteur des sources instrumentales simplement additionnées. Mon but est d'obtenir des sources de haute qualité, c'est-à-dire autant que possible sans distorsion et sans bruit en provenance des autres sources.

4 Caractéristiques du son instrumental

On distingue deux grandes familles d'instruments, selon que l'oscillation produisant le son est libre (percussions, piano) ou entretenue (violon, hautbois, flûte, saxophone, trompette).

Une note isolée d'un instrument entretenu se modélise généralement en trois parties, ou modèle *ASR* : l'attaque, la partie soutenue, et la décroissance (*release*). Durant l'attaque, le système commence à osciller à des fréquences presque harmoniques. Un bruit transitoire est généralement présent tant que l'excitation n'est pas parfaitement périodique. La partie soutenue est la poursuite de cette oscillation. La décroissance voit l'énergie de l'oscillation diminuer. Pour un instrument à oscillations libres, seul le transitoire d'attaque est présent. Ce transitoire contient de l'énergie à des fréquences qui peuvent être harmoniques ou non.
Dans une situation normale de jeu, les notes ne sont plus isolées mais connectées par exemple par des transitions *legato*. Pour les instruments

entretenus, ces transitions ne sont pas la simple superposition de la décroissance d'une note N et de l'attaque de la note $N+1$ [Str85]. La durée de la transition est souvent inférieure à celle d'une attaque , et la forme exacte de la transition peut dépendre de l'intervalle entre les deux notes, voire des deux notes elles-mêmes. De plus, la réverbération des notes 1 à N se superpose au son de la note $N+1$ proprement dit.

Le son instrumental se décrit *grosso modo* grâce à quatre paramètres : la hauteur, le volume, la durée et le timbre. Ces paramètres peuvent se relier aux variations du signal, mais rien n'est direct. Par exemple, une variation de hauteur ne se manifeste en général pas seulement par le changement de la fréquence fondamentale : la forme globale du spectre varie aussi. Et les variations de hauteur n'ont pas lieu seulement entre notes différentes : il existe des micro-variations au sein de chaque note comme le *vibrato*.
Le timbre est un terme fourre-tout ne pouvant être défini par une simple valeur. Sa nature complexe est étudiée dans de nombreux cadres, par exemple celui de la reconnaissance d'instruments [Mar99]. La forme des attaques et des transitions, et en particulier celle de l'énergie inharmonique qu'elles contiennent, a une part importante dans la perception du timbre [Dup99].

Plusieurs catégories de modèles ont été proposées pour le son instrumental, des modèles de production (*i.e.* de l'instrument lui-même) aux modèles de signal (*i.e.* du son produit par l'instrument). Les trois méthodes de séparation de sources présentées dans les parties suivantes utilisent trois modèles de signal distincts.

5 Analyse de scènes auditives

Introduite par Bregman [Bre90] et constituant l'état de l'art en séparation de sources monocapteur, l'analyse de scènes auditives (*auditory scene analysis*, ASA) cherche à expliquer la manière dont l'ensemble oreille-cortex auditif perçoit un ensemble complexe de sons par une série de règles psychoacoustiques. Ces règles s'appliquent sur une représentation partiels + bruit : le signal est modélisé par un ensemble d'éléments, pouvant être des groupes de partiels harmoniques ou des "nuages" de bruit coloré.
De nombreuses règles ont été implémentées et testées dans la thèse d'Ellis [Ell96]. Un mélange est analysé par tranches de temps successives : à chaque instant, plusieurs hypothèses de modélisations possibles sont faites en fonction du contenu du signal, et la probabilité

a posteriori de ces hypothèses est calculée en fonction des hypothèses des instants précédents et des règles en question.

Elégante en apparence, cette technique souffre d'un gros défaut : pour bien expliquer la création d'objets auditifs au sein d'une scène sonore, il faut considérer de nombreuses règles parfois antagonistes, ce qui ramène souvent à une série d'heuristiques complexes.
D'autre part, la probabilité d'une modélisation comportant plusieurs éléments de nature et de taille différentes est difficile à évaluer. Un élément plus gros représente mieux le signal, mais peut échouer à distinguer les informations significatives du bruit et doit donc être pénalisé. Ellis utilise le critère de longueur de description minimale (*minimum description length*, MDL), mais laisse de côté la question importante du choix de l'échelle relative entre la vraisemblance du modèle et la pénalité.

Les deux modèles suivants que j'ai étudié tentent de montrer qu'une modélisation unifiée des sons peut également servir à la séparation de sources et nécessiter moins de paramètres à régler.

6 Analyse en sous-espaces indépendants

L'analyse en composantes indépendantes [Car98] (*independent component analysis*, ICA) modélise les sources comme des processus aléatoires i.i.d. indépendants.

Lorsqu'on possède autant d'observations du mélange que de sources et que le mélange est linéaire instantané et inversible, cette technique permet de retrouver les sources avec de très bons résultats.
Notons B la matrice représentant l'opération de démixage, x le vecteur aléatoire centré représentant les observations et $y = Bx$ celui des sources estimées. On définit le B optimal comme celui qui maximise l'indépendance des sources estimées y. En notant $\tilde{y}$ le vecteur aléatoire dont les marginaux sont indépendants et de même distribution que ceux de y, le problème se ramène à minimiser le critère dit d'information mutuelle, exprimé par la divergence de Kullback $K(y|\tilde{y})$. Cette divergence est toujours positive et nulle ssi y et $\tilde{y}$ sont indentiques.
La minimisation est effectuée par descente de gradient, en approchant les densités de probabilité par des modèles paramétriques. Le modèle le plus populaire fait intervenir les variances $E(y_i^2)$ et les *kurtosis* $E(y_i^4) - 3E^2(y_i^2)$. Lorsqu'une distribution est parcimonieuse (*sparse*), c'est-à-dire lorsqu'elle génère beaucoup d'observations proches de zéro, son

kurtosis est élevé. Pour les applications audio, la plupart des sources sont parcimonieuses (zones de silence ou de faible volume sonore), et cette propriété tend à diminuer après mélange : l'ICA rend les sources les plus parcimonieuses possible.

Dans le cas monocapteur, les hypothèses de l'ICA s'adaptent en remarquant que les valeurs de deux sources distinctes à différentes fréquences ou à différents instants sont indépendantes. Une partie des règles d'ASA formulent cette idée sous un angle non probabiliste, en proposant de regrouper en une seule source des partiels qui ont les mêmes variations en amplitude et en fréquence.
Casey [Cas00] exprime le spectrogramme d'amplitude de chaque source i comme le produit d'un vecteur normalisé de caractéristiques fréquentielles F_i et d'un vecteur normalisé de caractéristiques temporelles T_i, ainsi que d'une amplitude a_i. Le spectrogramme X des observations se décompose alors en $X = \sum_i a_i F_i T_i^T + bruit$.
Pour retrouver les F_i et T_i, une décomposition en valeurs singulières de X fournit un premier résultat de ce type avec des F_i et T_i décorrélés. Il prend ensuite pour hypothèse tantôt l'indépendance des F_i, tantôt celle des T_i, et applique une ICA à ce résultat. Lorsqu'une source présente plusieurs comportements, il la représente à l'aide de plusieurs vecteurs caractéristiques, regroupés suite à l'ICA par des critères de distance. On parle alors d'analyse en sous-espaces indépendants. Le spectrogramme de chaque source est finalement inversé pour reconstituer les sources proprement dit.

Étudié lors de mon stage de DEA [Vin01], ce modèle s'est révélé efficace pour l'analyse de bruit environnementaux et l'extraction d'attaques. L'utilisation d'une représentation unifiée (le spectrogramme) a effectivement permis de réduire de beaucoup le nombre de paramètres à régler.
Mais il semble inutilisable pour une séparation de qualité dans un contexte musical. L'expression des sources grâce aux F_i et T_i est adaptée aux structures temps-fréquence "horizontales" (partiels de fréquence stable) ou "verticales" (bruits), mais pas aux "diagonales", pourtant souvent présentes en musique (*vibrato* par exemple). Cette même expression conduit à l'extraction de sources au son très distordu.

La partie la plus criticable de ce modèle concerne les hypothèses mises en jeu.
L'addition des sources dans le domaine de l'amplitude ne reflète pas la réalité. J'ai effectué des tests avec une addition en énergie, mais les sources de volume sonore important "cachent" alors les autres.

Le choix heuristique entre l'indépendance des F_i ou celle des T_i est aussi peu réaliste. J'ai montré dans certains cas simples qu'une combinaison de ces deux critères donnait de meilleurs résultats, et que le choix du paramétrage des densités de probabilité avait une influence décisive.
J'ai enfin montré que l'ajout de connaissances *a priori* telles que le spectre approximatif des sources ou leur support temporel pouvait également améliorer les résultats.

7 Modèle de Markov caché

Le modèle de Markov caché (*hidden Markov model*, HMM) est un moyen de décrire une suite d'observations (X_t) par un nombre réduit de causes sous-jacentes (états) s'enchaînant dans le temps (S_t). Les hypothèses sont simples : X_t ne dépend que de l'état pris par S_t, et (S_t) est une chaîne de Markov, d'ordre 1 généralement, à valeurs dans un ensemble fini. Le modèle M est la donnée des lois d'observation $P(X_t|S_t = i)$, des probabilités initiales $P(S_1 = i)$ et des probabilités de transition $P(S_t = j|S_{t-1} = i)$.
En traitement du signal audio, les X_t sont souvent les spectres d'amplitude du signal au cours du temps et les S_t prennent des significations physiques particulières (phonème, note...)

On appelle reconnaître un signal le fait de retrouver les états cachés décrivant cet signal à partir d'un modèle fixé. Selon les applications, il s'agit de retrouver le chemin (la suite d'états) I le plus probable $argmax_I\ P((S_t) = I|(X_t), M)$ ou la probabilité *a posteriori* de chaque état i à chaque temps t $P(S_t = i|(X_t), M)$.
Avant cela, il faut apprendre les paramètres du modèle sur un ensemble d'exemples (Z_t). Le modèle optimal (parmi une famille de modèles) est celui qui maximise la vraisemblance de ces exemples, soit $argmax_M\ P((Z_t)|M)$. Souvent, cet apprentissage est supervisé, c'est-à-dire que des chemins sous-jacents aux exemples fixés manuellement servent à initialiser l'apprentissage, et donnent à chaque état une signification physique.

Cette modélisation s'applique aux mélanges de plusieurs sources grâce à une sous-catégorie de modèle : le modèle de Markov caché factoriel (*factorial hidden Markov model*, FHMM) [Gha97]. L'observation du mélange X_t dépend non plus d'une mais de deux chaînes de Markov cachées indépendantes S_t et U_t, chaque chaîne décrivant une source. Le modèle contient les paramètres de chaque chaîne, ainsi

que la loi d'observation combinée $P(X_t|S_t = i, U_t = j)$. Il peut se ramener à un HMM classique en considérant comme états les couples $(S_t = i, U_t = j)$. L'expression en tant que FHMM est une forme factorisée.
Le FHMM permet de reconnaître des mélanges de deux sources en trouvant le couple de chemins (I, J) de probabilité maximale $argmax_{(I,J)}\ P((S_t) = I, (U_t) = J|(X_t), M)$ ou la probabilité de chaque couple d'états (i, j) à chaque temps t $P(S_t = i, U_t = j|(X_t), M)$.

Roweis [Row00] a défini une méthode permettant d'utiliser le FHMM en séparation de sources de parole. Premièrement, un HMM classique est appris séparément pour chaque source en situation isolée. Deuxièmement, les lois d'observations de chaque HMM sont combinées pour former un FHMM et reconnaître ces sources dans un mélange. Troisièmement, les sources proprement dit sont estimées par filtrage non stationnaire du mélange à l'aide du modèle et des résultats de la reconnaissance.
Son modèle, non supervisé, représente les sources par leur spectrogramme en décibels, et transforme leur addition en maximum point par point sur le spectrogramme. Les lois d'observation sont paramétrées par des gaussiennes partageant une covariance diagonale unique. Enfin, les sources sont estimées en attribuant chaque point temps-fréquence à la source la plus puissante en ce point, et en inversant les spectrogrammes ainsi obtenus.

J'ai adapté ce modèle à la séparation de sources musicales en y apportant plusieurs modifications ou améliorations [Vin02b].
En particulier, une méthode a été définie pour permettre un apprentissage supervisé, utilisant un algorithme existant de segmentation de morceaux en notes grâce à la partition [Sch01]. Les états correspondent alors à des couples "transition de note N à note $N + 1$" et "partie soutenue note $N + 1$ plus réverbération note N". La phase d'estimation des sources après reconnaissance a été modifiée pour minimiser la distorsion. Et l'utilisation d'une transformée temps-échelle calculée sur des intervalles de temps fixes à la place du spectrogramme a permis de mieux séparer des sons de fréquences fondamentales proches sans trop augmenter la dimension des observations.
En plus de la séparation haute qualité des sources, deux autres buts ont été étudiés : la retranscription de partition (*wav2midi*) et la modification de scène sonore, qui consite à créer un "remix" de bonne qualité du mélange en modifiant un peu chaque source après démixage avant de remélanger [Vin02a].
J'ai obtenu des résultats encourageants pour ces trois problèmes, et

particulièrement pour le dernier. Encore une fois, le nombre de paramètres à régler est faible, ce qui constitue un atout.

Cependant, l'initialisation de l'apprentissage supervisé s'est révélée assez difficile, et j'ai été confronté à un manque de données d'apprentissage comparé à la taille élevée du modèle. Par conséquent, ces résultats ne sont pas concluants.
Enfin, ce modèle procédant par filtrage du mélange ne permet pas toujours la séparation haute qualité des sources. Lorsqu'une source en masque une autre dans une zone temps-fréquence, le mélange perd les informations sur la source masquée et seule une synthèse permet de la reconstituer. Et la synthèse de sons perceptivement acceptables est un problème difficile, nécessitant un modèle bien plus complexe que celui-ci.

8 Conclusion et perspectives

La séparation de sources est un des problèmes fondamentaux du traitement du signal audio. L'analyse de scènes auditives, approche classique du problème dans le cadre monocapteur, est pénalisée par l'utilisation de plusieurs représentations différentes du son et la nécessité de régler de nombreux paramètres heuristiquement.
L'analyse en sous-espaces indépendants et le modèle de Markov caché échappent à ces problèmes, avec ou sans apprentissage préliminaire. Les résultats semblent prometteurs pour le modèle de Markov caché, mais pas encore concluants. Des difficultés subsistent principalement quant à l'initialisation de l'apprentissage.

Sans être au cœur du modèle, l'utilisation d'une représentation partiels + bruit pourrait aider cette initialisation et permettre ainsi plus facilement l'obtention de grands ensembles d'apprentissage. Si les résultats se révèlent alors bons, on pourra étudier d'autres problèmes, comme reconnaître le nom des instruments présents dans un mélange ou resynthétiser les informations perdues lors du mélange.

Références

[Bre90] A.S. Bregman. Auditory scene analysis : the perceptual organization of sound. *MIT Press*, 1990

[Car98] J.-F. Cardoso. Blind signal separation : statistical principles. *Proc. IEEE*, 90(8) :2009-20026, oct. 1998

[Cas00] M.-A. Casey and A. Westner. Separation of mixed audio sources by independent subspace analysis. *Proc. ICMC2000*, 2000

[Dup99] S. Dupuis. Le rôle des transitions legato dans la reconnaissance des instruments de musique. *Rapport de DEA de Sciences Cognitives*, 1999

[Ell96] D. Ellis. Prediction-driven computational auditory scene analysis. *PhD thesis, MIT*, 1996

[Gha97] Z. Ghahramani and M.I. Jordan. Factorial hidden Markov models. *Machine Learning*, 29 :245-273, 1997

[Gri02] R. Gribonval et E. Vincent. Analyse critique des bases de données et critères de performance proposés à ICA'99. *Rapport interne, Action jeunes chercheurs GdR ISIS*, 2002, inachevé

[ISIS] IRCAM, IRISA et IRCCyN. Action jeunes chercheurs GdR ISIS. http ://www.ircam.fr/anasyn/ISIS/

[Mar99] K.D. Martin. Sound-source recognition : A theory and computationnal model. *PhD thesis, MIT*, 1999

[Row00] S. Roweis. One microphone source separation. *Proc. NIPS00*, 793-799, 2000

[Sch01] D. Schwarz and N. Orio. Alignment of Monophonic and Polyphonic Music to a Score. *Proc. ICMC2001*, 2001

[Str85] J. Strawn. Modeling musical transitions. *PhD thesis, CCRMA, Stanford University*, 1985

[Vin01] E. Vincent. Séparation de signaux audio : principes statistiques de l'analyse en composantes indépendantes et applications au signal monophonique. *Rapport de DEA ATIAM*, 2001

[Vin02a] E. Vincent. Problèmes typiques en séparation de signaux audio. *Rapport interne, Action jeunes chercheurs GdR ISIS*, 2002

[Vin02b] E. Vincent. Modélisation par modèles de Markov cachés pour la séparation de sources musicales dans un enregistrement monocapteur. *Rapport de première année de doctorat*, 2002