Modèles hiérarchiques pour la modélisation de signaux audio

Matthieu KOWALSKI[1], Bruno TORRÉSANI

[2]Laboratoire d'Analyse, Topologie et Probabilités
CMI, 39 rue Joliot-Curie, 13453 Marseille Cedex 13, France
kowalski@cmi.univ-mrs.fr,Bruno.Torresani@cmi.univ-mrs.fr

Résumé – On considère le problème de régression parcimonieuse et structurée des signaux dans un dictionnaire temps-fréquence. On choisit pour cela une modélisation aléatoire hiérarchique des signaux, dans laquelle des propriétés de persistance sont introduites via un modèle simple à états cachés sur les coefficients de synthèse d'un développement temps-fréquences, avec modélisation explicite des corrélations sur ceux-ci. L'étude du modèle conduit à des méthodes de classifications de type CEM des coefficients d'analyse, qui exploite les différences de comportement des matrices de variance/covariance mise en évidence par le modèle.

Abstract – We consider the structured and sparse regression problem in time-frequency dictionaries. We develop a hierarchical signal model, in which persistence properties are introduced through a simple hidden state model for synthesis coefficients in a time-frequency decomposition, with an explicit modelling of the corresponding correlations. The model directly leads to CEM type classification of analysis coefficients, which exploits de diversity of behaviors of the covariance matrices.

1 Introduction

Il est maintenant bien connu que de nombreuses classes de signaux, parmi lesquelles les signaux audio, peuvent être représentées de façon efficace dans des *dictionnaires temps-fréquence*, c'est à dire des familles surcomplètes de formes d'ondes élémentaires (atomes de Gabor, ondelettes,...). L'efficacité se juge souvent en termes de parcimonie, qui mesure la « concentration » de la suite de coefficients représentant le signal.

Le problème de « régression parcimonieuse » consiste à sélectionner, parmi une famille de représentations, celle qui concentre au mieux l'information dans un nombre limité de coefficients. Différentes approches ont été proposées pour résoudre ce problème, parmi lesquelles on peut notamment mentionner les approches variationnelles, les approches basées sur des algorithmes de poursuite, ou encore celles basées sur une modélisation probabiliste.

Nous nous intéressons ici aux représentations des signaux qui, en plus d'être parcimonieuses, sont *structurées*, au sens où les coefficients de la décomposition sur le dictionnaire ne sont plus supposés *i.i.d.*, mais dépendants. Plus précisément, les dépendances sont organisées sous formes de *lignes persistantes* de coefficients significatifs, la persistance pouvant être soit temporelle soit fréquentielle. Le problème d'estimation de tels modèles peut être abordé sous l'angle de formulations variationnelles [2], ou d'algorithmes de poursuite. Nous nous intéressons ici à l'approche probabiliste, et décrivons de nouveaux modèles explicites, ainsi que des méthodes d'estimation correspondantes. Nous décrivons également une nouvelle approche pour l'identification des cartes de signifiance, basée sur des estimateurs de corrélations de coefficients d'analyse à l'intérieur de groupes de coefficients d'analyse, généralisant les modèles de Bernoulli et Bernoulli hiérarchique étudiés dans [3].

2 Modèles hiérarchiques et estimation

Nous nous intéressons ici à des modèles hiérarchiques, dans lesquels les coefficients sont modélisés par des variables aléatoires, dont la distribution est gouvernée par un état caché.

On considère un dictionnaire $\mathcal{D} = \{\varphi_k\}$ de formes d'ondes, et un signal modèle de la forme

$$x = \sum_k \alpha_k \varphi_k \ ,$$

où les *coefficients de synthèse* α_k forment un vecteur aléatoire. Dans le modèle hiérarchique le plus simple, les coefficients α_k sont supposés indépendants, conditionnellement à un état caché X_k. Dans le modèle de Bernoulli, les X_k sont *i.i.d.*, et le problème d'estimation de la carte de signifiance, c'est à dire des X_k se ramène à un problème de seuillage adaptatif des *coefficients d'analyse* $\langle x, \varphi_k \rangle$ (voir [3]). L'introduction de structures, c'est à dire de relations de dépendance entre coefficients de synthèse, peut s'effectuer en introduisant de telles relations dans la carte de signifiance, tout en conservant l'indépendance des coefficients conditionnellement à la carte. C'est par exemple l'approche suivie dans [4, 3], où des modèles de Markov ou de Bernoulli hiérarchique sont étudiés.

On s'intéresse ici à une autre approche, dans laquelle ces structures sont introduites dans les coefficients eux mêmes, et

des cartes de signifiance simplifiées sont utilisées. Le dictionnaire considéré est un dictionnaire d'atomes temps-fréquence $\mathcal{D} = \{\varphi_{tf}\}$, le signal modèle étant de la forme

$$x = \sum_{t,f} \alpha_{t,f} \varphi_{t,f} \ .$$

Les cartes de signifiance simplifiées sont alors les marginales temporelle et fréquentielle, ce qui définit des groupes fréquentiel et temporel de coefficients.

2.1 Modèles, coefficients d'analyse

Modèle hiérarchique et structuré La distribution des coefficients aléatoires est supposée « structurée », dans le sens suivant. La carte de signifiance (i.e. l'ensemble des états cachés) est supposée ne dépendre que de l'un des deux indices (indice temporel ou indice fréquentiel). Pour fixer les idées, supposons que cet indice soit l'indice temporel. Conditionnellement à la séquence des états cachés $\{X_t\}$, les vecteurs de coefficients de synthèse à t fixé $\underline{\alpha}_t = \{\alpha_{t,f}, f = 1, \dots F\}$ sont des vecteurs aléatoires Gaussiens indépendants, dont la distribution dépend de X_t.

Plus spécifiquement, les modèles hiérarchiques considérés supposent deux instances pour les états cachés ($X_t \in \{0, 1\}$), et un modèle conditionnel

$$\underline{\alpha}_t \sim \begin{cases} \mathcal{N}(0, \Sigma) & \text{si } X_t = 1 \\ 0 & \text{sinon} \end{cases}$$

Pour contraindre le modèle, il est nécessaire d'introduire des hypothèses supplémentaires sur la matrice de variance/covariance Σ, permettant la classification. Les hypothèses sur lesquelles nous nous focaliserons ici concernent essentiellement l'amplitude moyenne des termes diagonaux, ainsi que la vitesse de décroissance en dehors de la diagonale. plus précisément, les coefficients étant supposés fortement corrélés, on supposera que Σ décroît lentement quand on s'éloigne de la diagonale. Un cas particulier est celui des signaux pour lesquels on suppose que les coefficients dans un groupe forme une suite stationnaires en moyenne quadratique, ou stationnaire après correction d'une tendance lente.

On définira de même des modèles structurés à fréquence fixée, en supposant la matrice de covariance correspondante $\tilde{\Sigma}$ lentement décroissante en dehors de sa diagonale également.

Modèles multicouches Le modèle que nous considérons est finalement le suivant : le signal est supposé être la somme de deux signaux hiérarchiques structurés, l'un à temps fixé (couche transitoire) et l'autre à fréquence fixée (couche tonale). Notons $U = \{u_{tf}\} = \{\underline{u}_f\}$ et $V = \{v_{tf}\} = \{\underline{v}_t\}$ les deux dictionnaires temps-fréquence correspondants, où $\underline{u}_f$ est le vecteur constitué des vecteurs de base $\{u_{tf}\}$ à temps fixé t (et une notation similaire pour $\underline{v}_t$). On écrira donc le signal modèle sous la forme

$$\begin{aligned} x &= \sum_t X_t \sum_f \alpha_{tf} v_{tf} + \sum_f \tilde{X}_f \sum_t \beta_{tf} \underline{u}_{tf} \\ &= \sum_t X_t \underline{\alpha}_t \cdot \underline{v}_t + \sum_f \tilde{X}_f \underline{\beta}_f \cdot \underline{u}_f \end{aligned} \quad (1)$$

où X et $\tilde{X}$ représentent les cartes de signifiance (binaires) associées aux deux couches. On notera Σ et $\tilde{\Sigma}$ les matrices de variance-covariance des vecteurs $\underline{\alpha}_t$ et $\underline{\beta}_f$, et on se limitera ici au cas de cartes X et $\tilde{X}$ distribuées selon un modèle de Bernoulli. L'objet de ce travail est de prolonger l'analyse de [3] à ce nouveau cadre.

Comportement des coefficients d'analyse Les coefficients d'analyse sont définis comme les produits scalaires du signal avec les atomes du dictionnaire. En notation vectorielle

$$\begin{cases} \underline{a}_t = \langle x, \underline{v}_t \rangle = \underline{\alpha}_t X_t + \sum_{f'} \tilde{X}_{f'} \sum_{t'} \beta_{t'f'} \langle u_{t'f'}, \underline{v}_t \rangle \\ \underline{b}_f = \langle x, \underline{u}_f \rangle = \underline{\beta}_f \tilde{X}_f + \sum_{t'} X_{t'} \sum_{f'} \alpha_{t'f'} \langle v_{t'f'}, \underline{u}_f \rangle \end{cases}$$

Conditionnellement aux deux cartes X et $\tilde{X}$, les vecteurs $\underline{a}$ et $\underline{b}$ sont des vecteurs Gaussiens, dont la matrice de variance/covariance dépend des cartes. Focalisons-nous par exemple sur la première des deux équations ci dessus. Il est clair que la distribution de $\underline{a}_t$ est très significativement conditionnée par la valeur de l'état caché X_t. Plus précisément, conditionnellement à l'état caché X_t, la distribution du vecteur $\underline{a}_t$ est normale, de moyenne nulle, de matrice de covariance notée $\mathcal{C}_{tf;t'f'} = \mathbb{E}\left\{a_{tf} \underline{a}_{t'f'}\right\}$. En se limitant aux termes diagonaux en temps ($t = t'$), la covariance s'écrit

$$\mathbb{E}\left\{a_{tf} \underline{a}_{tf'}\right\} = X_t \Sigma_{ff'} + \left[\Gamma_t(\tilde{X})\right]_{ff'} \ ,$$

où l'on a posé

$$\left[\Gamma_t(\tilde{X})\right]_{ff'} = \sum_\nu \tilde{X}_\nu \sum_{s,s'} \tilde{\Sigma}_{ss'} \langle v_{s\nu}, u_{tf} \rangle \langle u_{tf'}, v_{s'\nu} \rangle \ .$$

La matrice (aléatoire, à travers sa dépendance dans la carte $\tilde{X}$) $\Gamma_t(\tilde{X})$ représente la contribution de la seconde couche à la covariance de la première. Le point essentiel est d'être capable de discriminer les cas $X_t = 0$ et $X_t = 1$. Dans le cas $X_t = 0$, seule la seconde composante est présente, et la covariance Γ_t est significativement différente de Σ. En effet, Γ_t est diagonale dominante, alors que Σ est supposée décroître lentement quand on s'éloigne de la diagonale.

Remarque. Le cas particulier où les deux bases sont identiques est assez instructif. En effet, dans ce cas, le calcul donne

$$\Gamma_t(\tilde{X}) = \Sigma_{tt} \operatorname{diag}(\tilde{X}) \ ,$$

i.e. une matrice de covariance diagonale. Dans le cas $U \neq V$, Γ_t n'est plus diagonale, mais décroît rapidement lorsque l'on s'éloigne de la diagonale. Ce comportement est radicalement différent du comportement supposé de Σ, supposée décroître lentement hors de la diagonale.

Une analyse en tout point similaire peut être faite à partir des vecteurs $\underline{b}_f$. L'identification des valeurs des états cachés $\tilde{X}_f$ s'effectue via l'analyse de la matrice de variance/covariance des vecteurs de coefficients d'analyse $\underline{b}_f$.

2.2 Estimation

Compte tenu des hypothèses de stationnarité, nous optons pour une estimation d'états cachés basée sur la distribution des vecteurs de coefficients, en nous limitant pour simplifier au cas de la détection de transitoires.

2.2.1 Classification par EM ou CEM

Considérant toujours le cas des lignes de coefficients à indice temporel fixé, nous sommes donc conformément à notre modèle dans une situation approximable comme mélange de lois normales multivariées. Suivant la stratégie de [3], on approchera la distribution des coefficients d'analyse $\{a_{tf}\}$ par un mélange pondéré de deux lois normales multivariées de moyenne nulle. L'apprentissage et la classification peuvent être effectués via des stratégies de type EM ou CEM (EM conjoint avec classification, voir [1]).

Nous donnons en figure 1 les matrices de variance/covariance des vecteurs de coefficients dans les deux classes, estimées par CEM, dans l'exemple du signal *mamavatu* (voir la section 3). Les caractéristiques premières que l'on y voit sont les suivantes

- Une décroissance en fonction de la fréquence, comme attendu ; la partie tonale est dominée par les basses fréquences, ce qui est aussi attendu.
- La décroissance en dehors de la diagonale est bien plus marquée pour les coefficients « transitoires » (c'est à dire $X_t = 1$), ce qui montre l'existence de corrélations significatives à travers les fréquences.

Faisant abstraction de cette décroissance en fonction de la fréquence, la différence « morphologique » entre ces deux comportements est suffisamment marquée pour qu'un algorithme de classification soit capable d'effectuer une discrimination stable. Dans les résultats numériques présentés plus loin en section 3, nous utiliserons un algorithme CEM, initialisé avec une classification préalable effectuée à l'aide du critère $Z_p^{(n)}$ ci-dessous.

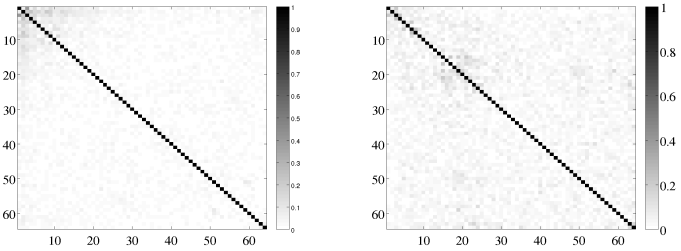

FIG. 1 – Matrices de corrélation des $\{\underline{a}_t\}$ estimés par CEM : parties non-transitoires (gauche) et transitoires (droite).

2.2.2 Classification par critère de parcimonie

La détection des « lignes transitoires » peut également être effectuée via une décision basée sur une statistique test bien adaptée. Compte tenu des caractéristiques supposées des distributions des vecteurs de coefficients dans les cas $X_t = 0$ et $X_t = 1$, des critères basés sur des mesures de diversité sont des candidats naturels. Par exemple, des normes ℓ_p de vecteurs de coefficients $\mathbf{a}_t = \{a_{tf}, f = 1, \dots F\}$ de la forme

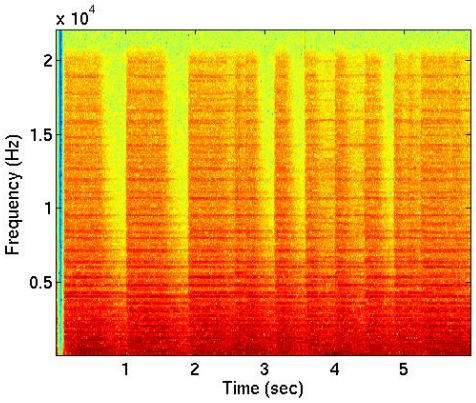

FIG. 2 – Coefficients MDCT d'un signal d'orgue.

$$Z_p(t) = \left(\sum_f |a_{tf}|^p \right)^{1/p} \qquad Z_p^{(n)}(t) = \left(\sum_f \left| \frac{a_{tf}}{\|\mathbf{a}_t\|_2} \right|^p \right)^{1/p},$$

ou des puissances de ces quantités. La fonction Z_p, pour $p < 2$, est une mesure « traditionnelle » de diversité, couramment utilisées dans les algorithmes de régression parcimonieuse. Elle présente le défaut de dépendre linéairement de la normalisation globale du signal. Pour s'abstraire de cet effet indésirable, on lui préfèrera ici la version normalisée, i.e. la fonction $Z_p^{(n)}$. La différence est illustrée ci-dessous par la détection de transitoires dans un signal d'orgue, riche harmoniquement et présentant des attaques bien marquées, comme on peut le voir en FIG. 2. On représente en FIG. 3, le signal original, ainsi que les fonctions Z_p et $Z_p^{(n)}$ calculées ici avec $p = 0.5$. On peut voir clairement (en particulier dans les premiers accords) que Z_p est très sensible à l'énergie locale du signal, et ne permet donc pas une détection précise des transitoires. $Z_p^{(n)}$ permet de se focaliser davantage sur les transitoires.

2.2.3 Algorithme de classification

La classification nous permet d'estimer les états cachés des coefficients, c'est-à-dire la carte de significance. On détaille ici l'algorithme utilisé pour l'estimation de la couche transitoire. Soit un signal x de la forme (1), pour lequel on a fixé une base $V = \{v_{tf}\}$ adaptée à la couche transitoire.

L'algorithme utilisé au final peut se résumer ainsi :

1. Calcul des coefficients d'analyse $a_{tf} = \langle x, v_{tf} \rangle$ et calcul des quantités $Z_p^{(n)}(t)$, avec $p \leq 1$.
2. Initialisation de la carte de signifiance par seuillage : si $Z_p^{(n)}(t) > \tau$ alors $\hat{X}_t = 1$, et $\hat{X}_t = 0$ sinon, où $\tau \in \mathbb{R}_+$ est un seuil fixé.
3. Classification des coefficients d'analyse par CEM, initialisé avec la classification précédente :

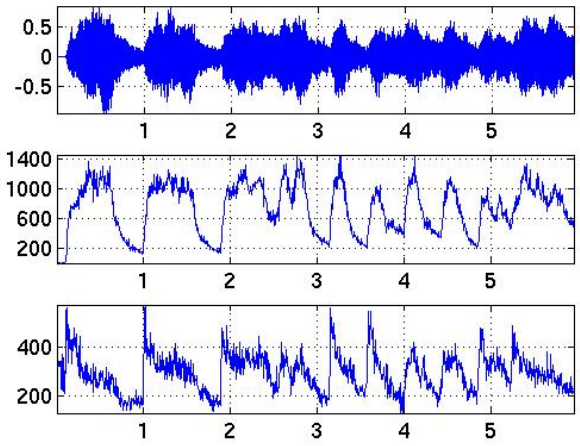

FIG. 3 – Signal d'orgue (haut), et indices Z_p (milieu) et $Z_p^{(n)}$ (bas).

- Estimation des poids du mélange par (notés respectivement p_1 et p_2) en calculant les proportions respective de $\hat{X}_t = 1$ et $\hat{X}_t = 0$.
- Estimation des matrices de covariance des $\{a_{tf}, X_t = 1\}$ et $\{a_{tf}, X_t = 0\}$ (notées respectivement Σ_1 et Σ_2)

(a) étape E : calcul des probabilités d'appartenance des a_{tf} à chacune des loi normales pondérées.

(b) étape M par classification : $\hat{X}_t = 1$ si $\mathbb{P}\{a_{tf} \sim p_1\mathcal{N}(0,\Sigma_1)\} > \mathbb{P}\{a_{tf} \sim p_2\mathcal{N}(0,\Sigma_2)\}$, puis estimation des poids et des matrices de covariance.

On obtient en sortie une estimation de la carte de signifiance de la couche transitoire. Une première estimation simple de cette dernière est alors donnée par $x_{trans} = \sum_t \hat{X}_t \underline{a}_t \cdot \underline{v}_t$.

3 Application, résultats numériques

Nous illustrons l'approche proposée avec un extrait de *mamavatu* de Susheela Raman. Le signal (échantillonné à 44.1 kHz dure environ 12 s, *i.e.* 2^{19} échantillons, voir FIG. 4), est formé d'un mélange de percussions et de guitare. Notre but est ici d'estimer la couche transitoire de ce signal. On choisit pour celà une base MDCT (Modified Discrete Cosine Transform) avec une fenêtre d'analyse bien localisée en temps (d'une durée de 3 ms environ, soit 128 échantillons), bien adaptée à l'estimation de transitoires. L'algorithme décrit en section 2.2.3 est alors appliquée, et l'estimation de la partie transitoire estimée est illustrée sur la figure 5.

L'algorithme de classification utilisée est très rapide : environ 2 secondes sous matlab tournant sur un PC QuadCore à 2.4 GHz (un seul coeur est utilisé). Les expériences supplémentaires menées montrent que l'algorithme est assez stable malgrés le nombre de minima locaux que peut rencontrer CEM. Les différentes initialisations utilisées donnent très souvent le même résultat ; et lorsqu'une mauvaise initialisation est choisie, le résultats renvoyé n'est pas réaliste.

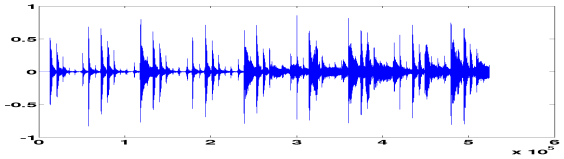

FIG. 4 – Échantillons du signal *mamavatu*.

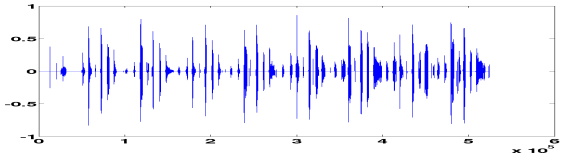

FIG. 5 – Couche transitoire estimée du signal *mamavatu*.

Il ressort des expériences que la partie transitoire est bien estimée sur des signaux comportant des percussions (et donc des transitoires rapides et bien « marqués »). Sur des signaux de nature plus « tonale », les transitoires estimés sont souvent «trop longs» par rapport au résultat attendu. Une origine vraisemblable de ce phénomène est une plus grande difficulté à différencier les matrices de variance-covariance.

4 Conclusion, perspectives

L'originalité de l'approche proposée ici est de prendre en compte la structure hiérarchique sur les coefficients à la fois par les états caché, mais aussi par la matrice de variance/covariance des coefficients de synthèse, et qu'on retrouve dans les coefficients d'analyse.

Une perspective envisagée, compte tenue des hypothèses de stationarité, est de construire des estimateurs sur les spectres des coefficients $\underline{a}_t$, après renormalisation permettant de corriger la décroissance de leurs valeurs en fréquence.

Références

[1] G. Govaert and G. Celeux. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14(3) :315–332, 1992.

[2] M. Kowalski. Sparse regression using mixed norms. *Applied and Computational Harmonic Analysis*, 2009. Doi : 10.1016/j.acha.2009.05.006.

[3] M. Kowalski and B. Torrésani. Random models for sparse signals expansion on unions of bases with application to audio signals. *IEEE Transactions on Signal Processing*, 56(8) :3468–3481, 2008.

[4] S. Molla and B. Torrésani. An hybrid audio scheme using hidden Markov models of waveforms. *Applied and Computational Harmonic Analysis*, 18(2) :137–166, 2005.