

Modèles de Markov cachés pour l'estimation de plusieurs fréquences fondamentales

Francis Bach*, Michael I. Jordan**

* Centre de Morphologie Mathématique
Ecole des Mines de Paris
35, rue Saint Honoré
77305 Fontainebleau, France
francis.bach@mines.org

** Computer Science Division
and Department of Statistics
University of California
Berkeley, CA 94720, USA
jordan@cs.berkeley.edu

1 Introduction

Le suivi de la fréquence fondamentale est un problème important du traitement de la parole et de la musique, et le développement d'algorithmes robustes pour la détermination d'une ou plusieurs fréquences fondamentales est un sujet actif de recherches en traitement du signal acoustique (Gold et Morgan, 1999). La plupart des algorithmes d'extraction de la fréquence fondamentale commencent par construire un ensemble de caractéristiques non linéaires (comme le corrélogramme ou le "cepstrum") qui ont un comportement spécial lorsqu'une voyelle est prononcée. Ensuite, ces algorithmes modélisent ce comportement afin d'extraire la fréquence fondamentale. En présence de plusieurs signaux mixés additivement, il est naturel de vouloir modéliser directement le signal ou une représentation linéaire de ce signal (comme le spectrogramme), afin de préserver l'additivité et de rendre possible l'utilisation de modèles destinés à une seule fréquence fondamentale pour en extraire plusieurs.

L'utilisation directe du spectrogramme nécessite cependant un modèle probabiliste détaillé afin de caractériser la fréquence fondamentale. Dans cet article, nous considérons une variante de modèle de Markov caché et utilisons le cadre des modèles graphiques afin de construire le modèle, apprendre les paramètres à partir de données et développer des algorithmes efficaces d'inférence. En particulier, nous utilisons des développements récents en apprentissage automatique (machine learning) pour caractériser les propriétés adéquates des signaux de parole et de musique; nous utilisons des probabilités *a priori* non-paramétriques afin de caractériser la régularité de l'enveloppe spectrale et nous améliorons la procédure d'apprentissage grâce à l'apprentissage discriminatif du modèle.

2 Modèle graphique pour l'extraction de la fréquence fondamentale

Dans cet article, nous utilisons des signaux sonores échantillonnés à 5.5 KHz. Etant donné un signal uni-dimensionnel x_t , $t = 1, \dots, T$, le *spectrogramme* s est défini comme la transformée de Fourier à fenêtres de x ; i.e., le signal x est découpé en N morceaux de longueur M qui se recouvrent, et le spectrogramme s est défini comme la matrice $N \times P$ dont la n -ième colonne $s_n \in \mathbb{R}^P$ est la FFT à P points du n -ième morceaux. Dans cet article nous modélisons le module du spectrogramme et référons à ce module du spectrogramme simplement comme le spectrogramme. Comme les signaux sonores sont réels, la FFT est symétrique et nous utilisons seulement les $P/2$ premières fréquences.

2.1 Modèle additif

La variable d'entrée de notre modèle de recherche de fréquence fondamentale est la suite $s_n \in \mathbb{R}^P$, $n = 1, \dots, N$, où N est le nombre de fenêtres, égal à une constante fois la durée T du signal x . Nous utilisons un modèle additif du spectrogramme, i.e., si K personnes sont présentes, nous modélisons la n -ième fenêtre comme la superposition des K signaux $u_n^k \in \mathbb{R}^P$ plus du bruit, i.e., $s_n = \sum_{k=1}^K u_n^k + \varepsilon_n$.

2.2 Modèle harmonique

Nous utilisons un modèle harmonique dans le domaine des fréquences, ce qui correspond à modéliser le spectrogramme comme un peigne de Dirac avec modulation d'amplitude. Nous modélisons chaque personne k à l'instant n à l'aide de quatre variables

- *Voyelles* : v_n^k est une variable binaire qui est égale à un si la personne k prononce une voyelle à l'instant n , et égale à zéro sinon.
- *Fréquence fondamentale* : ω_n^k est la fréquence fondamentale, définie de telle sorte qu'elle est égale à la distance entre deux pics dans le spectrogramme.
- *Harmoniques* : h_n^k est un ensemble de vecteurs d'amplitudes harmoniques lorsque $v_n^k = 1$. Il y a un vecteur $h_{n\omega}^k$ pour chaque valeur ω . La dimension de $h_{n\omega}^k$ est égale à $\lfloor P/2\omega \rfloor$.
- *Terme constant* : c_n^k est l'amplitude constante des portions sans voyelle ($v_n^k = 1$).

Le modèle graphique pour une personne est un simple modèle de Markov caché (voir figure 1). Les probabilités conditionnelles, qui sont nécessaires pour définir complètement le modèle, reflètent les propriétés psycho-acoustiques et statistiques connues des fréquences fondamentales (Gold et Morgan, 1999). En particulier, la propriété de régularité de l'*enveloppe spectrale* est explicitement prise en compte à l'aide d'outils de statistique non paramétrique. Pour plus de détails, voir Bach et Jordan (2005).

2.3 Modèle de Markov cachés factoriel

Les modèles pour chaque personne peuvent être combinés en un unique modèle graphique, un modèle de Markov caché "factoriel", où les $2K$ chaînes de Markov évoluent indépendamment (voir figure 1 pour deux personnes). Le paramètre λ_n est la variance

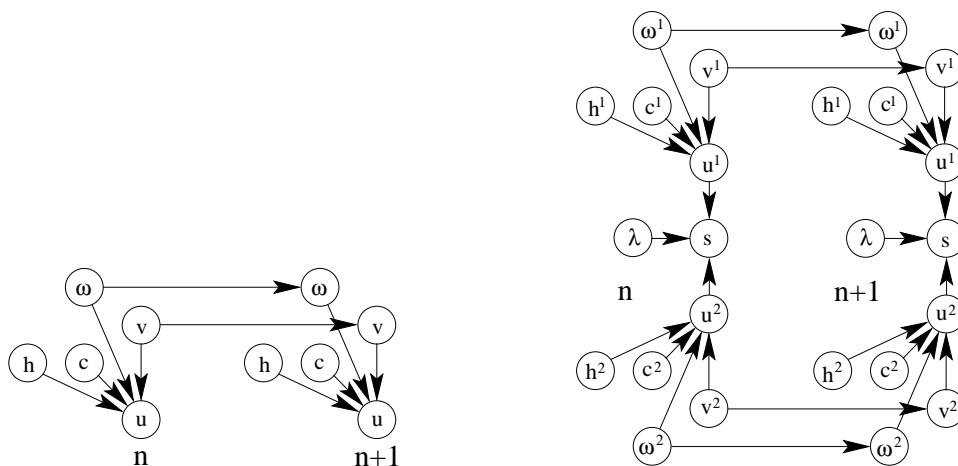


FIG. 1 – (Gauche) Modèle pour une personne pour deux fenêtres n et $n+1$. (Droite) Modèle pour deux personnes pour deux fenêtres n et $n+1$. les indices de temps sont omis.

du bruit Gaussien ε_n au temps n . Nous faisons l'hypothèse que λ_n a une distribution uniforme et est discrétisée sur une grille logarithmique à $n_\lambda = 10$ éléments.

3 Inférence

Dans les sections suivantes, nous utilisons le raccourci x pour dénoter l'ensemble des variables $(x_n^k)_{k,n}$ pour tous k et n , et le raccourci x^k pour dénoter l'ensemble des variables $(x_n^k)_n$ pour tous n . Si nous définissons $z = (\omega, v, h, c, \lambda)$, alors la tâche d'inférence est de calculer, étant donné s , $\arg \max_z p(z|s)$. La minimisation par rapport à (h, λ) peut s'effectuer en formule fermée, donc nous sommes ramenés à l'optimisation par rapport à (ω, v) .

Avec une personne, nous sommes face à l'inférence classique dans un modèle de Markov caché, où le nombre d'états cachés est proportionnel à n_ω , et la complexité pour un signal de durée T est $O(Tn_\omega^2)$ pour l'algorithme de Viterbi (Ghahramani et Jordan, 1997). Avec m personnes, nous avons un modèle factoriel avec $2m$ chaînes découplées, chacune avec n_ω ou 2 états; la complexité d'un algorithme de Viterbi structuré est alors $O(Tn_\omega^{m+1})$ (Ghahramani et Jordan, 1997). Comme n_ω est grand, nous utilisons la procédure d'approximation suivante :

1. Estimer récursivement les m fréquences en estimant une fréquence à la fois et soustrayant le modèle harmonique correspondant.
2. Construire un ensemble de p_ω candidats pour la fréquence, pour chaque instant, constitué des minima locaux dans chaque algorithme de Viterbi de l'étape 1.
3. Effectuer l'inférence exacte en utilisant le petit nombre de candidats.

La complexité de l'algorithme précédent est $O(mn_\omega^2 T + Tp_\omega^{m+1})$. En pratique p_ω est choisi assez petit (autour de 10), de telle sorte que la complexité est $O(mn_\omega^2 T)$, mais assez grand pour que l'approximation par rapport à l'inférence exacte soit minime (i.e., très peu de différences avec $p_\omega = n_\omega$).

4 Apprentissage des paramètres

Si z est défini comme $z = (\omega, v)$, alors notre modèle pour s est un modèle avec variable latente z . En présence de données déjà annotées, pour lesquelles à la fois s et z sont disponibles, nous pouvons utiliser un apprentissage discriminatif, i.e., nous optimisons la vraisemblance conditionnelle $p(z|s)$ au lieu de la vraisemblance jointe $p(s, z)$ (Bach et Jordan, 2005).

Nous utilisons la base de données annotées de Keele (Plante et al., 1995), qui est composée de 10 personnes différentes, enregistrées séparément. Nous pouvons mixer artificiellement les enregistrements de deux personnes différentes pour obtenir des données mixées.

5 Conclusion

Nous avons présenté un algorithme pour l'extraction de plusieurs fréquences fondamentales, à base de modèles graphiques. L'utilisation de probabilités *a priori* appropriées et d'apprentissage discriminatif permet d'obtenir une meilleure performance. Le temps de calcul de notre algorithme est fonction linéaire de la durée du signal initial. Bien que notre implémentation en Matlab soit 10 fois plus lente que le temps réel, il n'y a pas d'obstacles majeurs pour une implémentation en temps réel de notre algorithme.

Références

- Bach, F. R. et Jordan, M. I. (2005). Discriminative training of hidden Markov models for multiple pitch tracking. In *ICASSP*.
- Ghahramani, Z. et Jordan, M. I. (1997). Factorial hidden Markov models. *Machine Learning*, 29 :245–273.
- Gold, B. et Morgan, N. (1999). *Speech and Audio Signal Processing*. Wiley Press.
- Plante, F., Meyer, G. F., et Ainsworth, W. A. (1995). A pitch extraction reference database. In *Proc. EUROSPEECH*.