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

Francis Bach*, Michael I. Jordan**

* Computer Science Division
University of California
Berkeley, CA 94720, USA
fbach@cs.berkeley.edu
http://www.cs.berkeley.edu/~fbach

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

Résumé. Un algorithme d'estimation de la fréquence fondamentale de signaux sonores est introduit: il utilise une modélisation du spectrogramme du signal à l'aide d'un modèle de Markov caché factoriel, dont les paramètres sont estimés de manière discriminative à partir de la base de données de Keele (Plante et al., 1995). Les algorithmes présentés permettent de suivre plusieurs fréquences fondamentales et de déterminer le nombre de fréquences présentes à chaque instant. Les résultats de simulations, effectuées sur des mélanges de signaux de parole et du bruit, illustrent la robustesse de l'approche présentée.

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, McAulay et Quatieri, 1990, Wu et al., 2003, Li et al., 2004, Walmsley et al., 1999, Tabrikian et al., 2004). 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. Dans cet article, nous utilisons le module du spectrogramme; ce module n'est pas une représentation linéaire du signal de départ mais grâce à la parcimonie des signaux de pa-

role et de musique (Yilmaz et Rickard, 2004), notre représentation peut être considérée de manière approximative comme linéaire.

L'utilisation directe du spectrogramme nécéssite cependant un modèle probabilitiste 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 développments récents en apprentissage automatique (machine learning) pour caractériser les propriétés adéquates des signaux de parole et de musique; nous utilisons des probabilités antérieures 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 (Lafferty et al., 2001). Le modèle graphique est introduit en section 2, l'algorithme d'inférence en section 3 et l'agorithme d'apprentissage en section 4. En section 5, nous testons nos algorithmes sur une palette de tâches difficiles d'extraction de fréquences fondamentales.

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, $t = 1, \ldots, T$, 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 spectogramme s est défini comme la matrice $N \times P$ dont la n-ième colonne $s_n \in \mathbb{R}^P$ est la FTT à P points du n-ième morceaux.[1] 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 symmétrique 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 $s_n \in \mathbb{R}^P$, $n = 1, \ldots, N$, 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., $s_n = \sum_{k=1}^{K} u_n^k + \varepsilon_n$. Il est important de noter que l'acoustique n'est pas additive pour le module du spectrogramme, mais comme les signaux correspondant à deux personnes différentes exhibent peu de recouvrement (Yilmaz et Rickard, 2004), la linéarité est une approximmation raisonnable. L'avantage du module est que la modélisation de la régularité de l'enveloppe spectrale est simple à accomplir à l'aide de techniques de régularisation à base de splines (voir section 2.2).

1. Pour les simulations, nous usilisons des fenêtres de temps 40ms prises toutes les 10ms, et une FFT à 512 points.

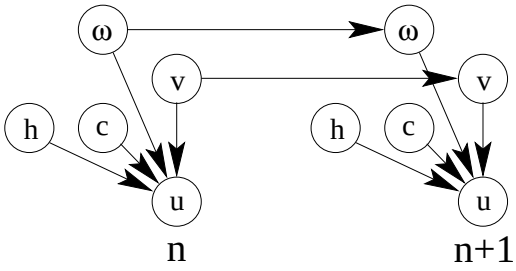

FIG. 1 – *Modèle pour une personne pour deux fenêtres n et $n+1$ (les indices de temps sont omis).*

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 (McAulay et Quatieri, 1990). 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 (Bregman, 1990, Gold et Morgan, 1999, McAulay et Quatieri, 1990):

- $p(v_{n+1}^k | v_n^k)$ est une matrice de transition T_v constante avec quatre éléments.
- $p(\omega_{n+1}^k | \omega_n^k)$: la fréquence est discrétisée sur une grille avec $n_\omega = 300$ éléments, et le logarithme d'une ligne de la matrice de transition est égale à (à une constante additive près) $\alpha_1(\omega_{n+1}^k - \omega_n^k)^2 + \alpha_2 \omega_{n+1}^k + \alpha_3 (\omega_{n+1}^k)^{-1}$.
- $p(h_{n\omega}^k)$: pour chaque valeur ω, $h_{n\omega}^k$ est modélisé comme la restriction d'une fonction régulière sur [0,P/2]—i.e., une fonction dont la dérivée seconde est bornée—aux multiples de ω. C'est à dire, $(h_{n\omega}^k)_i$ est égal à $g(i\omega)$, où g est une fonction telle que $\int |g^{(2)}|^2$ est borné. La fonction g est appelée l'*enveloppe spectrale* (McAulay et Quatieri, 1990).

  Il en découle (Wahba, 1990) que $h_{n\omega}^k$ peut être modélisé comme un processus Gaussien sur l'intervalle [0,P/2], observé aus multiples de la fréquence fondamentale ω; ceci implique que $h_{n\omega}^k$ s'exprime comme $h_{n\omega}^k = K_\omega a_{n\omega}^k + T_\omega b_{n\omega}^k$, où K_ω est

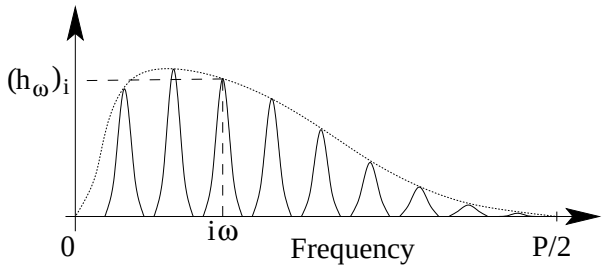

FIG. 2 – *Enveloppe spectrale (pointillé) et modèle harmonique (plein).*

la "matrice de noyau" définie comme $(K_\omega)_{ij} = (\frac{1}{2}ij\min\{i,j\} - \frac{1}{6}\min\{i,j\}^3)\omega^3$, et T_ω est une matrice avec deux colonnes, une constante et une fonction linéaire de la fréquence. Les variables auxilliaires $a^k_{n\omega}$ et $b^k_{n\omega}$ sont normales avec espérance 0 et matrice de covariance $(\alpha_4 K_\omega + \alpha_5 I)^{-1}$ et $\alpha_6 I$.

- La variable c^k_n est normale avec espérance α_4 et variance α_5.
- Modèle d'observation: sachant ω^k_n, h^k_n, c^k_n et v^k_n, le signal u^k_n est égal à $B(\omega^k_n)h^k_{n\omega^k_n}$ si $v^k_n = 1$, et égal à $c^k_n e$ si $v^k_n = 0$, où e est le vecteur constant égal à un. La i-ième colonne de la matrice $B(\omega)$ est une "cloche" centrée à la fréquence $i\omega$. Voir figure 2. Ainsi, le spectrogramme est modélisé comme une somme pondérée de cloches positionnées aux multiples de ω, dont les amplitudes sont les valeurs de l'enveloppe specrale.

2.3 Modèles de Markov cachés factoriels

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 3 pour deux personnes). Le paramètre λ_n est la variance 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. Notre modèle est similaires à certains modèles déjà proposés. Pour une comparaison détaillée, voir Bach et Jordan (2005).

3 Extraction de la fréquence fondamentale

Dans les sections suivantes, nous utilisons le raccourci x pour dénoter l'ensemble les variables $(x^k_n)_{k,n}$ pour tous k et n, et le raccourci x^k pour dénoter l'ensemble des variables $(x^k_n)_n$ 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).

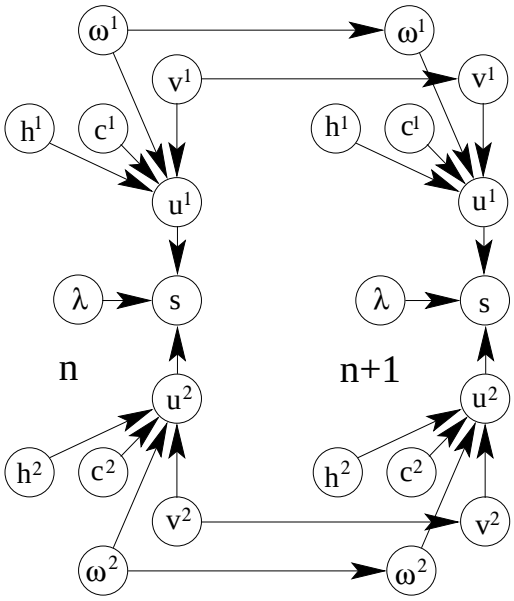

FIG. 3 – *Modèle pour deux personnes pour deux fenêtres n et $n+1$ (les indices de temps sont omis).*

3.1 Une personne

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'agorithme de Viterbi (Ghahramani et Jordan, 1997).

3.2 Deux personnes et plus

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'approximmation suivante, qui généralise les techniques de Tabrikian et al. (2004):

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 minimas 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 précédant est $O(mn_\omega^2 T + Tp_\omega^m)$. 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 paramêtres

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, il y a deux types d'apprentissage à l'aide du maximum de vraisemblance, la méthode générative et la méthode discriminative.

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 enregistrement de deux personnes différentes pour obtenir des données mixées. Avec deux personnes, nous avons donc deux ensembles de variables latentes, (ω^1,v^1), (ω^2,v^2), un par personne.

4.1 Apprentissage génératif

Dans ce type d'estimation, nous maximisons simplement la vraisemblance jointe. Comme dans tous les modèles graphiques avec graphe dirigé, la maximisation de la vraisemblance se découple en maximisations indépendantes pour chaque paramêtre.

Cette méthode est très efficace (aucune inférence dans un modèle de Markoc caché n'est nécessaire). Cependant, le but final du modèle est de prédire z à partir de s, non pas d'obtenir un modèle joint de s et z. Il est en général beaucoup plus performant d'optimiser la vraisemblance conditionelle $p(z|s)$ au lieu de la vraisemblance jointe $p(s,z)$ (Bahl et al., 1986, Lafferty et al., 2001), comme présenté dans la section suivante.

4.2 Apprentissage discriminatif

Au lieu de maximiser $p(s,z)$, nous maximisons $p(z|s)$. Maximiser la vraisemblance conditionnelle ne se découple pas en général pour les modèles graphiques et maximiser la vraisemblance nécessite donc d'effectuer plusieurs procédures d'inférence dans un modèle factoriel. Comme l'inférence exacte est trop coûteuse pour ces modèles factoriels, nous maximisons une "pseudo log vraisemblance", qui est définie comme la somme des log vraisemblances de sous-problèmes et a des propriétés asymptotiques similaires à la log vraisemblance classique (Liang et Yu, 2003). Nous définissons ainsi la pseudo log vraisemblance comme ceci: les données disponibles sont $(\omega^1,v^1,\omega^2,v^2)$; soit

$$q(\omega,\omega',v,v') = \max_{h^1,h^2,c^1,c^2,\lambda} p(\omega,\omega',v,v',h^1,h^2,c^1,c^2,\lambda).$$

Nous maximisons par rapport aux paramêtres la log vraisemblance définie comme:

$$\log \frac{q(\omega^1,\omega^2,v^1,v^2)}{\sum_{\omega,v} q(\omega,\omega^2,v,v^2)} + \log \frac{q(\omega^1,\omega^2,v^1,v^2)}{\sum_{\omega,v} q(\omega^1,\omega,v^1,v)}$$

La maximisation est effectuée par descente de gradient et nécessite d'effectuer l'inférence dans un modèle de Markov caché avec un nombre d'états proportionnel à n_ω (et non pas n_ω^2).

5 Simulations

Dans cette section, nous illustrons comment les différentes caractéristiques de notre modèle permettent d'obtenir une performance robuste. Dans toutes nos simulations, l'apprentissage a été fait sur les six premières personnes de la base de données, tandis que les tests sont opérés sur les quatre dernières personnes.

La mesure utilisée pour comparer deux fréquences ω et ω' est $d(\omega,\omega') = 1 - e^{-(\omega-\omega')^2/\sigma^2}$, où σ^2 est la variance empirique de la fréquence sur l'ensemble d'apprentissage. Cette mesure est équivalente à la distance Euclidienne au carré pour des fréquences proches et tend vers 1 pour des fréquences éloignées. Avec cette mesure, si la fréquence estimée est très distante de la vraie fréquence, la pénalité est bornée par un.

Le temps de calcul pour extraire la ou les fréquences fondamentales est fonction linéaire de la durée des signaux. Avec l'implémentation actuelle en Matlab, avec un processeur séquencé à 2GHz, le temps de calcul est égal à 30 fois la durée du signal pour extraire une fréquence, et 130 fois pour extraire deux fréquences.

5.1 Effet de la probabilité antérieure de régularité

Pour la tâche simple d'extraction d'une fréquence fondamentale à partir de fenêtres indépendantes (i.e., en ignorant la chaîne de Markov), nous avons comparé notre approche à une approche sans probabilité antérieure de régularité: avec la probabilité antérieure de régularité, l'erreur moyenne (comme définie ci-dessus), est égale à 0,28, alors que l'erreur sans probabilité antérieure de régularité est égale à 0,57; et dans la plupart des erreurs additionnelles la fréquence estimée est exactement la moitié de la vraie fréquence, ce qui constitue un problème classique des algorithmes d'estimation de la fréquence fondamentale. Il a été montré que pour les modèles harmoniques comme celui que nous avons présenté, la connaissance détaillée de l'enveloppe spectrale permet d'éviter ce problème (McAulay et Quatieri, 1990); les résultats que nous avons présentés suggèrent qu'une simple probabilité antérieure de régularité, qui ne nécessite pas de connaître exactement l'enveloppe spectrale, est suffisant.

5.2 Effet du type d'apprentissage

Nous avons comparé les performances des apprentissages discriminatifs et génératifs sur des tâches d'extraction d'une seule fréquence fondamentale. Le modèle estimé génèrativement a un taux d'erreur de 27,4% pour la décision de présence d'une voyelle (i.e., prédiction correcte de v_n) et une erreur pour l'estimation de la fréquence de 0,022; alors que pour le modéle estimé discrimativement, les erreurs sont seulement 5% et 0,016. L'apprentissage discriminatif permet donc d'obtenir une meilleure performance.

5.3 Deux fréquences fondamentales

Dans cette série d'expériences, nous mixons deux signaux de même énergie, provenant de personnes différentes. La performance pour des personnes de même sexe ou de sexe différent est présentée en figure 4.

	présence de voyelles	fréquence
homme - femme	22%	0,03
femme - femme	32%	0,08
homme - homme	31%	0,07

FIG. 4 – *Extraction de deux fréquences fondamentales: taux d'erreur pour la décision de présence de voyelle et erreur d'estimation des fréquences.*

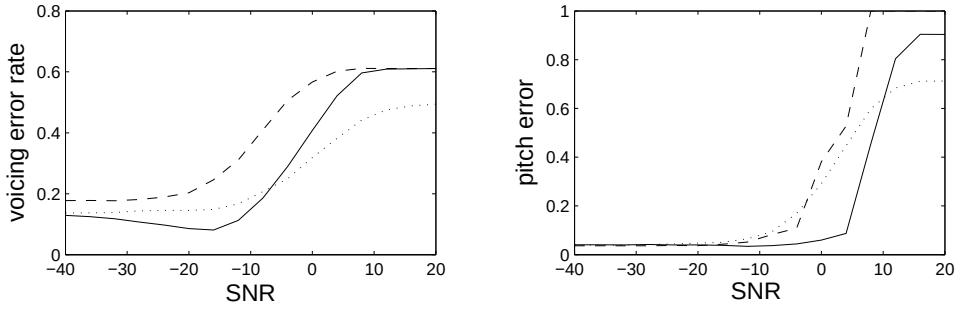

FIG. 5 – *Estimation d'une fréquence fondamentale en présence de bruit: taux d'erreur pour la décision de présence de voyelle (gauche) et erreur moyenne d'estimation de la fréquence (droite); bruit blanc (plein), bruit stationnaire coloré (dashed), bruit de fond de restaurant (pointillé).*

5.4 Estimation en présence de bruit

Dans cette série d'expériences, nous avons ajouté différents types de bruit au signal: un bruit blanc, un bruit stationnaire coloré et un bruit de fond de restaurant. Nous montrons la performance d'estimation en fonction du rapport signal sur bruit en figure 5, illustrant la robustesse au bruit de notre approche.

6 Conclusion

Nous avons présenté un algorithme pour l'extraction de plusieurs fréquences fondamentales, à base de modèles graphiques. L'utilisation de probabilités antérieures 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 130 fois plus lente que le temps réel, il n'y pas d'obstacles majeurs pour une implémentation en temps réel de notre algorithme.

Références

F. R. Bach et M. I. Jordan. Discriminative training of hidden Markov models for multiple pitch tracking. In *ICASSP*, 2005.

L. R. Bahl, P. V. de Souza P. F. Brown et, et R. L. Mercer. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In *Proc. ICASSP*, 1986.

A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, 1990.

Z. Ghahramani et M. I. Jordan. Factorial hidden Markov models. *Machine Learning*, 29:245–273, 1997.

B. Gold et N. Morgan. *Speech et Audio Signal Processing*. Wiley Press, 1999.

J. Lafferty, A. McCallum, et F. Pereira. Conditional random fields: Probabilistic models for segmenting et labeling sequence data. In *Proc. ICML*, 2001.

X. Li, J. Malkin, et J. Bilmes. Graphical model approach to pitch tracking. In *Intl. Conf. Spoken Lang. Proc.*, 2004.

G. Liang et B. Yu. Maximum pseudo likelihood estimation in network tomography. *IEEE Trans. Sig. Proc.*, 51(8):2043–2053, 2003.

R. J. McAulay et T. F. Quatieri. Pitch estimation et voicing detection based on a sinusoidal speech model. In *Proc. ICASSP*, 1990.

F. Plante, G. F. Meyer, et W. A. Ainsworth. A pitch extraction reference database. In *Proc. EUROSPEECH*, 1995.

J. Tabrikian, S. Dubnov, et Y. Dickalov. Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model. *IEEE Trans. Speech Audio Proc.*, 12(1), 2004.

G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.

P. J. Walmsley, S. J. Godsill, et P. J. W. Rayner. Polyphonic pitch tracking using joint Bayesian estimation of multiple frame parameters. In *Proc. IEEE Work. App. Sig. Proc. Acoust.*, 1999.

M. Wu, D. Wang, et G. J. Brown. A multipitch tracking algorithm for noisy speech. *IEEE Trans. Speech Audio Proc.*, 11(3), 2003.

O. Yilmaz et S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. Sig. Proc.*, 52(7):1830–1847, 2004.

Summary

We present a multiple pitch tracking algorithm that is based on direct probabilistic modeling of the spectrogram of the signal. The model is a factorial hidden Markov model whose parameters are learned discriminatively from the Keele pitch database (Plante et al., 1995). Our algorithm can track several pitches and determines the number of pitches that are active at any given time. We present simulation results on mixtures of several speech signals and noise, showing the robustness of our approach.