Le modèle harmonique stochastique et son application au rehaussement de signaux de parole

Eric GRIVEL[+], André FERRARI[†] et Olivier CAPPÉ[‡]

[+]ESI-ENSERB, BP 99, 33402 Talence Cedex
[†]UMR 6525 Astrophysique, Université de Nice, Parc Valrose, 06108 Nice Cedex
[‡]CNRS/ENST, Département TSI, 46 rue Barrault, 75634 Paris Cedex 13, France

eric.grivel@tsi.u-bordeaux.fr, ferrari@unice.fr, cappe@tsi.enst.fr

Résumé : Cette contribution est consacrée à l'étude d'un modèle localement périodique dans lequel les évolutions temporelles des paramètres sont modélisées par des processus aléatoires. L'originalité du travail réside dans le modèle considéré qui permet de représenter, de manière paramétrique mais relativement flexible des signaux localement périodiques. Une évaluation de différentes méthodes d'estimation des paramètres (EM, Whittle) et une propositon de solution sous optimale moins coûteuse à implémenter applicables dans le cas de signaux de parole sont en outre proposées pour rehausser le signal par filtrage de Kalman.
Abstract :This contribution deals with a stochastically modulated periodic model and its use for speech enhancement. The originality of our work stands in the stochastic modeling which offers a flexible way to represent quasi periodic signals. Various methods (EM, Whittle) are reviewed to estimate the model parameters and a suboptimal solution is proposed with a reduced computation load to perform a Kalman filter based speech enhancement.

1 Introduction

Deux catégories de modèles sont généralement envisagées dans le traitement de la parole : le modèle autorégressif (AR) tout particulièrement utilisé pour les codeurs CELP et les modèles sinusoïdaux ou harmoniques. Dans de nombreux travaux récents ces deux approches sont combinées en utilisant par exemple une décision de voisement.
Si le modèle AR fournit une représentation concise et efficace du signal dans le domaine spectral, le modèle sinusoïdal est plus adapté à l'analyse/synthèse de trames voisées et il a d'ores et déjà permis d'obtenir des avancées significatives notamment en codage et en synthèse. Cependant, pour des applications de rehaussement son utilisation n'a pas permis d'améliorer les performances des méthodes non-paramétriques d'atténuation spectrale court-terme [3, 7, 4]. Les méthodes de débruitage fondées sur un modèle AR gaussien ont été aussi étudiées mais l'estimation non biaisée des paramètres AR à partir d'observations bruitées est une première difficulté ; par ailleurs, un tel modèle n'est pas bien adapté à des trames quasi périodiques telles que les voyelles, les bruits voisés etc. [8]. Pour pallier ce problème, Goh et al. [10] proposent de considérer un modèle tout pôle dont l'excitation est une séquence blanche pour des trames non voisées ou un train d'impulsions périodiques pour des trames voisées. Une approche alternative [11, 1] repose sur l'utilisation d'un modèle sinusoïdal en cas de voisement. Elle est plus facile à implanter car en présence de bruit ce modèle est un modèle de régression que l'on peut traiter à partir de techniques standards telles que les moindres carrés. Les résultats sont prometteurs mais cette approche n'est pas robuste au bruit additif. Notons enfin que la modélisation harmonique peut engendrer des artefacts, tout particulièrement dans les régions du spectre où prédomine le bruit de fond stationnaire.
Notre objectif est de trouver une représentation des trames voisées du signal de parole moins contraignante que le modèle sinusoïdal déterministe et qui fournit une modélisation du comportement du signal plus précise que celle obtenue par un modèle AR.

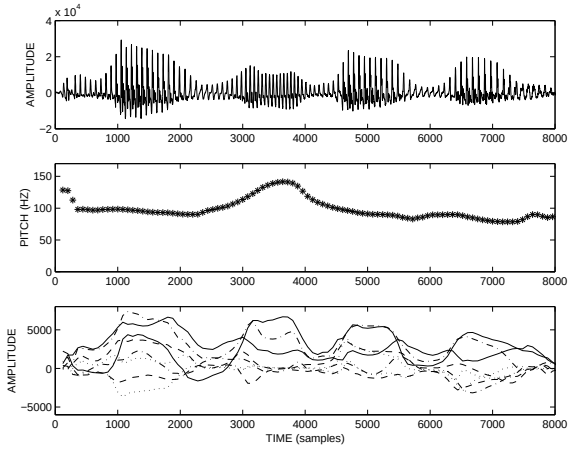

FIG. 1 – Signal de parole échantillonné à 8 kHz : évolution du pitch et des amplitudes des composantes en phase et quadrature des 4 premiers harmoniques

2 Le modèle

La modélisation harmonique stochastique a été retenue pour représenter les signaux quasi périodiques tels que les trames voisées. Les observations du signal suivent la relation suivante :

$$X_t = \underbrace{\sum_{k=1}^{K} A_t^k \cos(k\Phi_t) + B_t^k \sin(k\Phi_t)}_{S_t} + N_t. \quad (1)$$

où $(A_t^k)_{t\geq 1}$, $(B_t^k)_{t\geq 1}$ et $(\Phi_t)_{t\geq 1}$ sont des processus stochastiques et $(N_t)_{t\geq 1}$ est un bruit d'observation additif stationnaire.
Bien que la modélisation stochastique de la phase soit satisfaisante sur le plan conceptuel, elle pose des problèmes délicats d'estimation puisqu'il n'existe pas de procédure exacte de filtrage voire de lissage (i.e. calcul de l'espérance de S_t étant données les observations $X_1, \dots X_T$). Ce problème peut être contourné à partir de techniques numériques comme les simulations Monte Carlo par Chaîne de Markov [2], mais leur complexité calculatoire élevée ne semble pas adaptée à des applications sur la parole.
D'après la figure 1, on observe que la vitesse et l'échelle des variations de la fréquence du pitch sont bien moins importantes que celles des amplitudes des composantes harmoniques. La modélisation stochastique de la phase n'est pas aussi nécessaire que celle des amplitudes des composantes harmoniques. C'est pourquoi nous envisageons un traitement trame par trame et supposons la pulsation ω constante sur la trame d'analyse :

$$S_t = \sum_{k=1}^{K} A_t^k \cos(k\omega t) + B_t^k \sin(k\omega t), \quad (2)$$

Le modèle ainsi proposé est proche de celui traité dans [9] à la différence que chaque harmonique est ici défini à partir des deux composantes en phase et en quadrature. Mais, le modèle proposé dans [9] est cyclostationnaire.
Ici, si l'on suppose que $(A_t^k)_{t\geq 1}$ et $(B_t^k)_{t\geq 1}$ sont des processus stationnaires d'ordre 2 non corrélés et de même fonction d'autocovariance $r_k(d)$, la séquence $(S_t)_{t\geq 1}$ est un processus stationnaire d'ordre 2 de fonction d'autocovariance :

$$r_S(d) = \sum_{k=1}^{K} r_k(d) \cos(k\omega d). \quad (3)$$

De plus, si l'on suppose que les amplitudes $(A_t^k)_{t\geq 1}$ et $(B_t^k)_{t\geq 1}$ sont modélisables par des processus autorégressifs d'ordre 1 [2, 9], $r_k(d)$ vérifie :

$$r_k(d) = \sigma_k^2 \gamma_k^d / (1-\gamma_k^2), \quad (4)$$

où σ_k^2 est la variance de l'innovation pour le $k^{\text{ième}}$ harmonique et γ_k le paramètre AR associé.
Dans la suite de cette communication, le bruit additif $(N_t)_{t\geq 1}$ est supposé blanc et de variance η^2. Pour effectuer un rehaussement de signal de parole par filtrage de Kalman, le système (1) est représenté dans l'espace d'état :

$$\mathbf{S}_{t+1} = \mathbf{\Gamma} \mathbf{S}_t + \mathbf{\Sigma} \mathbf{d}_{t+1} \quad (5)$$

$$X_t = \mathbf{h}_t' \mathbf{S}_t + \eta N_t \quad (6)$$

Le vecteur d'état colonne $\mathbf{S}_t$ est la concaténation des amplitudes des composantes harmoniques :

$$\mathbf{S}_t = (A_t^1, B_t^1, \dots A_t^K, B_t^K)'. \quad (7)$$

Cette représentation permet d'obtenir une estimation de S_t au sens des moindres carrés, étant donné les observations $X_1, \dots, X_T$, grâce à un lissage de Kalman. Si l'on suppose $(A_t^k)_{t\geq 1}$, $(B_t^k)_{t\geq 1}$ et $(N_t)_{t\geq 1}$ gaussiens, cette procédure est équivalente à l'espérance de S_t conditionnée par les observations. Pour obtenir une estimation du signal S_t par filtrage/lissage de Kalman, les paramètres du modèle, $\{\sigma_k^2\}_{k=1,\dots,K}$, $\{\gamma_k^2\}_{k=1,\dots,K}$ et η doivent être estimés. Dans la section suivante nous proposons trois approches différentes d'estimation de ces paramètres.

3 Estimation des paramètres

3.1 Approximation de Whittle

D'après les hypothèses précédentes, le système (2,3,4) possède une représentation ARMA$(2K-1, 2K)$ dont les $2K$ pôles sont $\gamma_k \mathrm{e}^{\pm jk\omega}$. Plus précisément, la densité spectrale de puissance $f_S(\lambda)$ de S_t s'écrit :

$$f_S(\lambda) = \sum_{k=1}^{K} \left| \left(\frac{\mu_k (1 - \delta_k \mathrm{e}^{-j\lambda})}{1 - 2\gamma_k \cos(k\omega) \mathrm{e}^{-j\lambda} + \gamma_k^2 \mathrm{e}^{-j2\lambda}} \right|^2 \right. \quad (8)$$

avec

$$\delta_k \triangleq \frac{1}{2\gamma_k \cos(k\omega)} \left[(1+\gamma_k^2) - \sqrt{\gamma_k^4 - 2\gamma_k^2 \cos(2k\omega) + 1} \right],$$

et

$$\mu_k \triangleq \sigma_k \gamma_k \sqrt{\frac{1 + \cos(2k\omega)}{(1+\gamma_k^2) - \sqrt{\gamma_k^4 - 2\gamma_k^2 \cos(2k\omega) + 1}}}.$$

La relation (8) peut être exploitée pour estimer les paramètres en utilisant l'approximation de Whittle de la log-vraisemblance qui fait intervenir le périodogramme des données [6] :

$$L = \sum_{k=1}^{T} \left[\log f_S(\lambda_k) + \frac{1/T \left| \sum_{t-1}^{T} X_t \mathrm{e}^{jt\lambda_l} \right|^2}{f_S(\lambda_k)} \right]. \quad (11)$$

3.2 Algorithme EM

La deuxième approche envisagée consiste à estimer des paramètres $\{\gamma_k^2\}_{k=1,\dots,K}$ et $\{\sigma_k^2\}_{k=1,\dots,K}$ en maximisant la vraisemblance des observations. Pour cela, nous proposons d'utiliser l'algorithme EM (Expectation-Maximization)[5]. A chaque itération, l'étape de maximisation (M) effectue une remise à jour de l'estimation des paramètres.

Cela nécessite la connaissance de quantités que l'on obtient par lissage de Kalman, dans l'étape d'Espérance (E). Cette approche fournit donc conjointement une estimation itérative de $\{\gamma_k^2\}_{k=1,\ldots,K}$, $\{\sigma_k^2\}_{k=1,\ldots,K}$, η et du signal rehaussé.

3.3 Méthode des moments

L'équation (3) peut être écrite sous forme matricielle :

$$\underbrace{\begin{pmatrix} r_S(0)-\eta^2 \\ \vdots \\ r_S(L) \end{pmatrix}}_{\mathbf{r}} = \mathbf{M} \underbrace{\begin{pmatrix} \sigma_1^2 \\ \vdots \\ \sigma_K^2 \end{pmatrix}}_{\boldsymbol{\theta}} \tag{12}$$

$\mathbf{M}$ est une matrice de taille $(L+1)\times K$, dont l'élément (l,k) vaut $\gamma_k^{l-1}/(1-\gamma_k^2)\cos[(l-1)k\omega]$.
θ, le vecteur concaténant les variances des innovations des harmoniques, peut donc être estimé, étant donnés les autres paramètres, par la méthode des moments généralisés :
$\mathbf{r}$ est remplacé par une estimation empirique $\hat{\mathbf{r}}$ et le critère $\|\hat{\mathbf{r}} - \mathbf{M}\boldsymbol{\theta}\|_2^2$ est minimisée par rapport à $\boldsymbol{\theta}$ sous la contrainte que chaque élément de θ, doit être positif. Ce dernier problème, classique en traitement d'image, peut être résolu de manière itérative à partir d'un algorithme de gradient modifié simple à mettre en œuvre où chaque itération est définie comme suit :

$$\theta_k := \theta_k \exp\left(\hat{\theta}_k/\theta_k - 1\right). \tag{13}$$

θ_k est la $k^{\text{ième}}$ coordonnée de $\boldsymbol{\theta}$ et $\hat{\theta}_k$ est la $k^{\text{ième}}$ coordonnée de la solution des moindres carrés sans contrainte $(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\hat{\mathbf{r}}$.

4 Résultats

Les méthodes que nous avons présentées ont été testées sur différents segments du signal de parole contaminés par un bruit additif. Les estimations des $\{\sigma_k^2\}_{k=1,\ldots K}$ et des $\{\gamma_k\}_{k=1,\ldots K}$ au moyen de la vraisemblance de Whittle ou de l'algorithme EM se sont avérées lourdes à mettre en oeuvre. De plus des simulations menées sur des données synthétisées montrent que l'on peut difficilement se fonder sur les résultats d'estimation obtenus, à cause de leur forte variance. Pour cette raison nous avons décidé de réduire la paramétrisation du signal à $K+1$ paramètres, en prenant une valeur identique γ pour tous les γ_k, ce qui revient à supposer que les K modes ont la même largeur de bande. Dans ce cas, on constate que γ varie entre $0,97$ et $0,99$, ces variations ayant peu d'impact sur le signal rehaussé.
Sur l'exemple donné, nous avons pris $\gamma = 0,98$. La méthode proposée au paragraphe 3.3 donne un rehaussement de bonne qualité. Les spectrogrammes des signaux bruités et rehaussés sont donnés respectivement aux figures 2 et 3.
Il est à noter que l'utilisation de l'algorithme EM conduit à des résultats quasi identiques, mais avec une complexité calculatoire beaucoup plus grande.
Si l'on choisit $\gamma = 0,999$, la restauration du signal est entachée de quelques artefacts de fausse harmonicité dans les zones de bas rapport signal à bruit et aboutit à un résultat proche de celui obtenu avec un modèle sinusoïdal classique.Le résultat est donné à la figure 4.

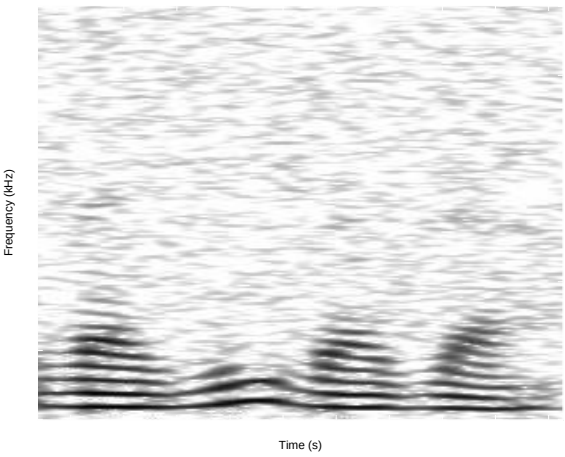

FIG. 2 – Spectrogramme du signal bruité

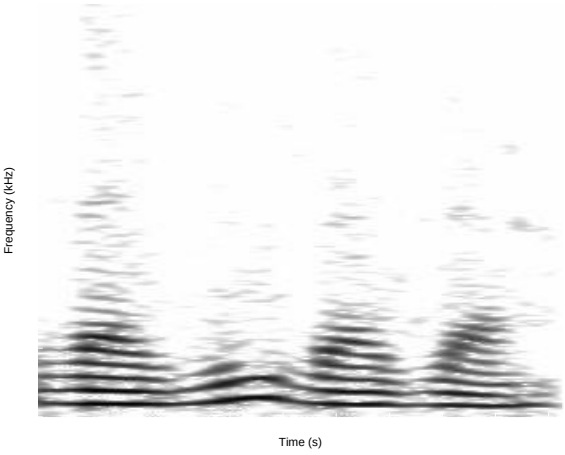

FIG. 3 – Spectrogramme du signal rehaussé avec $\gamma = 0,98$

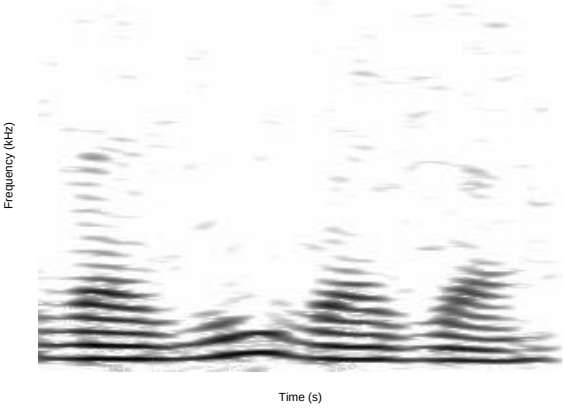

FIG. 4 – Spectrogramme du signal rehaussé avec $\gamma = 0,999$

Remerciements

Ce travail est soutenu par le GdR-PRC ISIS dans le cadre de l'action incitative *jeunes chercheurs*.

Références

[1] D. V. Anderson and M. A. Clements, "Audio signal noise reduction using multi-resolution sinusoidal modeling," IEEE ICASSP, 1999.

[2] C. Andrieu and A. Doucet, "Optimal Estimation of Amplitude and Phase Modulated Signals", technical report, Cambridge University, CUED/F-INFENG/TR395, 2000.

[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral substraction," vol. 27, no. 2, pp. 113–120, 1979.

[4] O. Cappé and J. Laroche, "Evaluation of short-time spectral attenuation techniques for the restoration of musical recordings," vol. 3, no. 1, pp. 84–93, Jan. 1995.

[5] L. Deng and X. Shen, "Maximum likelihood in statistical estimation of dynamic systems : Decomposition algorithm and simulation results," Signal processing, vol. 57, pp. 65–79, 1997.

[6] K. Dzhaparidze K., "Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series," 1986, Springer Verlag, New York.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," vol. 32, no. 6, pp. 1109–1121, 1984.

[8] M. Gabrea, E. Grivel and M. Najim, "A Single Microphone Kalman Filter-Based Noise Canceler," IEEE Signal Processing Letters, vol. 6, no. 3, pp. 55–57, 1999.

[9] M. Ghogho, A. Swami, and B. Garel, "Performance analysis of cyclic statistics for the estimation of harmonics in multiplicative and additive noise," vol. 47, no. 12, pp. 3235–3249, 1999.

[10] Z. Goh, K.-C. Tan and B. T. G. Tan, "Kalman-Filtering Speech Enhancement Method Based on a Voiced-Unvoiced Speech Model, " IEEE Trans. on Speech and Audio Processing, vol. 7, no. 5, pp. 510–524, 1999.

[11] T.F. Quatieri and R.J McAulay, "Noise reduction using a soft-decision sine-wave vector quantizer," IEEE ICASSP, Albuquerque, New Mexico, 1990.