Une nouvelle méthode d'estimation des spectres de puissance par un modèle source-filtre Application à l'analyse-synthèse de la parole

Thierry GALAS - Xavier RODET

LAFORIA - UA CNRS N°1095 Laboratoire de Traitement de la Parole
Université Pierre & Marie Curie - 4, place Jussieu - 75252 PARIS CEDEX 05

RESUME

Le problème étudié est l'approximation d'un spectre de puissance de raies par un modèle source-filtre et son application à l'analyse-synthèse du signal vocal. Nous considérons d'abord la prédiction linéaire discrète [El-Jaroudi 86] qui permet une telle modélisation par un filtre tout pôles. Nous présentons ensuite une nouvelle méthode où l'enveloppe spectrale appartient à une extension du domaine des enveloppes des filtres tout pôles. En appliquant cette méthode à des signaux de paroles réels nous mettons en évidence la necessité de résoudre le probléme en terme d'estimation ; ceci permet également d'introduire de nouvelles contraintes sur les enveloppes considérées. L'application de notre nouvelle méthode conduit à de bien meilleures estimations. Nous présentons finalement l'exploitation en synthèse des enveloppes spectrales générées.

SUMMARY

The problem we consider is the approximation of a discrete power spectrum in terms of a source-filter model, and its application to analysis and synthesis of speech. We first consider the Discrete Linear Prediction [El Jaroudi 86] which does an all-pole modeling. We then present a new method where spectral envelopes are taken from an extensions of the class of all-poles filter envelopes. When applying this method to real speech signals it appears that the problem has to be solved in terms of estimation : this also allow constraints on the envelopes so obtained to be taken into account. Whe show that our method leads to better estimations. We finally present the use of the obtained spectral envelopes for speech synthesis.

1. Introduction.

La modélisation d'un signal par un système source-filtre est classiquement utilisée dans le domaine de l'analyse-synthèse de la parole. On considère généralement dans le cas d'un son voisé, que le filtre et la source modélisent respectivement la fonction de transfert du conduit vocal et l'effet combiné de l'onde glottique et de la radiation aux lèvres.

L'identification des paramètres d'un tel modèle peut se faire directement dans le domaine temporel. En considérant un filtre tout pôles et en choisissant pour critère d'optimisation la minimisation de la norme quadratique du signal d'erreur Hedelin est conduit à une modélisation AR-X qui nécessite la connaissance de la phase de la source ainsi que celle du signal généré par le système à modéliser [Hedelin 84].

Le système est observé au travers d'une chaîne de transmission engendrant souvent des rotations de phase parfois inconnues et variables dans le temps. C'est le cas en traitement de la parole sous l'influence du milieu acoustique, de la position du locuteur, etc ... La compensation de ces rotations de phase et la recherche de la phase de la source constituent des problèmes difficiles à résoudre .

Par contre on peut facilement obtenir une bonne estimation des spectres de puissance du signal et de la source, d'où l'intérêt d'identifier les paramètres d'un modèle source-filtre pour approximer un spectre de puissance. On recherche alors les paramètres du spectre de puissance P (choisi dans une classe prédéfinie) qui, par produit avec le spectre de puissance S de la source supposée, fournit la meilleure approximation (au sens d'une distance spectrale) du spectre de puissance X du signal analysé ($X \approx S.P$).

Soit pour distance spectrale la distance d'Itakura et Saito, pour ensemble de filtres celui des filtres tout pôles et supposons le spectre de la source plat ($S(\omega)=1$ pour tout ω), alors on est conduit à une modélisation AR classique [Itakura 68], où l'on identifie les paramètres en résolvant le sytème linéaire d'équations de Yule-Walker.L'application de ce modèle se justifie pleinement dans le cas des sons non-voisés. Mais, si le signal est quasi périodique et si l'ordre du filtre n'est pas négligeable devant le nombre d'échantillons de la pseudo-période, alors les enveloppes spectrales estimées sont erronées [Atal 74] ; [El-Jaroudi 86] parce que les hypothèses sur la source ne sont plus vérifiées (par exemple pour les voix de pitch élevé).

Ce sont ces raisons qui mènent El-Jaroudi et Makhoul à modifier les termes du problème en ne prenant en compte qu'un nombre discret de points des spectres considérés [El-Jaroudi 86]. La prédiction linéaire discrète (Dicrete Linear Prediction Coding, DLPC) consiste en fait à considérer que la source et le signal modélisé présentent des spectres ayant même support discret $\Omega=\{\omega_i, i=1 \text{ à } n\}$ avec :

$$S=\sum_{i=1}^{n} \delta_{\omega_i} \qquad (1) \qquad\qquad X=\sum_{i=1}^{n} x_i \delta_{\omega_i} \qquad (2)$$

L'estimation des paramètres du filtre nécessite alors la résolution d'un système d'équations non linéaires présenté au paragraphe 2.

Au paragraphe 3 nous proposons une autre formulation qui conduit à un système d'équations linéaires.

2. Prédiction Linéaire Discrète.

2.1 Espace des filtres considérés.

L'espace des filtres considérés est celui des filtres tout pôles d'ordre p dont la réponse fréquentielle est de la forme $1/A(\omega)$ avec :

$$A(\omega) = \sum_{k=0}^{p} a_k e^{-j\omega k} \qquad (3)$$

Le spectre de puissance correspondant à un tel filtre s'écrit :

$$P(\omega) = \frac{1}{d_0+\sum_{i=1}^{p} 2d_i \cos(\omega i)} \qquad (4)$$

où les d_i sont les coefficients d'autocorrélation de la réponse impulsionnelle du filtre inverse c'est-à-dire :

$$d_i=\sum_{k=0}^{p-i} a_i\, a_{k+i} \qquad (5)$$

2.2 Critère d'erreur.

Le critère d'erreur sélectionné El-Jaroudi et Makhoul est l'expression pour un spectre discret de la distance d'Itakura et Saito. Nous ajoutons à cette expression une pondération en fréquence h_i, d'où l'expression de l'erreur :

$$E = \frac{1}{n}\sum_{i=1}^{n} h_i \left(\frac{x_i}{P(\omega_i)} - \log\left(\frac{x_i}{P(\omega_i)}\right)\right) \qquad (6)$$

2.3 Condition de minimisation de l'erreur.

Pour minimiser l'erreur on annule les dérivées partielles par rapport aux coefficients du filtre, le résultat est le système d'équations suivant :

$$\sum_{k=0}^{p} R'_{i-k}\, a_k = \sum_{k=0}^{p} R_{i-k}\, a_k \quad \text{pour } 0\le i\le p \qquad (7)$$

avec : $R'_i = \sum_{m=1}^{n} h_m P(\omega_m) e^{j\omega_m i}$ (8) et $R_i = \sum_{m=1}^{n} h_m x_m e^{j\omega_m i}$ (9)

Le système d'équations (7) peut se réécrire pour $0\le i \le p$:

$$\sum_{k=0}^{p} R'_{i-k} a_k = g(-i) \quad \text{avec} \quad g(-i) = \frac{1}{n}\sum_{m=1}^{n} \frac{h_m e^{-j\omega_m i}}{A(\omega_m)} \qquad (10)\ (11)$$

Le calcul des a_k nécessite la résolution de ce système d'équations qui est non linéaire et pour lequel un schéma de résolution itératif est décrit par El-Jaroudi et Makhoul [El-Jaroudi 86].

3. Nouvelle formulation.

3.1 Domaine des enveloppes spectrales considérées.

On a vu précédemment l'expression (4) du spectre de puissance de la réponse impulsionnelle d'un filtre tout pôles d'ordre p. Par extension, on choisit pour domaine des spectres de puissance, l'ensemble des fonctions de la forme (4) (les d_i étant alors des réels quelconques). Ces fonctions ont la particularité de ne pas présenter obligatoirement une énergie physiquement significative pour toute fréquence (énergie négative ou infinie). Cependant on peut remarquer que dans R^p le domaine des d_i, correspondant à des enveloppes physiquement significatives, est convexe.

3.2 Critère d'erreur.

On sélectionne le critère d'erreur suivant :

$$E=\sum_{i=1}^{n} h_i \left(\frac{x_i}{P(\omega_i)\, S_i} - 1\right)^2 \qquad (12)$$

où les h_i sont des coefficients positifs permettant de pondérer l'importance relative des différentes raies spectrales.

Si l'erreur sur une raie est suffisamment petite ($x_i / P(\omega_i) = 1+\varepsilon$ avec ε petit) on a :

$$(\log(x_i)-\log(p(\omega_i)S_i))^2 = \log\left(\frac{x_i}{P(\omega_i)\, S_i}\right)^2 = \log(1+\varepsilon)^2 \approx \varepsilon^2 = \left(\frac{x_i}{P(\omega_i)S_i} - 1\right)^2 \qquad (13)$$

Ainsi E apparait comme une somme pondérée des différences de deux spectres sur une échelle logarithmique, ce qui est une distance relativement pertinente sur le plan perceptif, et qui justifie le choix de notre critère d'erreur.

3.3 Identification des paramètres.

On considère l'espace de Hilbert des spectres de puissance de support inclus dans Ω muni du produit scalaire défini par :

$$\langle F,G\rangle = \sum_{i=1}^{n} h_i F_i G_i \qquad (14)$$

On note :

$$D_i = \delta_{\omega_i} \quad \text{et} \quad D = \sum_{i=1}^{n} D_i \qquad (15)$$

Les D_i forment une base de cet espace. On considère le sous-espace engendré par les vecteurs Y_j (j=0 à p) définis par $Y_0=X/S$ et $Y_j=(X/S).2.\cos(j\omega)$ (j=1 à p). On projette alors D dans ce sous-espace. En effet les coefficients d_i qui minimisent l'erreur E sont ceux qui minimisent :

$$\left\| \sum_{i=0}^{p} d_i Y_i - D \right\|^2 \qquad (16)$$

Nous calculons ces coefficients à l'aide de l'algorithme de Graam-Schmidt stabilisé.

3.4 Résultats.

Si l'on génère un signal par filtrage d'un peigne de Dirac de période M échantillons dans un filtre tout pôles d'ordre P, alors si 2P≤M la méthode permet de retrouver (aux erreurs numériques près) les paramètres du spectre de puissance du filtre utilisé. Par contre si 2P>M les Y_j forment un système lié, il existe alors une infinité de filtres d'ordre P dont les spectres de puissance correspondent à une erreur nulle. Les raies spectrales ne fournissent plus assez d'information sur l'enveloppe spectrale du filtre utilisé pour identifier celui-ci de façon univoque.

Si l'on perturbe les valeurs des différentes raies spectrales par un bruit de niveau fixe afin d'observer la sensibilité de la méthode aux erreurs de données [Vignes 86] on constate que cette sensibilité augmente quand le nombre de raies spectrales diminue. On représente figure 1 des familles d'enveloppes identifiées pour un niveau de bruit constant et différentes valeurs de la période, à comparer à la réponse en puissance du filtre utilisé.

4. Application à l'analyse de la parole.

Les valeurs (fréquence, amplitude) des différentes raies spectrales sont détectées sur une estimation spectrale calculée par FFT. Appliquée directement à ces valeurs la méthode fournit parfois des résultats non satisfaisants (enveloppes non physiquement significatives dites aberrantes). Cela tient à un ordre trop élevé du modèle et au fait que l'espace des enveloppes des spectres de puissance utilisé est bien plus étendu que celui des filtres tout pôles: des enveloppes aberrantes peuvent en effet passer par les valeurs spectrales X_i (ce problème affecte aussi la DLPC bien que de façon moins aiguë voir Fig 3). Pour éviter ces aberrations il est nécessaire d'introduire plus d'informations ou de contraintes sur le système.

Pour ce faire nous prenons en compte le fait que le modèle est simplifié par rapport au système réel, et le fait que les spectres de la source et du signal ne sont connus qu'approximativement. Il convient donc de remplacer la valeur déterministe nécessairement incertaine, de position ω_i et d'amplitude x_i de chaque raie par une distribution de probabilité $Pb_i(\omega, x)$. Le critère d'optimisation devient alors la minimisation de l'espérance mathématique E du critère d'erreur précédemment sélectionné. Si la source est supposée constituée de raies unitaires, cette espérance se met sous la forme :

$$E = \sum_{i=1}^{n} \iint Pb_i(\omega, x)\, h_i \left(\frac{x}{P(\omega)} - 1\right)^2 d\omega\, dx \qquad (17)$$

Cette **nouvelle méthode** permet d'**introduire des informations supplémentaires, ou des contraintes, sur la famille des enveloppes recherchées**. Ainsi, dans le cas de la parole, nous utiliserons les connaissances spécifiques dont nous disposons sur les enveloppes spectrales d'un signal vocal. On peut par exemple tenir compte du fait qu'une limite inférieure de largeur de bande des formants entraîne une limite supérieure de la pente de son enveloppe spectrale, donc de la pente de l'enveloppe globale recherchée, au voisinage des points (ω_i, x_i). Ceci est fait en modifiant la distribution de probabilité de façon à rendre beaucoup plus probable une enveloppe spectrale de pente localement inférieure au seuil.

Afin d'approcher la minimisation de l'expression (17) on procède de la façon qui suit: chaque raie (ω_i, x_i) est remplacée par plusieurs raies de valeurs proches (ω_k, x_k), chacune associée à un poid h_k correspondant à sa probabilité de représenter la raie (ω_i, x_i). L'amélioration des résultats ainsi obtenue peut être constatée en comparant les figures 4 et 5.

5. Exploitation en synthèse.

Une analyse centiseconde ou pitch-synchrone produit un ensemble discret de vecteurs acoustiques représentant :

- $A(t_n, \omega)$: l'enveloppe spectrale d'amplitude du signal pour la trame ou la période centrée sur l'instant t_n .
- $f_0(t_n)$: la fréquence fondamentale estimée du signal pour la trame ou la période centrée sur l'instant t_n .

Dans notre cas on a:

$$A(t_n, \omega) = \frac{1}{\sqrt{d_0 + \sum_{i=1}^{p} d_i \cos(\omega i)}} \qquad (18)$$

Effectuer une synthèse de bonne qualité nécessite la construction d'un signal dont le spectre à court-terme présente à l'instant t $(t_n \le t < t_{n+1})$ une enveloppe qui soit proche (au sens d'une distance spectrale) de $A(t_n)$ relativement à $(t - t_n)$. Cet impératif conduit à l'utilisation d'interpolations évitant des discontinuités de méthode lors de la génération du signal. Afin de parvenir à ce but deux solutions sont envisagées :

- la synthèse additive directe.
- la synthèse par filtre récursif.

5.1 Synthèse additive.

A l'aide de l'ensemble discret de valeurs dont on dispose et par interpolation on construit les fonctions $f_0(t)$ (fréquence instantanée) et $A(t, \omega)$ (enveloppe spectrale instantanée). On génère alors le signal x(t) à l'aide de la version discrétisée dans le temps de la formule :

$$x(T) = \sum_{n=0}^{\mathrm{int}(fe/(2f_0(t)))} A(T, n f_0(T)) \cos\left(\int_0^T 2\pi n f_0(t)\, dt\right) \qquad (19)$$

où fe est la fréquence d'échantillonnage.

Cette assimilation de l'organe vocal à un générateur de sinusoïdes se justifie dans le cas d'une voix de pitch élevé mais elle est inadéquate aux pitch bas, par exemple à la simulation d'un mode de fonctionnement tel que le mode "vocal fry".

5.2 Synthèse par filtre récursif.

Il s'agit en fait d'utiliser une synthèse LPC classique. Pour ce faire il est nécessaire de transformer $A(t_n)$ en paramètres d'un filtre récursif avec un paramètre de gain. Cette transformation s'effectue

simplement en appliquant la méthode d'autocorrélation au spectre de puissance $A(t_n)$.

Des comparaisons effectuées précedemment [Rodet 87a] ; [Rodet 87b] montre que cette méthode donne des résultats supérieurs à la synthèse additive. Il semble que la meilleure adéquation à la réalité physique du modèle de production utilisé en soit la cause.

6. Conclusion

Nous avons proposé une méthode d'estimation des paramètres d'un modèle source-filtre capable de traiter le cas des spectres de raies, particulièrement intéressante dans le cas du traitement des voix de pitch élevé, permettant d'intégrer des contraintes et des informations supplémentaires. Un développement possible serait d'intégrer les informations fournies sur la réponse impulsionnelle du conduit vocal par les évolutions du fondamental sur une courte durée.

Références

[Hedelin 84], P. Hedelin, "A glottal LPC-vocoder", ICASSP-84, San-Diego, Mars 1984.

[Itakura 68], F. Itakura, S. Saito, "Analysis Synthesis Telephony based on the Maximum Likelihood Method" Proc. 6eme Intern. Congr. Acoust., Tokyo, C17-20, 1968.

[Atal 74], B.S. Atal, "Recent Advances in Predictive Coding Applications to Speech Synthesis", Proc.Stockholm Speech Communications Seminar, 1974.

[El-Jaroudi 86], A. El-Jaroudi, J. Makhoul, "All-pole Modeling for discrete Spectra", IEEE ASSP Workshop on Spectrum Estimation and Modeling, Boston, MA, Nov. 1986, pp 29-32.

[Vignes 86], J. Vignes, "Approche stochastique de l'analyse de la propagation des erreurs de données dans les algorithmes numériques", Annales des Télécommunications, n°49 Juin 1986.

[Rodet 87a] X. Rodet, P. Depalle, G. Poirot, "Diphone Sound Synthesis", European Conf. on Speech Technology, Edinburgh, U.K., Sept 87.

[Rodet 87b] X. Rodet, P. Depalle, G. Poirot, "Analyse et synthèse de voix parlée et chantée par modélisation de l'enveloppe spectrale et de l'excitation", 16èmes JEP, Hammamet, 1987.

Notes :

- Figures 1x graduations verticales en dBx10
- Figures 2 à 5 axe vertical en dB, axe horizontal en Khz, modèle d'ordre 24.

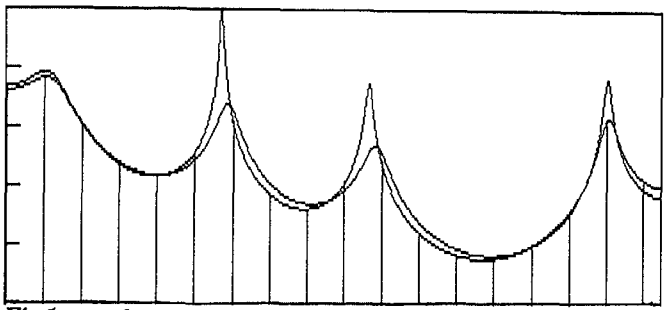

Fig 1a enveloppe du filtre d'ordre 10 utilisé et enveloppe identifée par autocorrélation pour un période de 35 échantillons.

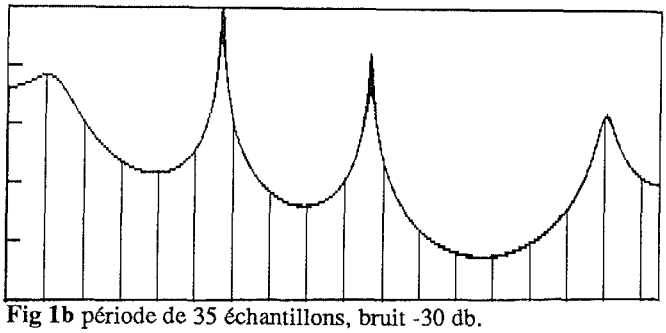

Fig 1b période de 35 échantillons, bruit -30 db.

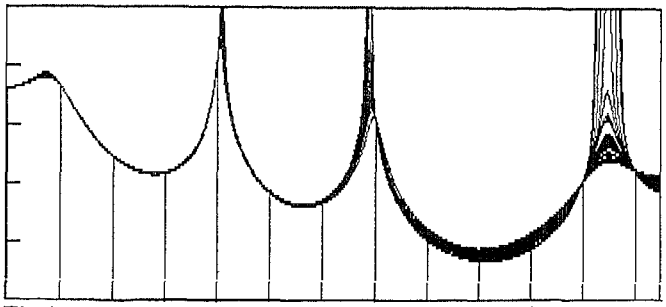

Fig 1c période de 25 échantillons, bruit -30db.

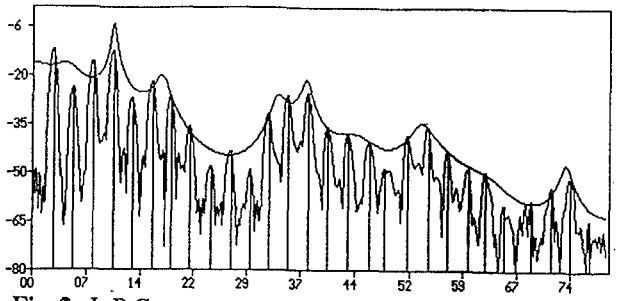

Fig 2 L.P.C

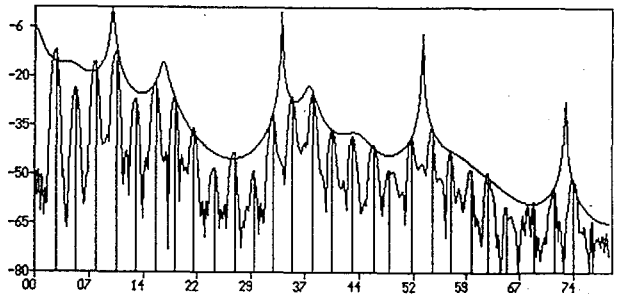

Fig 3 D.L.P.C

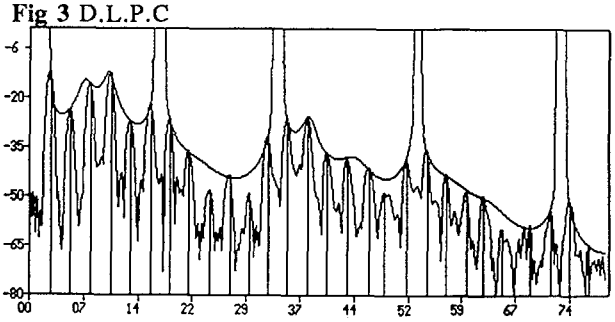

Fig 4 Méthode proposée appliquée directement.

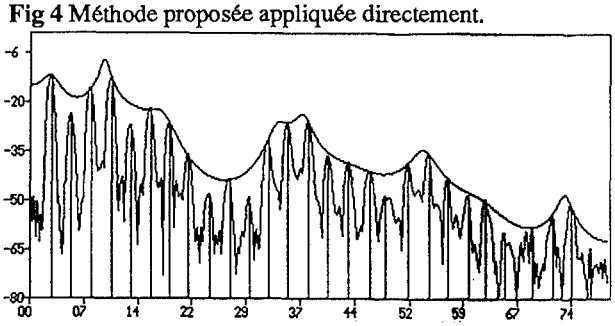

Fig 5 Méthode proposée appliquée en termes d'estimation.