# Estimation du signal glottique basée sur un modèle ARX

Damien VINCENT[1], Olivier ROSEC[1], Thierry CHONAVEL[2]

[1]France Telecom Division R&D TECH/SSTP
2, Avenue Pierre Marzin, 22307 Lannion Cedex, France

[2]ENST Bretagne, département Signal et Communications
B.P. 832, 29285 Brest Cedex, France
{damien.vincent,olivier.rosec}@francetelecom.com, thierry.chonavel@enst-bretagne.fr

**Résumé –** Le but de cet article est d'estimer à partir du seul signal de parole le signal de source glottique. L'utilisation du modèle ARX de production de la parole ainsi que d'un modèle de source glottique transforme ce problème de déconvolution en un problème d'optimisation non linéaire. Nous présentons une méthode efficace pour résoudre ce problème ainsi que des résultats sur signaux synthétiques et réels.

**Abstract –** The goal of this paper is to estimate the glottal source signal from the sole speech signal. By using the ARX model of speech production and a glottal source model, this deconvolution problem is turned into a complex nonlinear optimization problem. We present an efficient method to solve this problem, and some experiments on synthetic speech as well as on natural speech signals.

## 1 Introduction

La caractérisation de la qualité vocale à l'aide du signal de source glottique a fait l'objet de nombreuses études [1] et permet d'envisager des applications intéressantes visant à caractériser des voix ou à réaliser des transformations de la voix. L'information issue de la source glottique peut par exemple se révéler utile dans des domaines tels que la conversion de voix ou la synthèse de parole expressive.

La mesure directe de l'onde glottique ne peut être réalisée que par des méthodes intrusives et donc incompatibles avec la plupart des applications visées. Il convient par conséquent d'estimer le signal de source glottique à partir du seul signal de parole. Pour traiter ce problème de déconvolution, un modèle ARX (Auto-Regressive eXogenous) est utilisé pour représenter les mécanismes de production de la parole. De plus, une information a priori sur la source glottique est introduite par le biais du modèle LF (Liljencrant-Fant, [2]). Dans ce cadre, l'estimation conjointe de la source glottique et du conduit vocal s'apparente à un problème complexe d'optimisation non-linéaire. Pour limiter la complexité de l'estimation, nous proposons une méthode efficace visant notamment à réduire drastiquement le domaine d'exploration des paramètres. L'algorithme proposé est appliqué à l'analyse de signaux de parole synthétique et naturelle.

## 2 Modèle de production de la parole

Le modèle ARX est couramment utilisé pour représenter de manière linéaire le processus de production de la parole [3] :

$$s(n) = -\sum_{k=1}^{p} a_k(n)s(n-k) + b_0(n)u(n) + \epsilon(n) \quad , \quad (1)$$

où $s(n)$ correspond au signal de parole et $u(n)$ à la dérivée du signal de débit glottique, la dérivation correspondant à une première approximation de l'effet de radiation des lèvres. Le filtre tout pôle de coefficients $a_k(n)$ et d'ordre $p$ supposé fixé permet de représenter la réponse du conduit vocal, $b_0(n)$ est le facteur d'amplitude du signal de source glottique et $\epsilon(n)$ correspond à l'erreur de prédiction du modèle ARX.

Dans le cas d'une modélisation AR du signal de parole, le résidu est issu de la minimisation de l'erreur de prédiction. Aucun a priori n'est ajouté au signal d'excitation et celui-ci ne correspond donc pas à une approximation satisfaisante du signal de source glottique. Le modèle ARX corrige ce défaut en permettant de contraindre la source glottique $u(n)$ à rester dans un espace qui reste proche de la réalité physique.

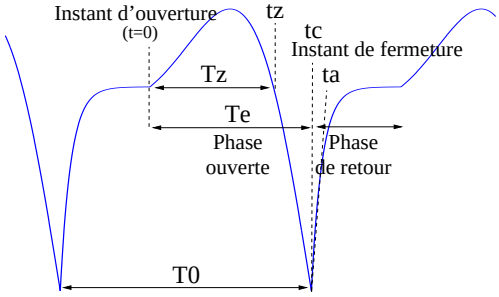


FIG. 1 – Le modèle LF

La contrainte sur la source glottique peut passer par l'utilisation d'un modèle paramétrique, celui-ci étant établi à partir de mesures directes du débit glottique ainsi qu'à partir de la modélisation de certains phénomènes physiques intervenant dans le processus de production de la parole. Parmi ces modèles paramétriques, le modèle LF, dont la dérivée du signal glottique et ses instants caractéristiques sont représentés sur la figure 1, est le plus répandu. Le modèle LF est un modèle à cinq degrés de liberté : un pour la position temporelle (la référence étant généralement l'instant de fermeture $t_c$), un pour l'amplitude (déjà intégré au modèle de production sous la forme du

coefficient $b_0$) et trois sur la forme du signal représentée par le vecteur $\theta = (O_q, \alpha_m, Q_a)$, où le quotient ouvert est défini par $O_q = \frac{T_e}{T_0}$, le coefficient d'asymétrie par $\alpha_m = \frac{T_z}{T_e}$ et le coefficient de phase de retour par $Q_a = \frac{T_a}{(1-O_q)T_0}$. Le domaine de variation admissible du vecteur de forme $\theta$ sera noté $\Theta$.
L'expression analytique du modèle LF fait intervenir un autre jeu de paramètres (non normalisés) qui est implicitement relié aux paramètres décrits précédemment :

$$\begin{array}{ll} u(t) = E_1 e^{at}\sin(wt) & 0 \le t \le T_e \\ u(t) = -E_2\left[e^{-b(t-T_e)} - e^{-b(T_0-T_e)}\right] & T_e \le t \le T_0 \end{array} \quad (2)$$

L'équation (2) permet de calculer analytiquement le spectre de l'onde glottique dont on peut déduire un comportement asymptotique représenté sur la figure 2 : le spectre est caractérisé par la présence d'une résonnance, aussi appelée formant glottique par analogie avec les résonnances du conduit vocal, et par une atténuation supplémentaire de 6dB/oct pour des fréquences situées au delà de la fréquence de coupure $F_a = \frac{1}{2\pi}\frac{1}{Q_a(1-O_q)T_0}$.

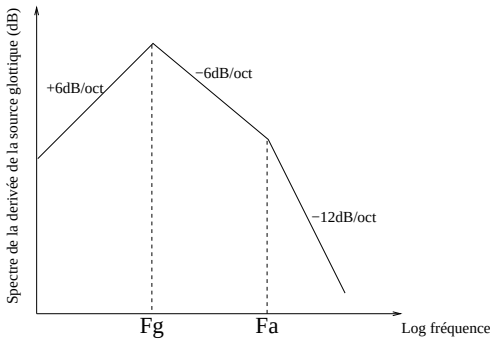


FIG. 2 – Comportement asymptotique de la dérivée de l'onde de débit glottique.

# 3 Estimation des paramètres

## 3.1 Principe général

En utilisant le modèle ARX et le modèle LF de source glottique, le problème de déconvolution est ramené à un problème d'optimisation non linéaire sur les paramètres de source et les coefficients $a_k(n)$, $b_0(n)$. La période fondamentale $T_0$ peut être estimée par des méthodes telles que l'algorithme YIN [4] et ceci de manière indépendante des autres paramètres de source. Pour pouvoir considérer le filtre AR comme stationnaire, l'analyse sera réalisée sur un intervalle $[t_c - T_0; t_c + T_0]$ (l'instant de fermeture détermine ainsi l'instant d'analyse) et en utilisant une fenêtre de Hanning. Le modèle ARX étant linéaire à $u$ fixé, l'estimation des $a_k$ et $b_0$ s'obtient simplement par minimisation du critère des moindres carrés $\sum \epsilon^2(n)$ qui s'écrit aussi sous la forme $\|M_u A - S\|^2$ où :
$S = (s(t_c - N), \cdots, s(t_c + N))^T$,
$A = (a_1, a_2, \cdots, a_p, b_0)^T$,
$U = (u(t_c - N), \cdots, u(t_c + N))^T$,
$M_u = [D|U]$,

avec $D = \begin{pmatrix} s(t_c - N - 1) & \cdots & s(t_c - N - p) \\ \cdots & \cdots & \cdots \\ s(t_c + N - 1) & \cdots & s(t_c + N - p) \end{pmatrix}$.

Ce minimum se calcule à l'aide d'algorithmes standards de résolution de systèmes linéaires et s'écrit analytiquement sous la forme suivante :
$E(\theta, t_c) = \min_{a_k, b_0} \sum \epsilon^2(n) = \|S - M_u(M_u^T M_u)^{-1} M_u^T S\|^2$.

## 3.2 Quantification et estimation des paramètres de forme

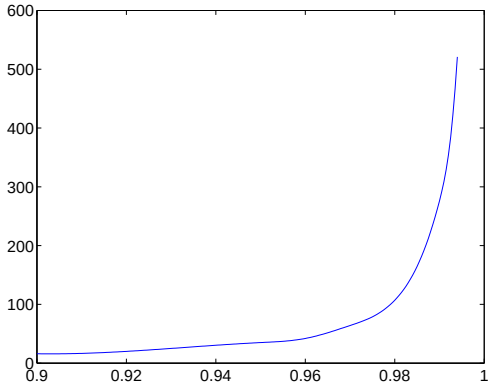


FIG. 3 – Nombre de sources de $\tilde{\Theta}_{\rho_m}$ en fonction de $\rho_m$

La minimisation de $E(\theta, t_c)$ vis à vis du vecteur $\theta$ des paramètres de forme (les autres paramètres de source étant fixés) reste un problème d'optimisation non-linéaire complexe. Afin de réduire cette complexité tout en gardant une solution proche de l'optimum global, l'erreur $E(\theta, t_c)$ ne sera évaluée que sur un dictionnaire fini $\tilde{\Theta}$ de vecteurs de forme. La construction d'un tel dictionnaire est basée sur une mesure de similarité des différentes formes d'onde glottique, la similarité entre 2 sources de paramètres $\theta$ et $\tilde{\theta}$ étant mesurée à l'aide de leur coefficient de corrélation $\rho_{u_\theta, u_{\tilde{\theta}}}$. Ainsi, étant donné un coefficient de corrélation minimal $\rho_m$, la création du sous-ensemble $\tilde{\Theta}_{\rho_m}$ devra respecter les 2 conditions suivantes :
- condition pour couvrir l'ensemble $\Theta$ :
  $\forall \theta \in \Theta : \exists \tilde{\theta} \in \tilde{\Theta}_{\rho_m}$ tel que $\rho(u_\theta, u_{\tilde{\theta}}) \ge \rho_m$,
- condition pour éviter la redondance :
  $\forall \tilde{\theta}_1, \tilde{\theta}_2 \in \tilde{\Theta}_{\rho_m} : \rho(u_{\tilde{\theta}_1}, u_{\tilde{\theta}_2}) < \rho_m$ .

En pratique, l'algorithme 1 permet de créer un tel sous-ensemble. La figure 3 montre les variations du cardinal du sous-ensemble $\tilde{\Theta}_{\rho_m}$ en fonction de $\rho_m$.

**Algorithme 1** : Algorithme de création de $\tilde{\Theta}$

```
Θ̃ ← ∅ ;
Δ ← 0 ;
répéter
    Générer un vecteur de forme θ ;
    si max_{θ̃∈Θ̃} ρ_{u_θ̃, u_θ} < ρ_m alors
        Θ̃ ← Θ̃ ⋃ {θ};
        Δ ← 0;
    sinon
        Δ ← Δ + 1;
tant que Δ < Δ_max;
```

L'estimation des paramètres de forme sera réalisée en deux temps : la première étape consiste à minimiser l'erreur sur l'ensemble fini $\tilde{\Theta}$ puis à affiner la solution obtenue en réalisant quelques itérations d'un algorithme d'optimisation non linéaire tel que celui décrit dans [5] qui se base sur des transformations géométriques élémentaires pour minimiser une fonction de coût. Le choix de $\rho_m$ (et donc du nombre de sources dans le dictionnaire) doit être basé sur les considérations suivantes : un nombre trop élevé de sources augmente la complexité sans améliorer la solution car l'algorithme d'optimisation non linéaire permet déjà d'affiner la solution ; au contraire, un nombre trop faible de sources nuira à la qualité de l'estimation finale. Prendre $\rho_m = 0.99$ s'est révélé être un bon compromis entre ces deux considérations : le cardinal du dictionnaire est alors de $L = 280$ sources.

Pour estimer les paramètres de forme, nous sommes donc amenés à évaluer $E(\theta, t_c)$ et donc à inverser $M_u$ pour différentes valeurs de $\theta$. Une modification des paramètres de forme n'affectant que le signal de source, seule la dernière colonne de $M_u$ est modifée lorsque $\theta$ parcourt $\tilde{\Theta}$. Nous pouvons donc tabuler un certain nombre d'opérations en mettant en oeuvre une décomposition QR de $M_u$ pour le calcul de $E(\theta, t_c)$, et ainsi réduire l'ordre de complexité de $O(LNp^2)$ à $O(LNp)$ (où $L$ correspond au cardinal de $\tilde{\Theta}$) puisque seule la dernière colonne de $Q$ et de $R$ doivent être recalculées pour chaque point de $\tilde{\Theta}$. Au final, les opérations sont réparties de la manière suivante :

- pour le premier vecteur $\theta$, $O(Np^2)$ opérations sont nécessaires,
- pour les $L-1$ vecteurs suivants, le calcul de $E(\theta, t_c)$ requiert $O(Np)$ opérations.

Lors de l'estimation des paramètres, il faut générer de manière explicite le signal de source glottique LF ce qui est coûteux si on utilise directement l'équation (2). Une méthode plus rapide consiste à générer ce signal de manière récursive en utilisant un filtre AR d'ordre deux pour la phase ouverte et un filtre AR d'ordre un pour la phase de retour.

### 3.3 Résumé de la procédure d'estimation

L'utilisation d'un algorithme du type « délai de groupes » tel que celui présenté dans [6] permet d'initialiser de manière robuste la position des instants de fermeture de glotte. L'algorithme 2 détaillé dans l'encadré ci-dessous limite donc la recherche de l'instant de fermeture à un intervalle centré autour de cette position. Pour chaque position testée, l'estimation du paramètre $\theta$ optimal est effectuée suivant les deux étapes décrites dans la section précédente, à savoir une minimisation de $E(\theta, t_c)$ sur le sous-ensemble $\tilde{\Theta}_{\rho_m}$ suivi d'un raffinement par un algorithme d'optimisation non linéaire.

**Algorithme 2** : Algorithme d'estimation des paramètres de source

$T_0 \leftarrow$ Estimation par l'algorithme YIN;
$\tilde{t_c} \leftarrow$ Estimation par un algorithme de délais de groupe;
**pour** $t_c = \tilde{t_c} - \Delta$ **à** $\tilde{t_c} + \Delta$ **de pas** $\delta t_c$ **faire**
  $E_{t_c} \leftarrow \min_{\theta \in \tilde{\Theta}_{\rho_m}} E(\theta, t_c)$;
  $\theta_{t_c} \leftarrow \text{argmin}_{\theta \in \tilde{\Theta}_{\rho_m}} E(\theta, t_c)$;
  $(E_{t_c}, \theta_{t_c}) \leftarrow$ Optimisation non linéaire avec $\theta_{init} = \theta_{t_c}$;
$\hat{t_c} \leftarrow argmin_{t_c} E_{t_c}$;
$\hat{\theta} \leftarrow \theta_{\hat{t_c}}$;

## 4 Expérimentations et résultats

### 4.1 Résultats sur signaux synthétiques

La mesure de la qualité de l'estimation du vecteur de forme $\theta$ est réalisée ici sur des signaux synthétiques. Le coefficient de corrélation moyen entre la source estimée et la source théorique donne une mesure globale de la qualité de l'estimation de la forme tandis que les variances des estimateurs de $O_q$, $\alpha_m$ et $Q_a(1-O_q)$ donnent une indication de la qualité des résultats sur chacun des paramètres de forme. Les biais sur chacun des trois paramètres de forme ne sont pas donnés car ils se sont révélés négligeables sur les expériences effectuées. L'expression de la fréquence de coupure $F_a = \frac{1}{2\pi} \frac{1}{Q_a(1-O_q)T_0}$ justifie l'étude sur $Q_a(1-O_q)$ plutôt que sur $Q_a$ qui n'a pas de signification physique réelle. Les signaux synthétiques ont été générés en tirant aléatoirement les paramètres de source glottique et en utilisant de manière aléatoire un filtre AR d'ordre huit parmi sept configurations correspondant à différentes voyelles francaises. Les résultats donnés dans le tableau 1 confirment que la réduction du domaine d'exploration de $\theta$ à l'ensemble $\tilde{\Theta}$ ne nuit pas à la qualité de l'estimation, l'algorithme d'optimisation non linéaire permet juste d'améliorer la précision de l'estimation.

TAB. 1 – Résultats de l'estimation des formes sur signaux synthétiques avec $f_0 = 100$Hz, HNR=25dB et en utilisant ou non un algorithme d'optimisation non linéaire pour affiner la solution.

| OptimNL | $\sigma_{O_q}$ | $\sigma_{\alpha_m}$ | $\sigma_{Q_a(1-O_q)}$ | $\rho_{mean}$ |
|---|---|---|---|---|
| Non | 0.048 | 0.021 | 0.013 | 0.993 |
| Oui | 0.023 | 0.008 | 0.003 | 0.999 |

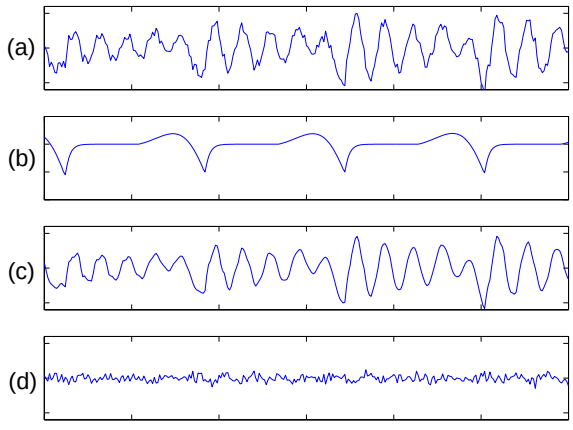


FIG. 4 – Analyse et resynthèse sur un signal de parole synthétique. (a) Signal de parole. (b) Source estimée. (c) Signal resynthétisé. (d) Différence entre le signal de parole et le signal resynthétisé.

Un signal synthétique a aussi été généré sur vingt périodes ($f0 = 100$ Hz) dans le cadre d'un test de resynthèse. La première période a été générée en utilisant les paramètres de sources

suivants : $Oq = 0.6$, $\alpha_m = 0.7$, $Q_a = 0.05$ et en utilisant un filtre représentant la voyelle 'A' tandis que la dernière période a été générée en utilisant $O_q = 0.4$, $\alpha_m = 0.85$, $Q_a = 0.05$ et un filtre AR correspondant à la voyelle 'E'. Pour les périodes intermédiaires, les paramètres de source ainsi que les coefficients LSF ont été interpolés linéairement. Comme le montre la figure 4, le signal resynthétisé en utilisant les paramètres de source estimés ne diffère de l'original que par la composante apériodique.

### 4.2 Résultats sur signaux réels

La voyelle 'A' issue du mot 'sable' prononcé par un homme a été analysée en utilisant un filtre AR d'ordre 14. La validité du résultat de l'analyse visible sur la figure 5 est partiellement confirmée par le signal DEGG (dérivée du signal électroglottographique) qui montre que les instants caractéristiques du signal de source glottique ont bien été estimés. Sur la voyelle analysée, l'erreur de prédiction est également faible, le résultat du filtrage inverse par l'AR estimé est donc très proche de la source LF estimée. D'autre part, un test informel d'écoute confirme que le signal resynthétisé est proche de l'original sans toutefois être complètement identique au niveau perceptuel, des résultats similaires ont été observés dans [7].

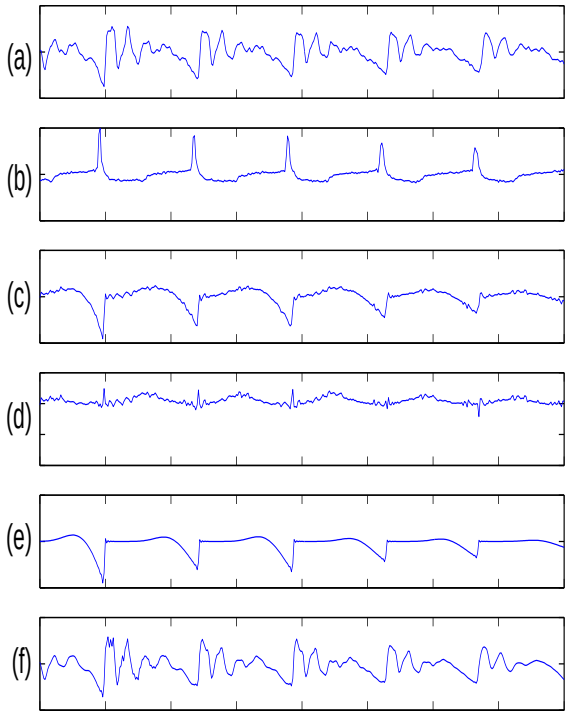


FIG. 5 – Analyse et resynthèse sur un signal de parole naturelle. (a) Signal de parole. (b) Signal DEGG. (c) Résultat du filtrage inverse. (d) Erreur de prédiction du modèle ARX. (e) Source LF estimée. (f) Signal resynthétisé.

## 5 Conclusion

L'estimation conjointe de la source glottique et du filtre AR modélisant le conduit vocal mène à un problème complexe d'optimisation non linéaire. Dans cet article, nous avons proposé une méthode efficace et robuste basée sur l'exploration systématique d'un ensemble fini de vecteurs de forme et suivie d'une procédure de raffinement local. Des optimisations algorithmiques ont également permis de réduire d'un facteur $p$ la complexité d'estimation du vecteur de forme.

La méthode proposée a été validée sur des signaux synthétiques et les expériences réalisées sur de la parole naturelle ont montré que les caractéristiques principales du signal glottique sont correctement estimées. La resynthèse n'étant toutefois pas complètement transparente, il conviendra de réaliser une modélisation acoustique du résidu, cette modélisation pouvant également servir dans le cadre de modifications ou transformations de voix de haute qualité.

## Références

[1] N. Henrich, "Etude de la source glottique en voix parlée et chantée," Ph.D. dissertation, Université de Paris 6, 2001.

[2] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, 1985.

[3] W. Ding, H. Kasuya, and S. Adachi, "Simultaneous estimation of vocal tract and voice source parameters based on an ARX model," *IEICE Trans. Inf. Syst.*, vol. E78-D, no. 6, pp. 738–743, June 1995.

[4] A. Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustic Society of America*, 2002.

[5] J. Nelder and R. Mead, "A simplex method for function minimisation," *Computer Journal*, vol. 7, pp. 308–313, 1964.

[6] A. Kounoudes, P. Naylor, and M. Brookes, "The dypsa algorithm for estimation of glottal closure instants in voiced speech," *IEEE ICASSP*, 2002.

[7] P. Hedelin, "High quality glottal LPC-vocoding," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 11, pp. 465–468, April 1986.