Estimation des instants de fermeture basée sur un coût d'adéquation du modèle LF à la source glottique

Damien Vincent [1], *Olivier Rosec* [1], *Thierry Chonavel* [2]

[1] France Telecom, Division R&D
2, Avenue Pierre Marzin - 22307 Lannion
{damien.vincent,olivier.rosec}@francetelecom.com
[2] École Nationale Supérieure des Télécommunications de Bretagne, Département Signal et Communication,
Technopôle Brest-Iroise, CS 83818, 29285 Brest Cedex, France.
thierry.chonavel@enst-bretagne.fr

ABSTRACT

An algorithm for GCI (Glottal Closure Instants) estimation is presented in this paper. It relies on a source-filter model of speech production using a LF model for the source component. From this source-filter decomposition, a ratio which measures the goodness of fit of the LF source model is introduced in the GCI estimation procedure together with fundamental frequency constraints. Then, a Viterbi algorithm is applied to extract the most likely GCI sequence. Experiments performed on a real speech database show that the proposed method outperforms existing approaches.

1. INTRODUCTION

L'estimation des instants de fermeture de glotte est un problème récurrent en traitement de la parole. Ces instants sont cruciaux en analyse du signal de parole, car leur localisation précise est nécessaire pour estimer le signal de source glottique ainsi que pour caractériser la qualité vocale. Une autre application liée à la synthèse vocale par concaténation concerne le marquage pitch-synchrone des bases de données acoustiques, opération nécessaire pour la mise en oeuvre d'algorithmes de modification prosodique tel que TD-PSOLA [6]. Pour la détermination de ces GCI, plusieurs méthodes ont été proposées telles que celles basées sur la fonction de retard de groupe [7] qui exploitent des propriétés basiques des signaux à phase minimale, ou encore celles reposant sur des algorithmes de programmation dynamiques pour estimer une séquence de GCI en accord avec une mesure préalable de la fréquence fondamentale F_0 [5].

Cependant, les approches existantes ne tiennent pas compte de façon explicite de la structure du signal glottique. Dans cet article, nous utilisons une mesure obtenue à partir d'une décomposition source-filtre du signal de parole afin de localiser les GCI potentiels et donc de mieux contraindre le problème d'estimation. Nous définissons alors une fonction de coût combinant cette information et une mesure de F_0 préalablement obtenue. Cette information combinée à une mesure de F_0 permet alors de définir une définition de coût. L'estimation de la séquence de GCI est alors obtenue par minimisation de cette séquence via un algorithme de programmation dynamique. Le papier est organisé comme suit. En section 2, nous présentons le modèle source-filtre utilisé et définissons une mesure d'adéquation à ce modèle. La section 3 détaille l'algorithme d'estimation des GCI proprement dit et la section 4 décrit les expériences destinées à valider la méthode proposée.

2. MODÈLE ARX

De nombreux modèles de production de la parole font l'hypothèse que le signal de parole résulte du filtrage linéaire de l'excitation glottique par le conduit vocal. Dans une telle décomposition source-filtre, la partie source, appelée Dérivée de l'Onde de Débit Glottique (DODG), correspond au signal produit au niveau de la glotte après prise en compte de l'effet de radiation des lèvres, approximé par une dérivation. La partie filtre désigne, quant à elle, les résonnances du conduit vocal.

Lors de la production de sons voisés, les cordes vocales entrent en vibration, ce qui se traduit par une DODG quasi-périodique. Plusieurs modèles ont été proposés pour modéliser la DODG ainsi produite. Nous considérons ici le modèle LF [4] qui permet une paramétrisation de la forme de la DODG à l'aide de trois paramètres. La figure 1 représente une onde obtenue par ce modèle.

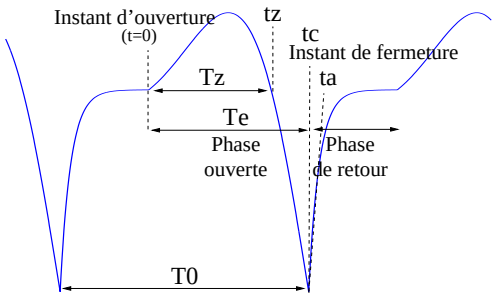

FIG. 1: Le modèle LF

Une composante stochastique est également présente pour modéliser les différents phénomènes aléatoires (irrégularité de la DODG, bruit de friction, etc...). Etant données ces hypothèses, un son $s(n)$ peut être modélisé par un processus ARX (Auto Regressive eXogenous) défini par l'équation suivante :

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

dans laquelle les a_k sont les coefficients du modèle AR caractérisant le conduit vocal, $u(n)$ désigne la DODG dont l'amplitude est contrôlée par le coefficient b_0 et $e(n)$ est le résidu.

L'estimation des paramètres du modèle ARX est très complexe, car l'optimisation selon les paramètres du modèle LF est un problème non-linéaire. Cependant, lorsque la source LF est fixée, le filtre peut être estimé par des méthodes de moindres carrés classiques. Sur la base de ce

constat, nous utilisons ici une méthode d'estimation efficace proposée dans [8]. Cette méthode consiste dans un premier temps à effectuer une recherche exhaustive dans un espace de DODG quantifiées, puis à procéder à une optimisation locale.

3. PROCÉDURE D'ESTIMATION DES GCIS

3.1. Principe de la méthode

La méthode que nous présentons dans cette section vise à estimer la séquence de GCI optimale au sens d'un critère combinant deux types de métriques. D'une part, les GCI estimés doivent être en accord avec le modèle de source défini précédemment, ce qui suggère d'associer à chaque instant potentiel, un coût cible. D'autre part, un coût dit de concaténation doit contraindre deux GCI consécutifs à être proche d'une mesure locale de la période fondamentale supposée connue : en pratique, la fréquence fondamentale sera estimée à l'aide de l'algorithme YIN [3] et en utilisant une interpolation par des splines cubiques pour obtenir une estimée pour chaque échantillon. Formellement, il s'agit donc de déterminer la séquence de GCI minimisant la fonction de coût définie par :

$$C = \sum_{l=1}^{L} C_{\text{cible}}(t_c^l) + \sum_{l=2}^{L} C_{\text{concat}}(t_c^l, t_c^{l-1}) \quad , \qquad (2)$$

où t_c^l désigne le $l^{\text{ème}}$ GCI candidat et où C_{cible} et C_{concat} désignent respectivement le coût cible et le coût de concaténation à satisfaire. Notons que le nombre L d'instants de fermeture de glotte n'est *a priori* pas connu. La résolution d'un tel problème d'optimisation peut être obtenue par programmation dynamique. Les détails relevant de l'implémentation algorithmique seront présentés en section 3.4.

3.2. Coût cible

Le modèle ARX présenté en section 2 permet de représenter le signal de parole comme la convolution d'une approximation LF de la source glottique par le filtre modélisant le conduit vocal à laquelle se rajoute un terme d'erreur de modélisation. Cette prise en compte explicite de l'information *a priori* sur la source glottique confère au modèle ARX une meilleure capacité de modélisation du signal de parole. De ce fait, l'erreur quadratique moyenne du résidu issu du modèle ARX est toujours inférieure à celle obtenue par une modélisation AR.

Cette constatation suggère de définir pour tout instant candidat t_c^l une mesure d'adéquation au modèle de la forme :

$$C_{\text{cible}}(t_c^l) = \frac{E_{LF}(t_c^l)}{E_0(t_c^l)} \quad , \qquad (3)$$

où $E_{\text{LF}}(t_c^l)$ désigne l'erreur quadratique moyenne issue du modèle ARX en utilisant la source optimale dont l'instant de fermeture est situé en t_c^l et où $E_0(t_c^l)$ est l'erreur quadratique moyenne obtenue par prédiction linéaire. Cette mesure normalisée tend vers 0 lorsque le son analysé est purement voisé et qu'il suit parfaitement le modèle LF ; elle tend vers 1 si le signal est non voisé ou si le modèle LF est très éloigné du signal glottique réel.

3.3. Le coût de concaténation

Le coût de concaténation $C_{concat}(t_c^l, t_c^{l-1})$ entre le $l^{\text{ième}}$ GCI t_c^l et le précédent t_c^{l-1} vise à pénaliser des distances $\Delta t_c^l = t_c^l - t_c^{l-1}$ entre ces deux GCI trop éloignées de la période fondamentale estimée en t_c^l. Le coût induit doit cependant être en adéquation avec la confiance accordée à la période fondamentale estimée : lorsque cette confiance est faible, l'estimation des GCIs sera d'avantage basée sur le coût cible ; tandis que si la confiance est élevée et si le coût cible ne permet pas de discriminer les instants de fermeture, le coût de concaténation sera privilégié, le processus d'estimation s'apparente dans ce cas à un mécanisme d'interpolation du GCI courant à partir du précédent.

Le distance CMNDF introduite dans [3] est ainsi utilisée comme mesure de confiance pour moduler le coût de concaténation. Cette distance est définie par :

$$d_n^2(\tau) = \begin{cases} 1 & \text{si } \tau = 0, \\ \frac{d_n(\tau)}{\sum_{k=1}^{\tau} d_n(k)} & \text{sinon,} \end{cases}$$

où $d_n(\tau) = \sum_{k=-K}^{K} \left(s(n+k) - s(n+k-\tau)\right)^2$ correspond à la fonction différence sur une fenêtre de longueur $2K+1$.

La modulation du coût de concaténation par cette mesure de confiance sera réalisée en deux étapes. Tout d'abord, la distance CMNDF sert à définir les périodes minimale et maximale entre deux GCI consécutifs, les fréquences associées étant obtenues à partir de la fréquence fondamentale estimée $f_0(t_c^l)$ par :

$$\begin{array}{l} \frac{f_0(t_c^l)}{f_0^{\min}} = \frac{f_0^{\max}}{f_0(t_c^l)} \\ \ln\left(\frac{f_0^{\max}}{f_0(t_c^l)}\right) = \gamma \frac{\min(d^2_{t_c^l}(T_0(t_c^l));1)+\delta}{1+\delta} \end{array} \quad . \qquad (4)$$

Le paramètre δ autorise une certaine variation de la période entre 2 GCI même si la distance CMNDF est nulle (c'est à dire correspondant à une confiance très élevée) tandis que le paramètre γ correspond à un facteur d'échelle : en pratique, $\delta = 0.15$ et $\gamma = 0.53$ ce qui donne $\frac{f_0^{\max}}{f_0(t_c^l)} = 1.07$ pour $d^2_{t_c^l}(T_0) = 0$ et $\frac{f_0^{\max}}{f_0(t_c^l)} = 1.70$ pour $d^2_{t_c^l}(T_0) \geq 1$. A partir de ces fréquences minimale et maximale, nous en déduisons un premier coût de concaténation représenté sur la figure 2 et donné par :

$$C^1_{\text{concat}}(t_c^l, t_c^{l-1}) = g\left(\frac{f_s}{t_c^l - t_c^{l-1}}\right) \quad . \qquad (5)$$

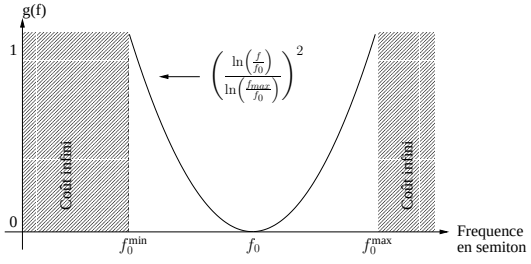

FIG. 2: Pénalité appliquée en fonction de l'écart par rapport à la fréquence fondamentale estimée.

Pour rendre l'estimation des GCI plus robuste aux erreurs d'estimation de la fréquence fondamentale, une seconde

modulation du coût de concaténation définie par

$$C^2_{\text{concat}}(t^l_c, t^{l-1}_c) = \min\left(d^2_{t^l_c}(\Delta t_c(l)) - \min_{\tau} d^2_{t^l_c}(\tau); 1\right) \quad (6)$$

est introduite de manière à favoriser des périodes Δt^l_c correspondant à des valeurs faibles de la fonction CMNDF. Nous obtenons au final la fonction de concaténation suivante :

$$C_{\text{concat}}(t^l_c, t^{l-1}_c) = C^1_{\text{concat}}(t^l_c, t^{l-1}_c)\, C^2_{\text{concat}}(t^l_c, t^{l-1}_c) \quad (7)$$

La figure 3 illustre l'intérêt du coût C^2_{concat} : la période T_1 se trouve favorisée tout autant que la période estimée T_0.

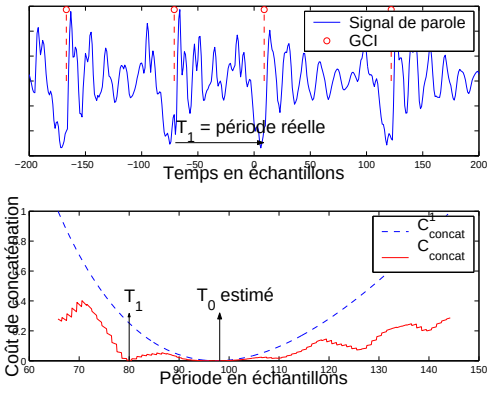

FIG. 3: Exemple de coût de concaténation sur une zone où les GCI du signal de parole sont irrégulièrement espacés

3.4. *Considérations pratiques*

Le nombre L de GCIs n'étant pas connu a priori, se pose le problème de la terminaison de l'algorithme. Un critère basé sur la minimisation de C (défini par l'équation 2) vis à vis de L n'est pas valable car les fonctions de coût cible et de concaténation sont toujours positives : le minimum de C par rapport à L serait donc obtenu en prenant $L = 1$ et en ne sélectionant que le GCI qui correspond au minimum global du coût cible. Le problème de terminaison peut se résoudre en considérant le problème d'estimation des GCI comme un problème de plus court chemin : les premier et dernier GCI sont d'abord contraints à être respectivement dans un intervalle de départ $[t^s_1, t^s_2]$ et d'arrivée $[t^e_1, t^e_2]$, l'algorithme d'estimation doit ensuite déterminer le plus court chemin entre ces 2 intervalles en plaçant L GCI le long du chemin optimal.

L'espace d'état associé à l'algorithme de programmation dynamique est composé de $N = t^e_2 - t^s_1$ échantillons. La complexité de chaque itération n'est cependant pas $O(N^2)$ car pour un état courant donné, un grand nombre d'états précédents est interdit : la complexité est réduite à $O(N(T^{max}_0 - T^{min}_0))$ où T^{max}_0 et T^{min}_0 sont les périodes associées à f^{min}_0 et f^{max}_0 définis par l'équation 4. Pour réduire d'avantage la complexité, l'algorithme pourra être appliqué sur chaque partie voisée du signal séparément ce qui permet de réduire le nombre d'états.

4. TESTS

L'évaluation a été réalisée sur la base *arctic* [1] qui fournit à la fois les signaux de parole et les enregistrements électroglottographiques (EGG) correspondants. La dérivee du signal EGG (DEGG) permet d'extraire facilement les instants de fermeture réels car ceux-ci correspondent à des pics très marqués sur ce signal. Ces GCI de référence sont nécessaires pour évaluer de manière objective les performances de notre algorithme.

4.1. *Détermination des GCI de référence*

L'évaluation étant réalisée sur une base entière de signaux de parole, une méthode manuelle d'extraction des pics correspondant aux GCI n'est pas réaliste. L'algorithme 1 permet de déterminer automatiquement les GCI, les GCIs étant extraits selon une confiance décroissante : la confiance est mesurée directement par l'amplitude du pic correspondant sur le signal DEGG. Il est à noter que l'algorithme utilise une estimée de la période fondamentale T_0 ce qui pourrait le rendre dépendant des performances de l'estimateur de T_0 ; l'algorithme s'avère cependant robuste aux erreurs d'estimation de T_0.

Algorithme 1 : Extraction de l'ensemble G des GCI de référence

$\forall t : c(t) = 1$
Tant que $M = \max(c \otimes y) \geq$ *seuil* **faire**
  $t_c \leftarrow \text{argmax}(c \otimes y)$
  $G \leftarrow G \bigcup \{t_c\}$
  $t_1 \leftarrow \max\left\{t \leq t_c - 1\text{ms} / \left(1 - \frac{t_c - t}{T_0(t_c)}\right) M \leq y(t)\right\}$
  $t_2 \leftarrow \min\left\{t \geq t_c + 1\text{ms} / \left(1 - \frac{t - t_c}{T_0(t_c)}\right) M \leq y(t)\right\}$
  $c([t_1; t_2]) = 0$

4.2. *Critères de performance*

A l'aide de l'algorithme 2, les quatre ensembles suivants sont construits : l'ensemble des indices de référence non détectés (ND), l'ensemble des GCI estimés qui sont des fausses alarmes (FA), l'ensemble contenant les appariements de GCI de référence et estimé qui correspondent à des erreurs supérieures à 2.5ms (erreurs grossières EG) et enfin l'ensemble E correspondant aux couples de GCI de référence et estimés dont l'erreur d'estimation est inférieure à 2.5ms. A partir de ces ensembles sont obtenus les mesures de performance suivantes : le taux de non détections TND $= \frac{\text{card}\{ND\}}{N_r}$ où N_r correspond au nombre de GCI de référence, le taux de fausses alarmes TFA $= \frac{\text{card}\{FA\}}{N_r}$, le taux d'erreurs grossières TEG $= \frac{\text{card}\{EG\}}{N_r}$ et la variance sur E de l'erreur d'estimation des GCI.

Algorithme 2 : Association entre les GCI de référence et les GCI estimés

K ensemble des indices des GCI de référence
L ensemble des indices des GCI estimés
Tant que $K \neq \emptyset$ *et* $L \neq \emptyset$ **faire**
  $(k_m, l_m) = \text{argmin}_{(k,l) \in K \times L}\, t_c(k) - \hat{t}_c(l)$
  $\Delta = t_c(k_m) - \hat{t}_c(l_m)$
  $K \leftarrow K \backslash \{k_m\}$ et $L \leftarrow L \backslash \{l_m\}$
  Si $\Delta >$ 5ms : ND $=$ ND $\bigcup k_m$ et FA $=$ FA $\bigcup l_m$
  Si $\Delta \in [2.5; 5\text{ms}]$: EG $=$ EG $\bigcup (t_c(k_m), \hat{t}_c(l_m))$
  Si $\Delta < 2.5ms$: $E = E \bigcup (t_c(k_m), \hat{t}_c(l_m))$
ND $=$ ND $\bigcup K$
FA $=$ FA $\bigcup L$

Deux types d'évaluation sont réalisés : l'une en utilisant

l'ensemble des GCI $t_c(l)$ de référence, l'autre en n'utilisant que les GCI qui sont régulièrement espacés. Le GCI l est dit irrégulier si l'une des 2 conditions suivantes est réalisées : $\frac{t_c(l)-t_c(l-1)}{(t_c(l+2)-t_c(l-2))/4} \notin [0.8; 1.2]$ ou $\frac{t_c(l+1)-t_c(l)}{(t_c(l+2)-t_c(l-2))/4} \notin [0.8; 1.2]$; c'est à dire si la période à gauche ou à droite dévie trop de la période moyenne.

4.3. Résultats

Les performances de l'algorithme proposé sont comparées à celles de l'algorithme DYPSA [5]. L'estimation des GCI par DYPSA est basée sur l'utilisation des délais de groupe pour déterminer une liste d'instants de fermeture candidats : la fonction de délai de groupe EW (*Energy Weighed Group Delay*) définie dans [2] est appliquée au résidu LPC en utilisant une longueur de fenêtre de 10ms pour les voix d'homme et de 7ms pour les voix de femme. La séquence la plus probable est ensuite déterminée à l'aide d'un algorithme de programmation dynamique qui permet entre autres d'introduire des contraintes de régularité des périodes fondamentales obtenues à partir de la différence entre deux GCI consécutifs.

Le tableau 1 présente les écarts-types, les taux d'erreurs grossières, de non détections et de fausses alarmes pour l'algorithme proposé et la méthode DYPSA. Dans la configuration de test C1 comprenant les GCI irréguliers, l'algorithme DYPSA présente un taux de mauvaise détection plus élevé (3.42%). En supprimant les GCI irréguliers des statistiques, les performances des deux algorithmes sont bien meilleures, l'algorithme proposée présente cependant une variance d'estimation plus faible que DYPSA et les taux TEG, TFA et TND sont également plus faibles. La figure 4 montre sur un exemple que l'algorithme proposé est capable d'estimer correctement les GCI sur les zones stationnaires mais aussi sur des zones où les GCI sont irrégulièrement espacés. A notre sens, deux raisons peuvent expliquer les meilleures performances de notre algorithme. Tout d'abord, la contrainte de régularité des GCI est beaucoup plus souple dans l'algorithme proposé ce qui permet d'obtenir de bons résultats sur les zones où les GCI sont irrégulièrement espacés ; ensuite, le coût cible utilisé semble être plus discriminant que les fonctions de délai de groupe qui d'une part amènent à prendre en compte certains instants d'ouverture très prononcés comme GCI candidat, d'autre part peuvent aboutir à une mauvaise précision lorsque l'instant de fermeture est peu marqué.

Test	Algorithme	σ	TEG	TFA	TND
C1	Proposé	0.37	0.73	0.95	0.57
	DYPSA	0.38	1.20	1.35	3.42
C2	Proposé	0.25	0.09	0.09	0.08
	DYPSA	0.30	0.35	0.25	0.90

TAB. 1: Comparaison des deux méthodes : variance d'estimation (en ms), taux TEG, TFA et TND (en %) pour la configuration de test C1 (utilisant tous les GCI de références) et la configuration de test C2 (utilisant uniquement les GCIs régulièrement espacés).

5. CONCLUSION

La méthode proposée permet d'estimer les instants de fermeture avec une bonne précision, tout en gardant des taux de fausses alarmes et de non détections faibles. Ces bonnes

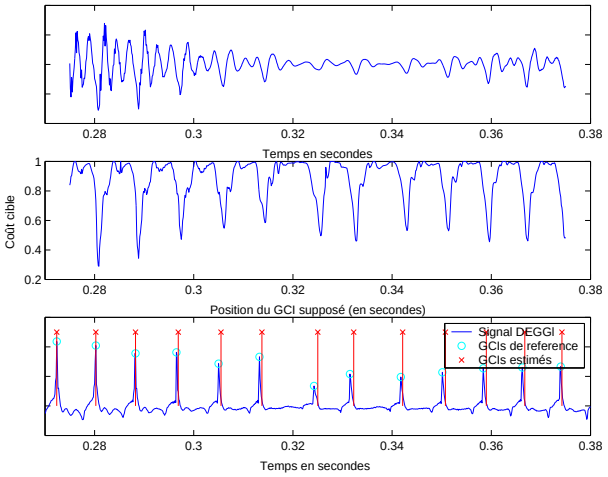

FIG. 4: Résultat de l'estimation des GCI sur un signal correspondant au mot anglais '*was*'. De haut en bas : (a) le signal de parole, (b) le coût cible $C_{\text{cible}}(t_c)$, (c) le signal DEGG et les GCI estimés.

performances sont imputables : i) au choix d'un coût cible en accord avec les mécanismes de production de la parole et ii) à l'utilisation de contraintes de continuité permettant un bon compromis entre le respect de la période fondamentale estimée et l'adéquation au coût cible. A partir de ces coûts, la séquence optimale de GCI peut être déterminée en utilisant un algorithme de plus court chemin appliqué sur l'ensemble des échantillons du signal. Des études complémentaires restent néanmoins nécessaires afin d'une part de caractériser plus précisément les erreurs produites par l'algorithme proposé et d'autre part de valider les résultats obtenus sur d'autres bases de parole.

RÉFÉRENCES

[1] Arctic speech database. http ://festvox.org/cmu_arctic/.

[2] M. Brookes, P.A. Naylor, and J. Gudnason. A quantitative assessment of group delay methods for identifying glottal closures in voiced speech. *IEEE Trans. on Speech and Audio Processing*, 2006.

[3] A. de Cheveigné and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.*, 111(4) :1917–1930, 2002.

[4] G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. *STL-QPSR*, 4 :1–13, 1985.

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

[6] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9 :453–467, 1990.

[7] R. Smits and B. Yegnanarayana. Determination of instants of significant excitation in speech using group delay function. *IEEE Trans. on Speech and Audio Processing*, 3(5) :325–333, 1995.

[8] D. Vincent, O. Rosec, and T. Chonavel. Estimation of LF glottal source parameters based on ARX model. *Interspeech*, pages 333–336, 2005.