Gilles Degottex (gilles.degottex@ircam.fr), Axel Röbel, Xavier Rodet
Équipe Analyse-Synthèse - CNRS-UMR9912-STMS
IRCAM – Institut de Recherche et Coordination Acoustique/Musique – Paris

Estimation du filtre du conduit-vocal adaptée à un modèle d'excitation mixte pour la transformation et la synthèse de la voix

CONTEXTE

- « Séparation de la source glottique des influences du conduit vocal » (Ph.D.)
- Recherche principale: Méthodes d'estimation de paramètres de la source glottique (Rd[Degottex09a,Lu02], GCI[Degottex09b])
- Applications: Transformation / Synthèse de la parole / Conversion d'identité

RESUME

En transformation et synthèse de la voix, le filtre du conduit-vocal est habituellement supposé être excité par un spectre d'amplitude plat. Nous proposons d'utiliser un modèle de source mixte: un modèle de Liljencrants-Fant (LF) et un bruit Gaussien. L'estimation du conduit-vocal doit donc être adapté à cette source en prenant en compte les amplitudes du modèle LF dans les basses fréquences et le bruit dans les hautes fréquences. Le modèle de production vocal résultant peut ensuite être utilisé pour contrôler indépendamment le conduit-vocal et la source dans le cadre de la transformation de la voix et pour l'apprentissage de ses paramètres dans la synthèse vocale HMM.

MODELE DE PRODUCTION VOCAL

$$S(\omega) = \left(H^{f_0}(\omega) \cdot G^{Rd}(\omega) + N^{\sigma_g}(\omega)\right) \cdot C^{c}(\omega) \cdot L(\omega)$$

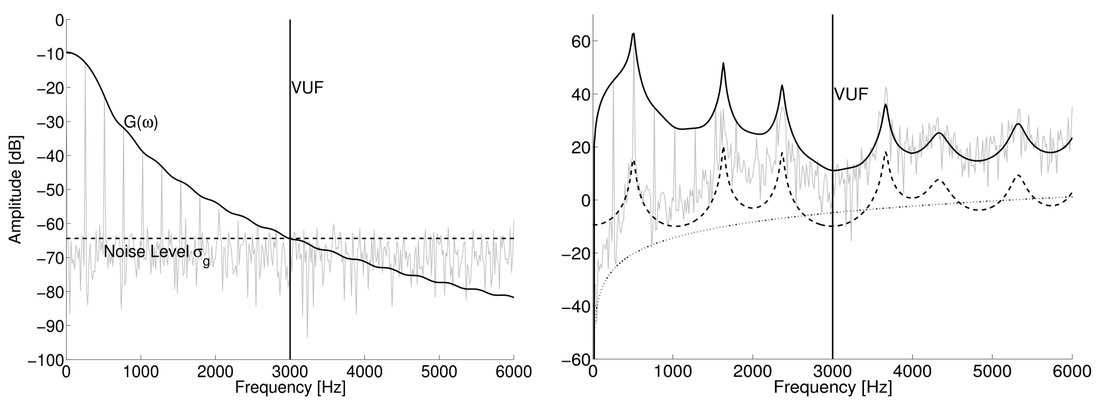
H^{f0} est une structure harmonique due à la périodicité en *1/f0* du signal de parole.

G^{Rd} est la source déterministe. Un spectre à phase mixte définit par le modèle LF [Fant85]. Dans nos applications, ce modèle est contrôlé par un paramètre de forme *Rd* [Lu02], le gain de cette source *Ee* et la période fondamentale *1/f0*.

$N^{\sigma g}$ est constitué d'un bruit blanc Gaussien d'écart-type σ_g.

C^{c} est le Filtre du Conduit-Vocal (VTF), un filtre à phase-minimal paramétrisé par des coefficients cepstraux *c*.

L est la radiation aux lèvres (≈dérivée temporelle L(ω)=jω)



PARAMETRES DETERMINISTES: *f0, Rd, Ee*

f_0 est calculé à partir du signal de parole (ex. YIN, Swipep, *Harmonic matching* [Yeh08]).

Rd est calculé en négligeant l'influence de la source sur les phases du conduit vocal [Degottex09a]. (robuste mais biaisé). Différentes méthodes peuvent être utilisées pour estimer ce paramètre (ex. IAIF, ZZT).

Ee est complètement dépendant de 2 autres gains: σ_g et la moyenne log-amplitude du VTF *c(0)*. Une contrainte est donc nécessaire. Nous utilisons: $G^{Rd}(\omega)$ normalisé par $G^{Rd}(0)$. *Ee* n'est donc pas nécessaire.

PARAMETRE STOCHASTIQUE: σ_g

Une estimation de VUF (Voiced/Unvoiced Frequency) est utilisée pour séparer S(ω) en une partie déterministe et une partie stochastique [Kim07] (voir figure). On suppose donc que $G^{Rd}(\omega)$ croise l'amplitude moyenne du bruit à hauteur de VUF. Par conséquent, σ_g peut être déduit à partir de VUF et $G^{Rd}(\omega)$

$$\sigma_g = |G^{Rd}(VUF)| \frac{\sqrt{2}}{\sqrt{\pi/2} \cdot \sqrt{\sum_t \mathrm{win}[t]^2}}$$

Cette partie du spectre est supposée excitée par un bruit Gaussien. Son écart-type doit donc être calculé à partir de l'amplitude espérée du modèle LF $|G^{Rd}(\omega)|$ en respectant la distribution de Rayleigh des amplitudes spectrales (*i.e.* $\sqrt{2}/\sqrt{\pi/2}$) [Yeh08].
De plus, le niveau de bruit est inversement proportionnel à l'énergie de la fenêtre d'analyse. La normalisation par $\sqrt{\Sigma_t \mathrm{win}[t]^2}$ est donc nécessaire pour rendre la mesure de σ_g indépendante de l'analyse.

ESTIMATION DU FILTRE DU CONDUIT VOCAL

Les deux bandes (déterministe et stochastique) séparées par le VUF sont modélisées par deux enveloppes différentes:

Dans la bande déterministe: la contribution de la radiation aux lèvres et de la source glottique sont retirées du spectre du signal observé S(ω) par division dans le domaine fréquentiel (eq. ci-dessous). Puis une enveloppe cepstrale $\mathscr{T}^o$ est alignée sur le sommet des harmoniques à l'aide d'une méthode itérative [Roebel05].

Dans la bande stochastique: S(ω) est divisé par L(ω) et G^{Rd}(VUF) pour assurer une continuité entre les deux bandes fréquentielles. Le cepstre de puissance $\mathscr{C}^o$ permet ensuite d'obtenir l'enveloppe de bruit qui doit être finalement alignée sur l'amplitude espérée en échelle linéaire calculée dans la partie déterministe (*i.e.* $\sqrt{\pi/2}/e^{0.058}$) [Yeh08].

$$C(\omega) = \begin{cases} \mathscr{T}^o\left(\frac{S(\omega)}{L(\omega)G^{Rd}(\omega)}\right) & \text{if } \omega < VUF \\ \mathscr{C}^o\left(\frac{S(\omega)}{L(\omega)G^{Rd}(VUF)}\right) \cdot \frac{\sqrt{\pi/2}}{e^{0.058}} & \text{if } \omega \geq VUF \end{cases}$$

Les enveloppes ne doivent pas modéliser la structure harmonique. Leur ordre est donc o=0.5·fs/f0.
Même s'il n'y a pas d'harmoniques dans la partie stochastique, des partiels distants de f0 (mais pas multiples de f0) apparaissent car le bruit glottique est modulé par l'aire glottique [Lu02].

CONCLUSIONS

- Les amplitudes du spectre observé |S(ω)| sont toujours modélisées car l'estimation du VTF ne fait que compléter le spectre de la source. Les phases sont par contre imposées par le modèle LF, le bruit blanc et les phases minimales du VTF.
- Lorsque le VTF est modifié (ex. étirement fréquentiel), les amplitudes de la source (ex. formant glottique) sont conservées. En terme de transformation, ces deux éléments sont donc indépendants.
- Contrôle simple du modèle de source (f0, Rd, σ_g).

EXEMPLES sur www.ircam.fr/anasyn/degottex

REFERENCES
[Fant85] G. Fant, J. Liljencrants, and Q. Lin, “A four-parameter model of glottal flow”, *STL-QPSR*, vol. 4, pp. 1-13, 1985.
[Lu02] H.-L. Lu, Toward a High-quality Singing Synthesizer with Vocal Texture Control, Ph.D. thesis, Stanford, 2002.
[Degottex09a] G. Degottex, A. Roebel, X. Rodet, "Shape parameter estimate for a glottal model without time position", SPECOM, p. 345--349, 2009.
[Roebel05] Roebel, A. and Rodet, X., "Efficient Spectral Envelope Estimation And Its Application To Pitch Shifting And Envelope Preservation", in DAFx, 2005.
[Degottex09b] G. Degottex, A. Roebel, X. Rodet, "Glottal Closure Instant detection from a glottal shape estimate", SPECOM, p. 226—231, 2009.
[Yeh08] Yeh, C., "Multiple fundamental frequency estimation of polyphonic recordings," Ph.D. thesis, UPMC - Paris VI, 2008.
[Kim07] Kim, S.-J. and Hahn, M., "Two-Band Excitation for HMM-Based Speech Synthesis", IEICE - Trans. Inf. Syst., vol. E90-D, no. 1, pp. 378--381, 2007.