# Amélioration de la précision de la resynthèse avec TD-PSOLA

*Vincent Colotte et Yves Laprie*

LORIA
Campus scientifique, BP 239, F-54506 Vandœuvre-lès-Nancy, FRANCE
Tél.: ++33 (0)3 83 59 20 74 - Fax: ++33 (0)3 83 41 30 79
Mél: Vincent.Colotte@loria.fr, Yves.Laprie@loria.fr - http://www.loria.fr/∼colotte/

## ABSTRACT

The paper describes techniques to improve the precision of prosodic modifications with TD-PSOLA. TD-PSOLA relies on the decomposition of the signal into overlapping frames synchronised with pitch period. The main objective is thus to preserve the consistency of marks between neighbouring frames with respect to the temporal structure of pitch periods. First, we improve pitch marking by eliminating mismatch errors which appear during rapid formant transitions. This is achieved by pruning pitch mark candidates. From the synthesis point of view we exploit a fast re-sampling method which allows signal frames to be shifted finely. Together with the pitch marking improvement, this fast re-sampling method enables very high quality transformations characterised by the absence of noise between harmonics.

## 1. INTRODUCTION

Les méthodes reposant sur le principe de synchronisation avec le fondamental sont utilisées pour réaliser des modifications temporelles ou fréquentielles d'un signal de parole, ou pour mettre en œuvre des systèmes de synthèse à partir du texte. Ces méthodes nécessitent au préalable un marquage des périodes du fondamental. Nos travaux se situent dans le cadre de l'apprentissage des langues, sur la modification des paramètres prosodiques. C'est ainsi que nous avons choisi d'utiliser la méthode TD-PSOLA (*Time Domain - Pitch Synchronous OverLap and Add*) : elle est rapide et permet de modifier simultanément F0 et le débit d'un signal de parole.

TD-PSOLA [6] est basée sur la décomposition du signal de parole en fenêtres recouvrantes synchronisées sur les périodes du fondamental. Les marques de synchronisation (ou du fondamental) indiquent le centre de ces fenêtres. Les modifications consistent à manipuler les marques d'analyse pour générer de nouvelles marques de synthèse. Cela correspond à la duplication ou l'élimination de fenêtres dont l'écartement peut être modifié.

La principale exigence, commune aux techniques de décomposition du signal en courtes fenêtres, est de garder la cohérence mutuelle des emplacements des marques pour préserver la structure temporelle originale du signal étudié. Par conséquent, il est crucial d'obtenir un marquage précis des périodes du fondamental car il influe directement sur la qualité du signal.

De nombreuses méthodes de marquage ont été décrites dans la littérature. Elles sont généralement basées sur la recherche d'évènements précis dans le signal de parole : instants de fermeture glottale [1], extrema du signal, instants d'excitation des modèles LPC, dernier passage par zéro avant un pic maximum…

Comme le souligne Veldhuis [8], ces techniques souffrent d'une certaine rigidité face au critère numérique exploité. En particulier, le critère numérique peut forcer le marquage d'échantillons qui satisfont ce critère mais dont la distance avec les marques voisines est éloignée de la période du fondamental. Dans le contexte de la synthèse de parole, il est concevable de corriger quelques erreurs à la main, ce qui n'est plus possible si l'on souhaite modifier des phrases de manière automatique pour l'apprentissage des langues.

Dans [3], nous avons proposé un algorithme de marquage qui exploite les résultats de l'extraction de F0 et assure la cohérence des marques sur l'ensemble de la phrase. Veldhuis [8] a proposé un algorithme élargissant légèrement notre approche qui, à partir d'un pré-marquage, prend en compte la corrélation du signal dans le voisinage des marques du fondamental.

L'amélioration de la robustesse et de la précision du marquage du fondamental est le premier objet de cet article.

Outre la qualité du marquage du fondamental, l'emplacement des marques de synthèse elles-mêmes peut être à l'origine de dégradations temporelles. Dans une première approche, les marques de synthèses peuvent être choisies parmi les instants d'échantillonnage. Pour les mêmes raisons que celles mentionnées précédemment, les marques de synthèse doivent être placées avec précision pour éviter les erreurs de phase. La figure 1 montre ce phénomène : une période du son /ε/ à 16 KHz a été dupliquée et la fréquence fondamentale du stimulus a été modifiée linéairement. Les marques d'analyse ont été placées exactement à la même place à l'intérieur de chaque période. Le premier spectrogramme montre le résultat obtenu avec l'algorithme classique de TD-PSOLA et le second avec une meilleure précision des marques de synthèse. Dans le premier cas, nous pouvons observer la répercussion des erreurs de phase sur la qualité des harmoniques.

Il est donc important de développer des algorithmes offrant une meilleure précision du marquage du fondamental et de la localisation des marques de synthèse. Tout d'abord, nous décrirons les améliorations apportées à l'algorithme proposé dans [3]. Ensuite, nous expliquerons comment le marquage précis peut être combiné avec une technique de ré-échantillonnage pour obtenir une meilleure précision de synthèse avec TD-PSOLA.

## 2. MARQUAGE DU FONDAMENTAL

### *2.1. Principe*

Le principe de notre algorithme est de sélectionner les marques du fondamental parmi les extrema locaux.

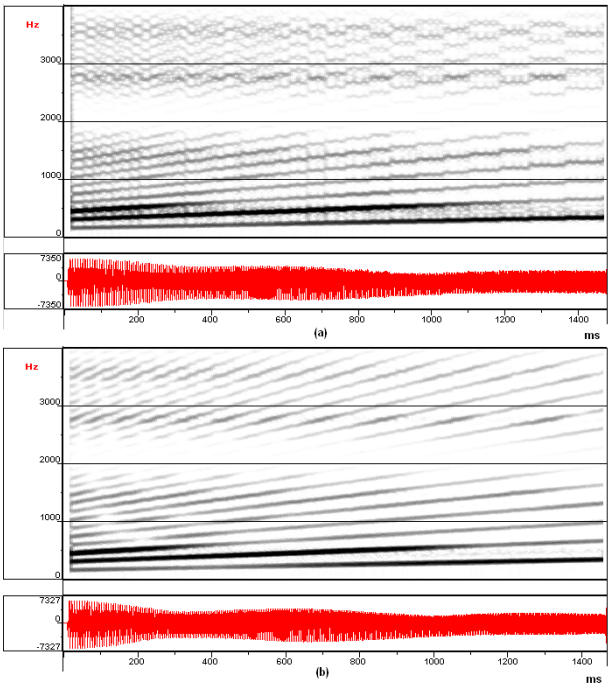


**Figure 1:** Le stimulus a été construit à partir d'une période du son /ɛ/. Sa F0 a été modifiée linéairement : (a) avec un algorithme classique de TD-PSOLA, (b) avec une meilleure précision de la synthèse.

Soit un ensemble de marques candidates qui sont toutes des pics négatifs, ou toutes des pics positifs :

$$C = \{c(i)\} = \{c(1) \dots c(i) \dots c(N)\}$$

où $c(i)$ est l'instant (en échantillons) du $i^{ème}$ pic, et $N$ le nombre de pics extraits (voir [3] pour plus de détails sur la recherche des candidats).

Les marques du fondamental constituent un sous-ensemble de $C$, et sont espacées par des périodes du fondamental données par un algorithme d'extraction de F0. Cette sélection peut être représentée par une séquence d'indices :

$$J = \{j(k)\} = \{j(1) \dots j(k) \dots j(K)\}$$

avec $K < N$. $J$ doit préserver l'ordre chronologique nécessitant la monotonie de $j$: $j(k) < j(k+1)$.

La séquence des indices parmi les pics correspondants définit l'ensemble des marques du fondamental :

$$\overline{C} = \{c(j(k))\} = \{c(j(1)) \dots c(j(k)) \dots c(j(K))\}$$

La détermination de $j$ nécessite un critère exprimant la fiabilité de deux marques consécutives par rapport à la valeur de F0 précédemment extraite. Le critère local est choisi comme suit :

$$d(c(l), c(i)) = |(c(i) - c(l)) - pitchPeriod(c(l))| \quad (1)$$

où $l < i$. Il prend en compte l'intervalle entre deux marques comparé à la période du fondamental. Ce critère retourne zéro si les deux pics sont exactement séparés de $pitchPeriod(c(l))$ et une valeur positive si les deux pics sont plus éloignés ou plus proches qu'une période de F0.

Le critère global est :

$$D = \sum_{k=1}^{K-1} d(c(j(k)), c(j(k+1))) - B(c(j(k+1)) \quad (2)$$

où $B$ est le bonus sélectionnant de préférence un extremum comme marque du fondamental. Dans un premier temps, nous avons choisi $B(c(j(k)) = \gamma \times |amplitude(c(j(k)))|$. Le coefficient $\gamma$ exprime le compromis entre la distance séparant le pic candidat d'un pic précédent (devant être proche de la période du fondamental) et l'amplitude du pic étudié. $D$ est minimisé par programmation dynamique.

Nous avons utilisé l'algorithme d'extraction de F0 proposé par Martin [4] pour évaluer le critère local défini dans Eq.1. Le signal a été filtré avec un filtre passe-bas de fréquence 2500 Hz.

### 2.2. *Amélioration de l'algorithme de marquage*

Le coefficient $\gamma$ qui contrôle le compromis entre la proximité avec la valeur du fondamental et l'amplitude de la marque a été expérimentalement fixé à $1/400$. Quand $\gamma$ est trop fort de nombreux pics sont gardés comme marques (voir la figure 2a à 1144 ms). La valeur retenue pour $\gamma$ est suffisamment générale pour donner de bons résultats avec la plupart des signaux de parole. Cependant, il apparaît que ce choix n'est pas approprié pour certains signaux contenant des transitions spectrales importantes.

Nous avons mis en place deux solutions. La première consiste à exploiter une fonction de similarité pour éviter les problèmes de déphasage entre les extrema voisins. Nous avons donc utilisé la corrélation comme bonus. Le bonus de l'Eq. 2 devient alors :

$$B(c(j(k))) = \gamma \times (\delta |amplitude(c(j(k))| + \delta' corr_n(c(j(k-1)), c(j(k))) \quad (3)$$

où $corr_n(c(j(k-1), c(j(k)))$ est la corrélation entre les segments de longueur $n$, centrés en $c(j(k-1))$ et $c(j(k))$. La durée, sur laquelle a été calculée la corrélation, est égale à une période du fondamental à l'instant $c(j(k-1))$. Les coefficients $\delta$ et $\delta'$ pondèrent indépendamment l'amplitude et la corrélation.

Le calcul de la corrélation pour chaque paire de candidats augmente fortement le temps d'exécution du marquage du fondamental. De plus, la corrélation est utilisée pour corriger des erreurs « évidentes » qui correspondent à l'association de deux extrema locaux dont l'écart diffère significativement de la période du fondamental local.

Pour ces raisons, nous avons mis au point une autre solution qui repose sur une stratégie d'élagage. Sans éliminer les candidats potentiels, il est intéressant d'augmenter drastiquement le poids de la distance entre les candidats comparée à la période du fondamental, quand cette distance est trop éloignée de la période du fondamental. De cette façon, le critère local (Eq.1) est privilégié par rapport au bonus.

### 2.3. *Résultats du marquage*

Dans un premier temps, nous avons testé le bonus basé seulement sur la corrélation. Le ratio entre la corrélation et le critère local est alors 40. Il s'avère que le compromis entre la distance et la corrélation donne approximativement le même nombre d'erreurs que la stratégie initiale. En effet, cette nouvelle stratégie favorise la sélection de deux pics très proches l'un de l'autre car la corrélation est alors proche de 1.

Ensuite, nous avons testé un bonus intégrant l'amplitude et la corrélation. Leur pondération respective face au critère local a été conservée. Les résultats sont très satisfaisants pour la plupart des signaux mais quelques erreurs ne peu-

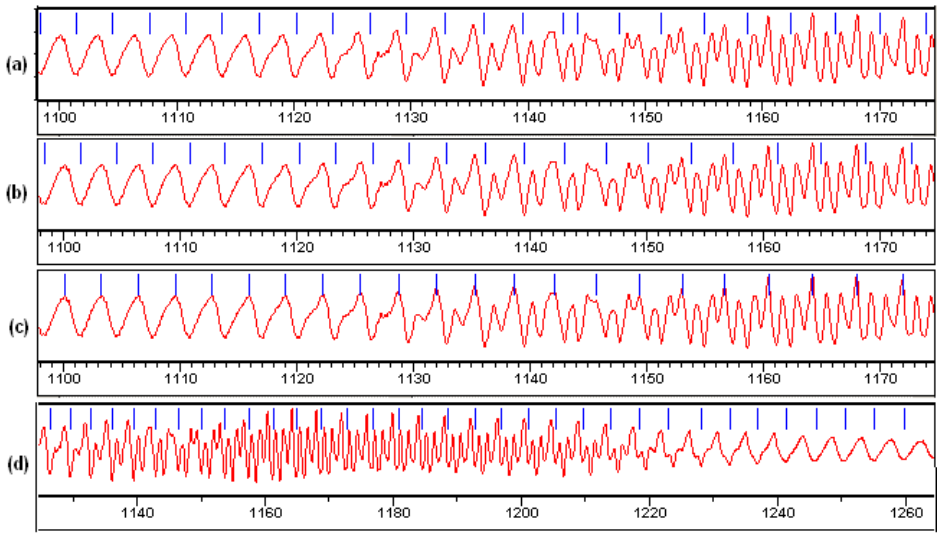


**Figure 2:** (a) Erreurs de marquage avec l'amplitude comme bonus sans l'élagage. (b) Marquage avec élagage et l'amplitude comme bonus. (c) Marquage avec élagage et la corrélation comme bonus. (d) Marquage sur un segment de parole avec des transitions de formants.

vent être éliminées. De plus, comme nous l'avons mentionné précédemment, ces erreurs sont évidentes car la distance entre deux marques est très éloignée de la période du fondamental. Ce problème provient de l'algorithme de programmation dynamique qui favorise localement des erreurs « grossières » si elles contribuent à abaisser le critère global sur l'ensemble de la partie à traiter. Ainsi, l'élagage a été mis en place en augmentant brutalement le poids du critère local pour les paires de candidats dont l'écart est de $20\%$ supérieur ou inférieur à la valeur attendue du fondamental. $\gamma$ est alors égal à $1/40000$ au lieu de $1/400$ dans l'Eq.2. Ce choix arbitraire empêche l'algorithme de programmation dynamique de sélectionner deux pics incohérents l'un par rapport à l'autre. De plus, la valeur de $20\%$ permet d'éliminer les erreurs grossières tout en préservant la possibilité de choisir des pics dont la distance est légèrement différente de la période du fondamental calculée, mais, qui sont cohérents avec le critère de corrélation. Ce choix ne nécessite pas de réglages fins ou dépendant du signal analysé.

Les résultats obtenus avec la stratégie d'élagage et avec l'amplitude, ou avec la corrélation ou avec les deux sont très bons. Ces résultats sont illustrés par la figure 2b et 2c. Ainsi, un marquage précis du fondamental peut être obtenu même dans le cas de transitions dans la structure temporelle du signal (voir 1145 et 1220 ms dans la figure 2d).

## 3. SYNTHÈSE DE PLUS HAUTE PRÉCISION

Dans le but de réaliser une synthèse de très haute précision, nous pouvons imaginer suréchantillonner le signal pour le marquage et la resynthèse. Nous avons en effet suréchantillonné le signal original pour le marquage du fondamental. Cependant, nous n'avons pas suréchantillonné le signal pour la synthèse car elle aurait nécessité une résolution trop élevée pour faire correspondre idéalement les marques de synthèse, pour n'importe quelle modification temporelle ou fréquentielle (c'est-à-dire pour réussir à ce que les marques de synthèse correspondent à un instant d'échantillonnage).

Voici comment notre algorithme fonctionne :

1. **Marquage du fondamental**
   - Suréchantillonnage et filtrage passe-bas (à la fréquence de 2500 Hz).
   - Marquage des périodes du fondamental sur le signal suréchantillonné.
2. **Resynthèse**
   - Application de l'algorithme TD-PSOLA avec ces marques pour obtenir la position exacte (théorique) des marques de synthèse et les associer avec une fenêtre d'analyse.
   - Recalage du segment par rééchantillonnage pour obtenir la fenêtre de synthèse réelle (voir figure 3).
   - Reconstruction du signal.

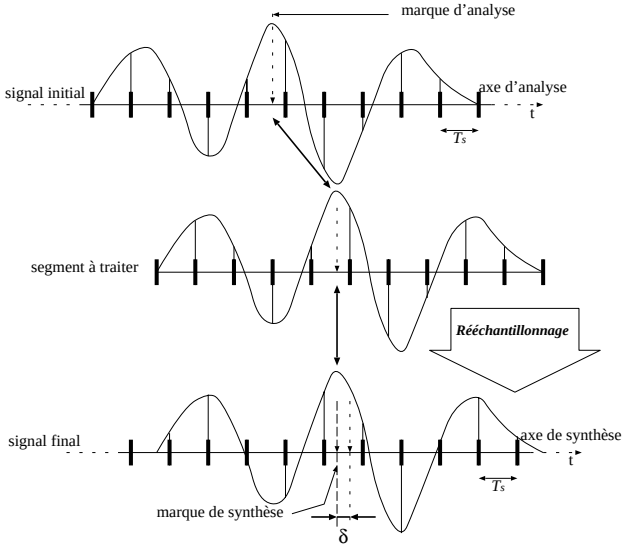


**Figure 3:** Recalage d'un segment d'analyse sur l'axe temporel de synthèse.

À partir de la marque du fondamental et de la marque de synthèse d'une fenêtre donnée, nous utilisons une méthode de rééchantillonnage rapide, décrite au-dessous, pour

recaler la fenêtre précisément à l'endroit où elle doit apparaître dans le nouveau signal.

Soit $x[n]$ la fenêtre originale, le signal rééchantillonné est donné par A. Oppenheim [7] comme suit :

$$x(t) = \sum_{n=-\infty}^{\infty} x[n] sinc\left(\frac{\pi(t - nT_s)}{T_s}\right) \quad (4)$$

où $T_s$ est la période d'échantillonnage.

Le calcul de la fenêtre finale $y[m]$ correspondant à la fenêtre $x[n]$ décalée d'un petit délai $\delta$ s'obtient en évaluant $x(mT_s - \delta)$ (voir figure 3). Ainsi, $y[m] = x(mT_s - \delta)$ i.e:

$$\begin{aligned} y[m] &= \sum_{n=-\infty}^{\infty} x[n] sinc\left(\pi f_s[(mTs - \delta) - nT_s]\right) \\ &= \sum_{n=-\infty}^{\infty} x[n] sinc\left(\pi f_s[(m-n)Ts - \delta]\right) \end{aligned} \quad (5)$$

où $f_s$ est la fréquence d'échantillonnage ($1/T_s$).

Maintenant, en remplaçant $sinc$ par $sin(x)/x$ et en utilisant la formule suivante :

$$sin\left(\pi f_s[(m-n)T_s - \delta]\right) =$$

$$cos\left(\pi f_s \delta\right) sin\left(\pi(m-n)\right) - sin\left(\pi f_s \delta\right) cos\left(\pi(m-n)\right)$$

Et sachant que $cos\pi(m-n) = \pm 1$ et $sin\pi(m-n) = 0$, on obtient alors :

$$y[m] = \sum_{n=-\infty}^{\infty} x[n] \frac{(-1)^{(m-n+1)} sin\left(\pi f_s \delta\right)}{\pi f_s[(m-n)T_s - \delta]} \quad (6)$$

Comme $0 < \delta < T_s$ (resp. $-T_s < \delta < 0$), on définit $\delta = \alpha T_s$, où $0 < \alpha < 1$ (resp. $-1 < \alpha < 0$). Alors l'équation devient :

$$y[m] = \sum_{n=-\infty}^{\infty} (-1)^{(m-n+1)} x[n] \left(\frac{\sin \alpha\pi}{\pi}\right) \frac{1}{(m-n) - \alpha} \quad (7)$$

La sommation ne pouvant être infinie, nous l'avons effectuée sur un court intervalle (1-2 ms $\simeq 50$ échantillons).

Enfin, la fenêtre obtenue est pondérée par une fenêtre de Hanning. Gimenez et Talkin [2] utilisent une fenêtre asymétrique pour réduire les phénomènes de distortion et de réverbération qui sont introduits par le fenêtrage.

La figure 4 montre l'amélioration de la qualité du signal obtenu. Le premier spectrogramme provient d'un signal modifié avec l'algorithme classique de TD-PSOLA. Le second est le spectrogramme du même signal modifié avec la méthode de haute résolution expliquée ci-dessus. En particulier, la structure harmonique apparaît plus clairement (par exemple vers 1100, 1500 et 2500 ms).

## 4. CONCLUSION

Notre algorithme de marquage du fondamental assure une meilleure précision à la fois pour les marques d'analyse et les marques de synthèse. Nous avons en particulier amélioré la robustesse du marquage pour les signaux présentant de très fortes transitions formantiques.

Deuxièmement, la combinaison de notre marquage du fondamental avec une méthode rapide de rééchantillonnage pendant l'étape de synthèse améliore la qualité du signal. Ce gain en précision évite la réduction de la qualité

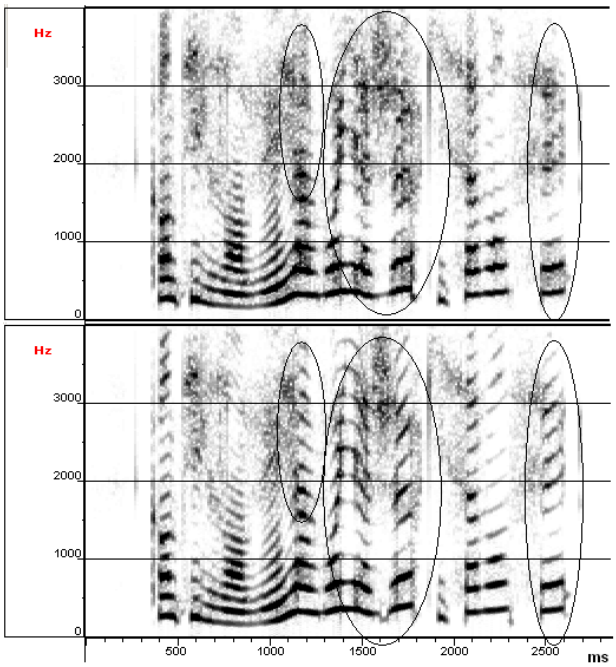


**Figure 4:** Haut : modification de F0 sans rééchantillonnage. Bas : avec rééchantillonnage.

entre le signal original et la signal de synthèse obtenu avec la méthode TD-PSOLA classique. Ce phénomène peut être clairement observé sur la qualité des harmoniques (figure 4) où le niveau de bruit entre les harmoniques est réduit grâce à notre méthode.

Dans des travaux futurs, nous chercherons à exploiter des algorithmes de très haute précision d'extraction de F0 [5] pour améliorer la détermination des marques du fondamental sans suréchantillonnage.

## BIBLIOGRAPHIE


[1] Y. M. Cheng and D. O'Shaughnessy. Automatic and reliable estimation of glottal closure instant and period. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-37(12):1805–1815, December 1989.

[2] F. Gimenez de los Galanes and D. Talkin. High resolution prosody modification for speech synthesis. In *Eurospeech*, pages 557–560, Rhodes, Greece, 1997.

[3] Y. Laprie and V. Colotte. Automatic pitch marking for speech transformations via TD-PSOLA. In *IX European Signal Processing Conference*, Rhodes, Greece, 1998.

[4] Ph. Martin. Comparison of pitch detection by cepstrum and spectral comb analysis. In *Proc. of Int. Conf. Acoust., Speech, Signal Processing 1982*, pages 180–183, 1982.

[5] Y. Medan, E. Yair, and D. Chazan. Super resolution pitch determination of speech signals. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, ASSP-39(1):40–48, January 1991.

[6] E. Moulines and F. Charpentier. Pitch synchronous waveform processing techniques for a text-to-speech synthesis using diphones. *Speech Communication*, 9(5,6):453–467, 1990.

[7] A. V. Oppenheim and R. W. Schafer. *Digital Signal Processing*. Prentice-Hall, Inc, 1975.

[8] R. Veldhuis. Consistent pitch marking. In *International Conference on Speech Language Processing*, Beijin, 2000.