# Synthèse vocale par sélection d'unité: une méthode pour la redéfinition de la courbe intonative

*Baris Bozkurt, Thierry Dutoit and Vincent Pagel*

MULTITEL ASBL
Initialis Scientific Park, Copernic Avenue, B-7000 Mons, Belgium
Mél: bozkurt@multitel.be, pagel@multitel.be - http://www.multitel.be/
Faculté Polytechnique De Mons
Rue de Houdain 9, B-7000 Mons, Belgium
Mél: thierry.dutoit@fpms.ac.be - http://www.fpms.ac.be/

## ABSTRACT

In this work, we propose a new algorithm for defining intonation curves from selected units in a non-uniform units-based text-to-speech synthesis system. Since the main trend in a non-uniform units-based system is to select the best and modify the least to achieve highly natural synthetic speech, the target intonation imposed on units is of great importance. We propose a 'shift-only' algorithm to re-define target intonation from selected units, which does not modify the general prosodic characteristics (micro-prosody, melodic movements) of units, while efficiently reducing F0 discontinuities at concatenation points. For the operation, a cost function is defined as a summation of discontinuities and shifts scaled by durations of the units. Minimizing this function for the shift variable, we optimize minimum shift and minimum discontinuity constraints.

## 1. INTRODUCTION

Les récents progrès dans le domaine de la sélection d'unités ainsi que dans les techniques de construction de corpus audio pour la sélection ont montré que l'on peut produire une excellente parole de synthèse selon la technique "select the best, modify the least" [Bal99, Coo00, Fuj98, Möb00, Beu99, Hun96]. Un facteur important, pour le naturel de la parole de synthèse obtenue, est la courbe intonative appliquée aux unités.

Bien que le module de sélection d'unités d'un tel système essaie de garantir la continuité de F0 aux points de concaténation en incluant les valeurs de F0 originales dans le coût de concaténation [Hun96], on obtient généralement des discontinuités si on concatène les segments audio sans retoucher leurs courbes intonatives originales. Le degré de discontinuité dépend fortement du critère de sélection et de la couverture prosodique du corpus. La préparation des corpus pour la sélection devient alors fort complexe quand en plus des séquences de phonèmes, on doit couvrir la réalisation des différents tons de la langue. Imposer une courbe intonative calculée par le module de traitement du langage naturel (TLN) à partir du texte n'est pas une solution pour résoudre notre problème, car cette solution a le désavantage de retoucher les unités originales, leur faisant perdre une partie de leur naturel selon la distance qui peut exister entre la courbe prédite et l'intonation réelle des unités sélectionnées.

Nous proposons ici l'approche suivante pour redéfinir la courbe intonative après la sélection des unités: nous définissons une fonction de coût qui est la somme des discontinuités après l'introduction de décalages verticaux de F0 et une valeur qui pénalise l'opération de décalage proprement dite. Cette deuxième partie de la fonction de coût est obtenue en sommant les décalages de F0 pondérés par la longueur des segments audio. Cette fonction est alors minimisée en tenant compte des variables indiquant le décalage F0 des segments pour obtenir un maximum de continuité avec le minimum de décalage. Avec ce traitement certaines discontinuités disparaissent sans dégradation de la qualité des unités concaténées. La pénalité introduite dans la fonction garantit que les unités sonores les plus courtes seront décalées en priorité (avec l'espoir que leur courte durée empêchera la perception d'éventuels artefacts ). Avec des sujets nous avons montré que lorsque la modification apportée par notre méthode est audible elle est considérée comme une amélioration. L'importance de cette amélioration de qualité dépend cependant de la courbe de F0 effective des unités sonores sélectionnées.

Dans des tests d'écoute informels que nous avons réalisés précédemment sur de courtes phrases de synthèse, nous avons remarqué que le naturel de la synthèse dépend fortement de l'algorithme utilisé pour modifier la prosodie et lisser le point de concaténation. Pour cette raison nous avons choisi pour cette expérience d'utiliser l'algorithme TD-PSOLA [Mou90], qui est populaire à juste titre pour la modification de durée et de fréquence fondamentale, mais qui n'effectue pas de lissage spectral au point de concaténation. En conséquence nos tests ne portent que sur les discontinuités de fréquence fondamentale.

## 2. MOTIVATIONS

Dans un système de synthèse par sélection d'unités, des segments audio de tailles variables sont sélectionnés dans un grand corpus de parole puis concaténés pour

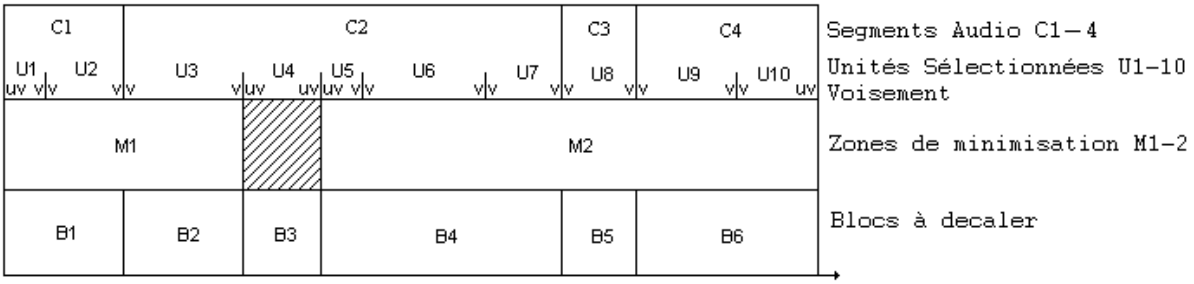


**Figure 1:** Séquence d'unité sélectionnée montrant les segments audio, les zones de minimisation et des blocs décalés

synthétiser un signal de parole extrêmement naturel. Les algorithmes de concaténation sont conçus pour modifier les segments sélectionnés et les concaténer de façon que les discontinuités (énergie, F0, formants, qualité de la source...) au point de concaténation soient réduites avec le moins d'artefacts possibles pour ne pas dégrader le naturel des segments de départ. Un des points importants de cette opération est la redéfinition de la courbe intonative pour réduire les discontinuités de F0 tout en évitant de distordre les segments originaux. Une des dimensions principales qui influe sur la dégradation audio est la distance entre la courbe de F0 originale et la courbe cible. L'algorithme TD-PSOLA que nous avons retenu pour cette expérience possède cette propriété intéressante que si le mouvement est nul, la dégradation de qualité est nulle, ce qui n'est pas le cas de l'algorithme MBROLA [Dut96] issu de notre laboratoire, qui introduit une dégradation constante quel que soit la modification de F0. De plus l'introduction de décalages de F0 sur des blocs de parole permet de conserver la structure micro-mélodique.

Dans notre approche basée sur des décalages par bloc, il est avantageux d'appliquer le lissage sur des ensembles de segment courts plutôt que sur la phrase complète afin de diminuer la charge de calcul et d'éviter la propagation de mouvements intonatifs qui provoqueraient des décalages en dehors des bornes de F0 admissibles. La première étape de notre algorithme consiste donc à déterminer les limites des zones à lisser dans les séquences de segments audio. Nous supposons que les discontinuités de F0 à gauche et à droite des segments non voisés sont de moindre importance, puisque masquées par la durée du segment non voisé. En conséquence, nous effectuons notre minimisation de façon indépendante sur des ensembles de segments audio délimités par des phonèmes non voisés. Nous appelons par la suite ces intervalles de calcul les "zones de minimisation" (Fig1).

Les blocs décalés sont formés par regroupement d'unités consécutives dans chacune des zones de minimisation. Ensuite pour chacune de ces zones, les décalages de F0 à appliquer sont calculés en minimisant la fonction de coût comme expliqué ci-après.

Dans notre système, la classification voisé/non-voisé des trames qui se trouvent aux frontières d'unités se base sur le ratio d'énergie harmonique/non-harmonique calculé pendant la minimisation par la méthode des moindres carrés des paramètres du modèle MBE [Dut96]. Pendant la synthèse aucune modification n'est appliquée aux segments non voisés, donc si un décalage est calculé pour un segment non voisé (parce que ses frontières droite et gauche contiennent des trames voisées), il est tout simplement ignoré.

## 3. CALCUL DES DÉCALAGES DE F0

La fonction *H* est définie comme la combinaison d'une mesure de discontinuité et d'une pénalité pour le décalage de segments audio dans une séquence de *K* unités;

$$H = \sum_{n=1}^{K-1} abs((T0_n^f + s_n) - (T0_{n+1}^i + s_{n+1})) + k * \sum_{n=1}^{K} abs(s_n * d_n)$$

$$d_n = \frac{duration_n}{\sum_{i=1}^{n} duration_i}$$

où $T0_n^f$ est la fréquence fondamentale d'estimée pour la dernière période de la $n^{\text{ème}}$ unité, et $T0_n^i$ est la fréquence fondamentale de la première période. Ces valeurs sont obtenues après le passage d'une filtre médian sur les valeurs de F0 obtenues aux frontières des unités. $s_n$ est le décalage à appliquer à la $n^{\text{ème}}$ unité, *k* est un facteur de pondération qui détermine l'importance de la pénalité de décalage, et $d_n$ est le ratio de la durée du segment et de la durée de la zone de minimisation. Minimiser cette fonction équivaut à minimiser simultanément chaque somme, ce qui amène le système linéaire suivant:

$$(T0_n^f + s_n) - (T0_{n+1}^i + s_{n+1}) = 0$$
$$n = 1,2,3....K-1$$

$$k * s_n * d_n = 0$$
$$n = 1,2,3...K$$

Il en découle la représentation matricielle suivante:

$$\begin{bmatrix} 1 & -1 & 0 & \dots & \dots & 0 \\ 0 & 1 & -1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 0 & 1 & -1 \\ k*d_1 & 0 & \dots & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & 0 & k*d_K \end{bmatrix} * \begin{bmatrix} s_1 \\ s_2 \\ \dots \\ \dots \\ \dots \\ s_K \end{bmatrix} = \begin{bmatrix} -(T0_1^f - T0_2^i) \\ \dots \\ \dots \\ -(T0_{K-1}^f - T0_K^i) \\ 0 \\ \dots \\ 0 \end{bmatrix}$$

$$A_{(2K-1)*K} * \vec{s}_{K*1} = \vec{d}_{(2K-1)*1}$$

Pour $K$>1, le nombre d'équations est supérieur au nombre d'inconnues et la méthode des moindres carrés convient (pour $K$=1, le décalage vaut tout simplement 0) :

$$\vec{s} = (A^T * A)^{-1}(A^T * \vec{d})$$

Le facteur de pondération $k$, qui sert à pénaliser les décalages, est fixé manuellement par essais successifs. Comme expliqué plus loin, on peut le régler en fonction du post-traitement appliqué après l'opération de décalage de F0.

Nous présentons dans la figure ci-dessous un exemple des courbes de T0 (période fondamentale) des unités avant et après l'opération de décalage (s1 et s6 sont des segments audio non voisés).

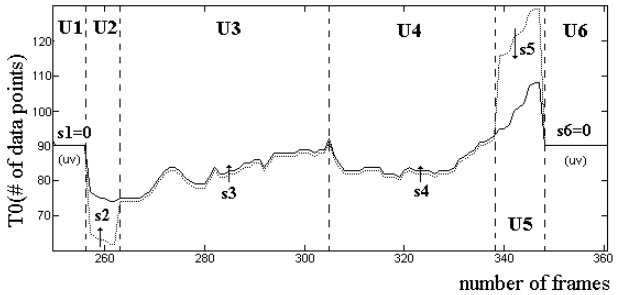


**Figure 2:** Courbes de T0 avant et après l'opération de décalage de fréquence fondamentale (courbe originale en pointillé). U1..U6 sont les unités.

Les décalages calculés dépendent des valeurs spécifiques de T0 observées sur les unités sélectionnées (et donc du contenu du corpus et de l'algorithme de sélection). Toutes les discontinuités ne peuvent pas être éliminées avec la méthode proposée, comme le montre la figure suivante avec de fortes discontinuités de T0 entre les unités. Malgré la réduction globale des discontinuités, certaines sont encore audibles dans la parole synthétisée.

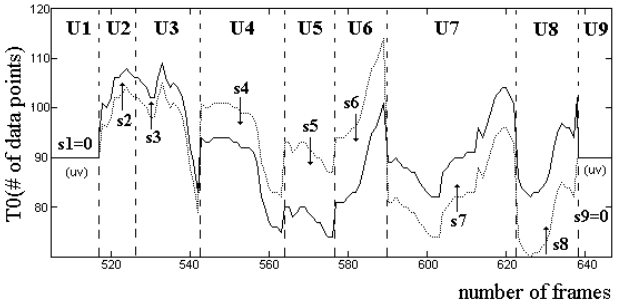


**Figure 3:** Courbe de F0 d'une zone de minimisation présentant de fortes discontinuités, dont certaines sont réduites par le décalage (courbe originale en pointillés).

Notre groupe de travail examinera l'intégration de cette méthode dans le coût de sélection proprement dit du système de sélection d'unité. En effet la connaissance des spécificités de l'algorithme de modification prosodique améliore le bénéfice de notre méthode de minimisation: les séquences d'unités pour lesquelles la méthode fournit un décalage de F0 satisfaisant seront préférées aux séquences pour lesquelles des discontinuités persistent.

## 4. POST-TRAITEMENT

L'algorithme proposé fait disparaître de nombreuses discontinuités de F0, cependant certains problèmes peuvent apparaître en fonction des unités sélectionnées, et introduire des mouvements intonatifs inattendus. Par exemple si trois unités consécutives présentent une intonation montante et qu'elles sont décalées pour assurer la continuité de F0, on obtient une longue courbe intonative montante. Pour éviter ce problème, on peut dans un premier temps augmenter le facteur de pondération $k$ de façon à pénaliser les décalages. Ce n'est cependant pas une solution universelle car tous les décalages, grands et petits, sont pénalisés. Pour cette raison nous avons limité le coefficient $k$ à 0.6 et appliqué un filtre aux décalages pour les conserver dans un intervalle acceptable (perceptible).

L'étape suivante consiste à appliquer un lissage aux frontières qui présentent toujours une discontinuité de F0 importante après l'opération de décalage. Parmi d'autres fonctions qui convenaient aussi, nous avons retenu un lissage linéaire de part et d'autre du point de concaténation (l'étendue du lissage correspond au tiers des unités impliquées). Une analogie très forte existant entre le lissage de F0 et le lissage spectral, d'autres méthodes utilisées normalement pour le lissage de coefficients spectraux pourraient donc s'appliquer (par exemple la méthode de [Wou01] qui permet de façon élégante de contrôler la dynamique des mouvements formantiques ).

## 5. TESTS PERCEPTIFS

Nous avons effectué des tests perceptifs préliminaires sur de courtes phrases synthétiques dont l'intonation est produite selon deux méthodes:

1-simple concaténation des unités sélectionnées,

2-simple concaténation des unités après décalage de F0 selon la méthode proposée.

L'algorithme de sélection utilisé dans cette expérience ne tient compte que de paramètres acoustiques dans son coût de sélection: contexte phonétique, durée/durée prédite, F0/F0 prédite par le TLN. Le test utilise une base de parole féminine de 60mn en français. Nous avons retenu pour le test les phrases pour lesquelles un décalage était effectivement appliqué, et nous n'avons pas mis en oeuvre le post-traitement afin de pouvoir juger l'algorithme de décalage. Notre lissage spectral était désactivé et une implémentation de TD-PSOLA était chargée des modifications de F0. Le marquage des périodes pour TD-PSOLA a été effectué en calculant les instants ou la phase de la première harmonique vaut 0 [Gri87], ceci en supposant que les portions du spectre les plus énergétiques sont proches des premières harmoniques, et que la phase du premier harmonique pour fournir

l'information nécessaire à la synchronisation de l'opération d'OLA [Sty98].

Vingt sujets, non familiers avec la parole de synthèse, ont choisi, au moyen d'une interface web, entre la version synthétisée par la méthode 1 et par la méthode 2, pour 14 phrases synthétiques courtes (test de préférence AB). Ils avaient la possibilité de réécouter chacune des 2 productions autant que nécessaire avant de choisir l'une d'elle, leur instruction étant de choisir la plus naturelle. Au total, la préférence pour la version utilisant la méthode 2 est de 75%. La plupart des sujets estiment que dans certains cas la différence de qualité est flagrante et que dans d'autre cas les 2 productions leurs paraissent équivalentes.

## 6. DISCUSSION

Nous proposons ici un nouvel algorithme pour redéfinir la courbe intonative des unités choisies par un système de synthèse par sélection. Les tests perceptifs ont montré que certaines des discontinuités de F0 aux points de concaténation pouvaient être éliminées sans altérer le naturel des segments de parole originaux. Bien sûr l'amélioration obtenue dépend fortement des caractéristiques intonatives des unités sélectionnées. Nous nous efforcerons dans nos prochains travaux d'intégrer cette méthode en aval dans la fonction de coût de sélection.

## REFERENCES


[Bal99] Balestri M., Paechiotti, A., Quazza, S., Salza, P. L., SandriS. (1999), "Choose the best to modify the least: a new generation concatenative synthesis system", Proc. of EUROSPEECH, Budapest, Hungary.

[Coo00] Coorman, G., Fackrell, J., Rutten, P., Van Coile, B. (2000),"Segment selection in the L&H Realspeak laboratory TTS system", Proc. of ICSLP, Beijing, China.

[Fuj98] Fujisawa, K., and Campbell, N. (1998), "Prosody based unit-selection for Japanese speech synthesis", Proc. of 3rd ESCA/COCOSDA Workshop on Speech Synthesis, Jenolan Caves, NSW, Australia.

[Möb00] Möbius, B. (2000),"Corpus-based speech synthesis: methods and challenges" Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart), AIMS 6 (4), 87-116.

[Beu99] Beutnagel, M., Conkie, A. and Schroeter, J., Stylianou, Y., and Syrdal, A. (1999),"The AT&T NextGen TTS system", Proc. of the Joint Meeting of ASA, EAA and DAGA , Berlin, Germany.

[Hun96] Hunt, A. and Black, A., (1996), "Unit selection in a concatenative speech synthesis system using a large speech database", Proc. of ICASSP, Atlanta, Georgia, p 373-376.

[Mou90] Moulines, E. and Charpentier F. (1990) "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", Speech Commun., Vol.9, p 453-497.

[Dut96] Dutoit,T. and Pagel, V., (1996), "Le projet MBROLA : vers un ensemble de synthetiseurs vocaux disponibles gratuitement pour utilisation non-commerciale", Actes des JEP, Avignon, pp.441-444.

[Wou01] Wouters, J. and Macon M.W. (2001), "Control of spectral dynamics in concatenative speech synthesis", IEEE Trans. on Speech and Audio Proc., Vol.9(1), p 30-38.

[Sty98] Stylianou, Y. (1998), "Removing phase mismatches in concatenative speech synthesis", Proc. 3rd ESCA Speech Synthesis Workshop, p 267-272.

[Sai01] Saito, T. and Sakamoto M. (2001), "Generating F0 contours by statistical manipulation of natural F0 shapes", Proc.of Eurospeech Scandinavia, p 1171-1174.

[Mon92] Monaghan, A.I.C. (1992), "Extracting microprosodic information from diphones, a simple way to model segmental effects on prosody for synthetic speech", Proc.of ICSLP, Banff,Canada, p 1159-1162.

[Kor68] Korn, A.G. and Korn T.M. (1968), Mathematical Handbook for Scientists and Engineers, McGraw-Hill.

[Cam97] Campbell, W.N. and A.W. Black, Prosody and the selection of source units for concatenative synthesis. In Jan van Santen, Richard W. Sproat, Joseph P.Olive, and Julia Hirschberg, editors, Progress in Speech Synthesis. Springer, New York, 1997, p 279-292.

[Gri87] Griffin, D.W. Multi-band excitation vocoder, PhD Dissertation, MIT, 1987.