Ralentisseur du signal de parole par autocorrélation

Philippe Martin

Département d'Études Françaises, Université de Toronto
Carr Hall, St Joseph St.
Toronto, Ontario, Canada M5S 1J4
Tél.: ++1 416 960 6122 - Fax: ++1 416 920 4634
Mél: philippe.martin@utoronto.ca
http://www.chass.utoronto.ca/french/ling/Homepage/martin.html

ABSTRACT

Speech rate changes - slow down or acceleration – have known for a long time important applications in language teaching, linguistic corpus transcription, office dictation, etc. Very good quality modifications are obtained by the phase vocoder, but at the expense of a somewhat high computing cost. Temporal methods such as PSOLA are more efficient, but highly dependent on a good pitch tracking algorithm.

A new approach is presented here, similar to PSOLA as working directly on the waveform, but relying to autocorrelation to align consecutive speech segments in the overlapping adding process. It is therefore simpler to implement and more reliable as bypassing the period marking process used in PSOLA.

1. INTRODUCTION

Le ralentissement de la parole naturelle constitue depuis de nombreuses années un outil indispensable pour de nombreuses applications telles que la transcription de textes oraux, la bureautique, l'enseignement des langues étrangères, etc. L'accélération de la parole a également connu des applications intéressantes notamment pour les malvoyants. Les premières réalisations déjà vieilles de 30 ou 40 ans sont mécaniques. Elles font appel à la répétition de segments de signal de parole de durée constante par un système de tête magnétique rotative installée sur un magnétophone à bande. On reconnaît là le principe de procédés plus récents et plus élaborés opérant directement sur le signal, procédés tels que PSOLA ou le vocodeur de phase. Des segments prélevés sur le signal original sont dupliqués à intervalles réguliers. La mauvaise qualité du système est évidemment due aux transitions brutales de pitch, de phase et d'intensité aux endroits de raccordement des segments.

1.1 Principe

Le principe de la modification du paramètre temporel du signal de parole est simple : il s'agit de dupliquer – ou de soustraire dans le cas de l'accélération – un segment relativement court du signal, d'une durée pour laquelle l'hypothèse de stationnarité de la configuration articulatoire reste valide, soit en pratique de l'ordre de 20 à 40 ms. Cette duplication – soustraction pose peu de problèmes de raccordement des segments pour les sons non voisés (pour lesquels l'hypothèse de stationnarité est suffisante). Il suffit dans ce cas « d'adoucir » la transition par une fenêtre temporelle appropriée. Par contre pour les sons voisés le raccordement aux frontières de segment implique un alignement de phase, ainsi que la réalisation d'une période de pitch inchangé par rapport a la période précédente. Toute l'astuce va donc consister à trouver un procédé optimal pour positionner chaque nouveau segment sur l'axe temporel par rapport au signal existant de manière à minimiser l'effet perceptif du raccord à l'endroit de la concaténation, effet portant sur trois discontinuités possibles : le pitch, la phase et l'enveloppe spectrale.

1.2 PSOLA

L'algorithme PSOLA [Mou90], qui opère en positionnant les segments autour d'une référence liée à la période de pitch du signal, tend à minimiser l'effet du raccord d'une part en utilisant de segments de durée sensiblement égale à la période, et de l'autre en utilisant une fenêtre temporelle également centrée sur la référence de la période de pitch. En opérant sur des variations de période – correspondant à l'écart entre deux segments successifs – relativement modestes (de l'ordre de 10 à 30%), on peut espérer une faible variation de phase des diverses composantes entre la fin du signal et le début du nouveau

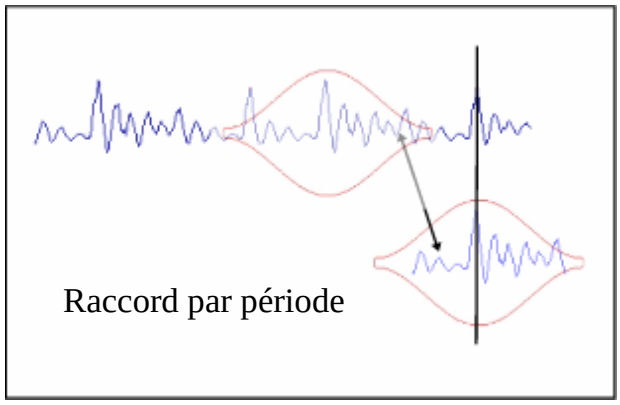

Figure1: Effet d'écho dû à une discontinuité de phase lors d'une augmentation importante du pitch (diminution de la période) dans la méthode PSOLA.

segment, celui-ci ayant été prélevé aux environs du signal original. Les discontinuités de pitch sont donc contrôlées par le principe même de la méthode, et les effets de discontinuités de phase, ne deviendront gênants que pour des variations de pitch importantes, provoquant un effet d'écho caractéristique.
D'une manière générale, ce procédé est limité par la fiabilité du marqueur de Pitch [Mar00], en particulier pour des signaux bruités ou multi sources, et ce malgré diverses améliorations possibles pour améliorer la synchronisation en phase des composantes spectrales [Pee02].

1.3 Vocodeur de phase

Le vocodeur de phase s'adresse directement au problème du raccord de phase, puisque chaque composante du nouveau segment concaténé est replacée en phase de manière à s'aligner avec le segment précédent. L'inconvénient de cette méthode réside dans le coût du calcul, et surtout dans les distorsions inhérentes à l'emploi d'une fenêtre d'analyse pour la FFT et à la non stationnarité du signal. Une autre difficulté provient de l'ambiguïté possible sur le calcul des phases des différentes composantes, chaque phase n'étant connue qu'à un facteur 2π près [Po76], [Por81] et [Mar01].

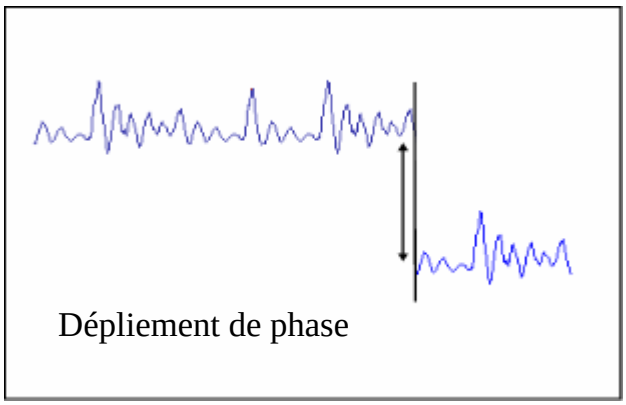

Figure 2: Vocodeur de phase.

Une étude comparative détaillée de différents procédés de ralentissement peut être trouvée dans [Ver00].

2. LE RALENTISSEMENT PAR AUTOCORRÉLATION

2.1 Principe

Le principe du ralentissement du signal de parole par autocorrélation est semblable à celui de la méthode par addition de segments alignés sur des marqueurs de pitch PSOLA : insertion d'un segment du signal prélevé à proximité immédiate du point d'insertion. Au lieu de contrôler la position du nouveau segment par un marqueur de pitch, c'est le maximum de la fonction d'autocorrélation sur cet intervalle qui est choisi comme point d'ancrage. L'intervalle de calcul de la fonction d'autocorrélation est constant (une valeur typique est de l'ordre de 16 ms), et les segments dans cet intervalle du signal original et du nouveau segment sont additionnés après pondération par une fenêtre temporelle triangulaire (Voir fig. 4).

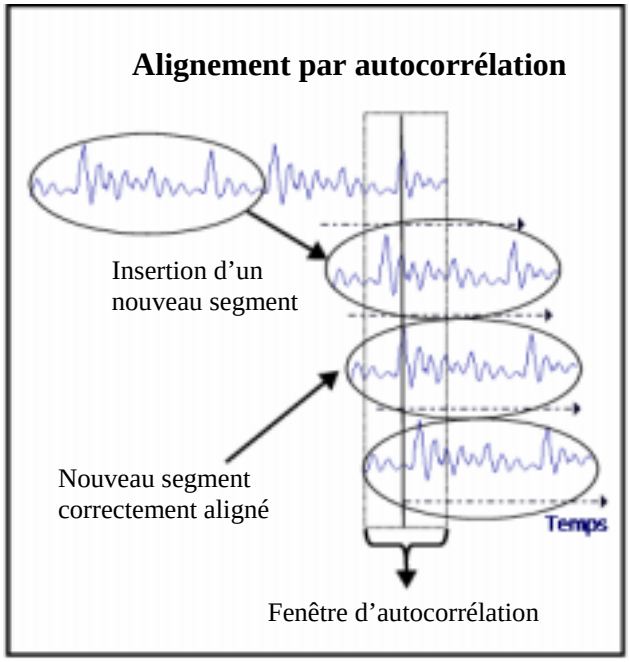

Figure 3: Principe du ralentisseur par autocorrélation. Un segment prélevé dans le passé immédiat du signal est aligné par autocorrélation dans la durée de recouvrement et inséré à la position optimale, correspondant à un pitch sensiblement constant

Ce fenêtrage théoriquement optimal assure la transition entre les parties finale et initiale des segments concaténés, et implique des erreurs de raccordement de phase sensiblement égales à celles obtenues par PSOLA, puisque la durée de recouvrement est, dans le cas du ralentissement sans changement de la fréquence fondamentale, égale à la moitié de la période.

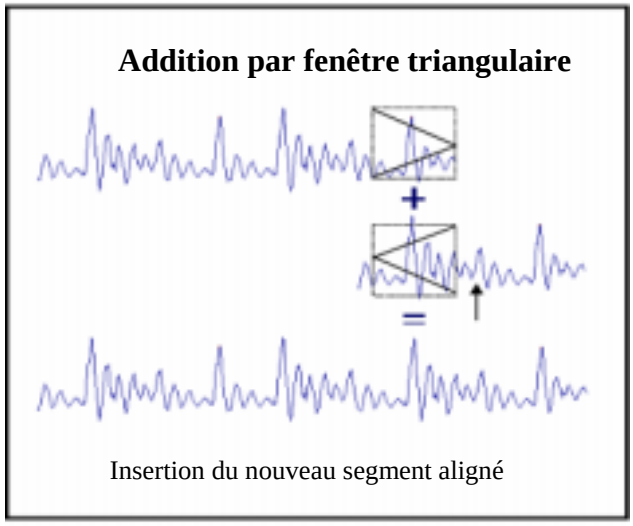

Figure 4: Addition par fenêtrage triangulaire dans la zone de recouvrement.

2.2 Coût computationnel

Le coût de calcul de la méthode de type PSOLA dépend de l'algorithme choisi pour la détection de la fondamentale et le marquage des périodes de pitch. Si une méthode spectrale est utilisée (par exemple le peigne spectral, [Mar00]) et donc implique une transformée de Fourier, on peut retenir comme unité de calcul la FFT d'une trame de signal, puisque le vocodeur de phase exige (au moins) une FFT et une FFT inverse, et que le coût du calcul de l'autocorrélation sur une demi période fondamentale peut être estimé à une demi FFT Une évaluation sommaire des coûts de calcul comparés les trois méthodes est résumée dans la table 1.

Table 1: Comparaison des coûts de calcul pour le ralentissement par PSOLA, par vocodeur de phase et par autocorrélation, l'unité de coût de calcul est une FFT.

	PSOLA	**PHASE**	**AUTO**
Nombre de FFT	1	2	1/1

Un paramétrage typique du calcul de la fonction d'autocorrélation pour l'évaluation du pitch implique une fenêtre de 256 points constante (voir fig. 3), soit une durée de 23 ms pour une fréquence d'échantillonnage de 11025 Hz. Dans l'hypothèse généralement admise d'une variation maximale de pitch de +/- s1% par ms, l'excursion maximale de période à considérer est de +/- 23 %, ce qui correspond à un déplacement de +/- 58 points. Le nombre total de multiplications – additions est donc de l'ordre de 30.000.

En fait l'alignement des segments insérés demande un balayage selon l'axe temporel de l'ordre d'une période fondamentale. Pour couvrir les cas des périodes les plus longues (soit 15 ms pour une fréquence fondamentale de 70 Hz), une excursion de 167 points est nécessaire, alors que dans les cas de Fo élevé (500 Hz), un balayage sur 2 ms ne demande que 21 points, et donc 21 évaluations de la fonction d'autocorrélation. L'utilisation d'une fenêtre à durée constante implique donc un alignement optimal au sens de l'autocorrélation prenant en compte plus d'une période de pitch.

2.3 Avantages

Le ralentissement par autocorrélation apparaît comme un cas particulier du raccordement par vocodeur de phase, dans lequel une optimisation globale est effectuée dans le domaine temporel sur l'ensemble des composantes spectrales. La méthode peut aussi être considérée comme de type PSOLA, le marquage du pitch assurant l'alignement des segments concaténés étant remplacé par l'autocorrélation.

3. CONCLUSIONS

Le ralentissement de la parole naturelle constitue un cas particulier de la modification des paramètres prosodiques. Les méthodes connues telles que PSOLA ou le vocodeur de phase demandent soir un marquage des périodes de pitch fiable (difficile à réaliser sans erreur sans correction manuelle), soit le dépliement des raccordements de phase des composantes spectrales analysées par une transformée de Fourier qui peut lui aussi s'avérer erroné.

La nouvelle méthode proposée, ASOLA, s'inspire du principe par recouvrement additif de PSOLA, mais aligne les segments concaténés par un calcul d'autocorrélation effectué sur une durée constante de recouvrement des segments. Une fois les segments correctement positionnés dans le temps, les parties communes de la fenêtre d'autocorrélation sont additionnées après pondération par une fenêtre triangulaire. Cette méthode est donc plus robuste, plus simple et plus rapide.

BIBLIOGRAPHIE

[Lar93] Laroche, J. (1993) “Autocorrelation Method for high quality pitch/time scaling”, *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics IEEE*, 200-204.

[Mou90] Moulines E. and Charpentier, F. (1990), “Pitch Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis”, *Speech Communication*, (9), no 5-6, 453-467.

[Mar00] Martin, Ph. (2000) "Peigne et brosse pour Fo : Mesure de la fréquence fondamentale par alignement de spectres séquentiels", *Actes des 23èmes JEP*, Aussois, France, juin 2000, 245-248.

[Mar01] Di Martino, J. and Laprie, Y. (2001) “Suppression of Phasiness for Time-Scale Modifications of Speech Signals Based on a Shape Invariance Property”, in *ICASSP, Salt Lake City, USA*. 2001.

[Por76] Portnoff, M.R. (1976) "Implementation of the Digital Phase Vocoder Using the Fast Fourier Transform" *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-24, no 3, June 1976.

[Por81] Portnoff, M.R. (1981) “Time-scale modification of speech based on short-time fourier analysis”, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-29(3), 374-390, June 1981.

[Pee02] Peeter, G. (2002), “Analyse et synthèse Sinola”, http://www.ircam.fr/anasyn/peeters/

[Ver00] Verhelst, W. (2000) “Overlap-add methods for time-scaling of speech”, *Speech Communication*, (30), 207-221.