# Analyse et synthèse des sons musicaux par la méthode PSOLA

Geoffroy Peeters
IRCAM Équipe analyse/synthèse 1, pl. Igor Stravinsky, 75004 Paris, France
Geoffroy.Peeters@ircam.fr

## Résumé

Dans cet article nous montrons l'application de la méthode PSOLA pour l'analyse et la synthèse de la voix et des sons musicaux. Nous exposons en particulier une procédure de marquage automatique s'adaptant aux caractéristiques du signal, ainsi que différentes méthodes visant à améliorer la synthèse par Superposition/Addition dans les zones voisées du signal, méthodes appelées TDI-PSOLA et FDI-PSOLA, et dans les zones non-voisées. Enfin, nous comparons ces méthodes.

## Introduction

La méthode d'analyse et de synthèse par Superposition/Addition de fenêtres synchrones à la période fondamentale du signal est désignée sous le nom de PSOLA (Pitch Synchronous OverLap-Add) [CS86]. Cette méthode offre de nombreuses possibilités de transformation du son, tout en préservant des caractéristiques fines difficiles à modéliser, telles la forme d'onde et donc certaines caractéristiques de phase du signal. Cette méthode est l'une des plus utilisées à l'heure actuelle pour la synthèse de la parole.

Nous décrivons dans un premier temps les techniques d'analyse et de synthèse. Les raffinements apportés sont ensuite exposés et comparés. Nous terminons par l'application de la méthode aux sons musicaux.

## 1 Méthode PSOLA

La méthode PSOLA repose sur le découpage d'un signal $s(t)$ en fenêtres successives $s_i(t)$ positionnées en fonction des périodes fondamentales du signal.

Ces fenêtres successives sont obtenues par placement de *marques de lecture* $t_r^i$ de manière synchrone au pitch du signal. Le signal est alors découpé à l'aide de fenêtres d'analyse centrées sur ces marques de lecture.

$$s_i(t) = s(t) \cdot h_i(t - t_r^i) \qquad (1)$$

Les fenêtres sont du type Hanning de longueur égale à deux fois la période locale (voir [MC90] pour une discussion du choix, du type et de la taille des fenêtres).

Le signal de synthèse $\tilde{s}(t)$ est alors obtenu par Superposition/Addition des signaux élémentaires centrés en de nouvelles positions $t_w^j$ que nous appelons *marques d'écriture* (fig. 1). Ce sont ces positions qui déterminent le pitch et la durée du signal de synthèse.

$$\begin{cases} \tilde{s}(t) = \sum_j \tilde{s}_j(t) \\ \tilde{s}_j(t) = s_i(t_r^i - t_w^j) \end{cases} \qquad (2)$$

Les signaux élémentaires peuvent au préalable être traités dans le domaine temporel ou dans le domaine fréquentiel. Ceci distingue les méthodes TD-PSOLA, TDI-PSOLA et FDI-PSOLA.

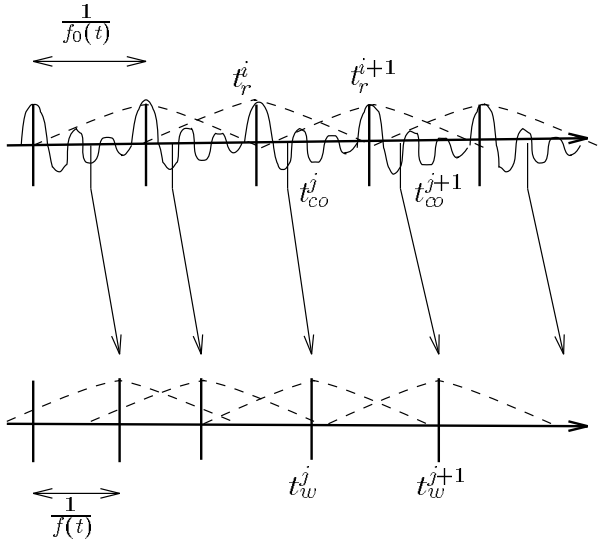


Figure 1: Placement des marques de lecture $t_r^i$, d'écriture $t_w^j$ et des temps de correspondance $t_{co}^j$

# 2 Analyse du signal

Différentes procédures de placement des $t_r^i$ sont utilisées selon les caractéristiques locales des composantes du signal : périodiques ou aléatoires, stationnaires ou transitoires. Une segmentation préalable du signal en zones de caractéristiques identiques permet d'orienter le marquage vers la méthode appropriée. En outre les résultats de cette segmentation seront nécessaires pour l'étape synthèse.

## 2.1 Caractérisation du signal

Quelque soit la méthode de synthèse utilisée il est indispensable de séparer les composantes périodiques du signal (dites voisées dans le cas de la voix), des composantes aléatoires (dites non-voisées).

Pour cela, nous calculons un coefficient de voisement $v(t, \Omega)$ [GL85] [RDP87] mesurant la périodicité du signal autour d'un temps donné et dans une bande de fréquence.

Il permet la séparation des zones *voisées* du signal des zones *non-voisées*.

Une deuxième segmentation est également nécessaire afin d'isoler les transitoires d'attaque (les plosives dans le cas de la voix).

## 2.2 Placement des marques de lecture

À partir des informations issues de la caractérisation du signal, nous orientons le marquage vers différentes méthodes.

La précision du placement des marques d'écriture détermine en grande partie la qualité du signal de synthèse obtenu [Kor97].

### 2.2.1 Dans les zones voisées

Les marques doivent y être placées non seulement de manière synchrone au pitch du signal, mais également de manière à ce que le fenêtrage préserve au maximum les caractéristiques temporelles du signal. Elles doivent donc aussi se trouver près des maxima locaux d'énergie. Nous utilisons l'algorithme suivant pour le marquage :

1. Estimation de la fréquence fondamentale du signal, $f_0(t)$, par la méthode du maximum de vraisemblance [DR91],
2. Estimation de l'énergie du signal $e(t)$,
3. Filtrage de l'énergie du signal par un filtre à temps variable passe-bande centré autour de la fréquence fondamentale : $e_{f_0}(t)$,
4. Détermination des maxima locaux de l'énergie filtrée $\max_{t \in T_0} e_{f_0}(t)$
5. Algorithme itératif de placement des marques par convergence vers les périodes vraies.

Le filtrage passe-bande permet d'éviter la détection de maxima dus à des harmoniques ou sous-harmoniques dont l'énergie serait plus importante que celle du fondamental. L'algorithme itératif vise à optimiser deux contraintes :

- un éloignement minimum des marques par rapport aux maxima d'énergie locaux
- un espacement entre deux marques successives égal à la période locale vraie

La position des marques est optimisée sur l'ensemble des marques appartenant à une région voisée.

### 2.2.2 Dans les zones non-voisées

Le marquage se fait dans ces zones de manière asynchrone. Les marques sont placées équi-distantes. La distance est égale à la moyenne des périodes contenues dans les zones voisées, soit $\overline{T_0(t)}$, ceci afin d'éviter une détérioration des transitions entre zones de caractéristiques différentes.

### 2.2.3 Pour les transitoires d'attaques

Les marques sont placées directement sur les maxima locaux du signal afin de préserver les caractéristiques temporelles de l'attaque.

# 3 Synthèse du signal

La synthèse en PSOLA permet d'effectuer très simplement des modifications du son, modifications de prosodies (dans le cas de la voix) ou d'une manière plus générale des transpositions ou des modifications de durée.

La synthèse est effectuée par Superposition/Addition des signaux élémentaires $\tilde{s}_j(t)$ (obtenus à partir des $s_i(t)$) placés en de nouvelles positions $t_w^j$. Ces positions sont déterminées par la hauteur et la durée du signal de synthèse voulu.

Nous appelons $f(t)$ la nouvelle fréquence fondamentale et $d(t)$ le coefficient de modification des durées.

Dans cette partie, nous allons comparer différentes méthodes dérivées de PSOLA permettant d'améliorer la qualité du signal de synthèse obtenu, méthodes appelées TD-PSOLA, TDI-PSOLA et FDI-PSOLA.

## 3.1 Placement des marques d'écriture

Ces marques déterminent les positions auxquelles seront recopiés les signaux élémentaires. La

manière dont ces marques sont placées est commune aux trois méthodes TD-PSOLA, TDI-PSOLA et FDI-PSOLA, mais diffère selon les caractéristiques locales du signal.

#### 3.1.1 Cas d'un signal voisé :

Nous désirons obtenir un signal de synthèse de fréquence $f(t)$. Pour ce faire nous plaçons les signaux élementaires (représentant chacun une période fondamentale) distants de $\frac{1}{f(t)}$. Ceci détermine alors le placement des marques d'écriture $t_w^j$.

Supposons connu $t_w^{j-1}$, on en déduit $t_w^j$ de la manière suivante :

$$\forall j, \quad t_w^j = t_w^{j-1} + \frac{1}{t_w^j - t_w^{j-1}} \int_{t_w^{j-1}}^{t_w^j} \frac{1}{f(t)}\, dt \qquad (3)$$

Afin de connaître la correspondance entre marques de lecture et d'écriture nous introduisons le temps dit de "correspondance" $t_{co}^j$. $t_{co}^j$ est la position correspondant à $t_w^j$ sur le signal original (fig. 1). Il dépend non seulement de $t_w^j$, mais également du facteur de dilatation $d(t)$.

Supposons connus $t_{co}^{j-1}$, $t_w^{j-1}$ et $t_w^j$, supposons de plus $d(t)$ constant sur la largeur d'une période, on en déduit $t_{co}^j$ de la manière suivante :

$$\forall j, \quad t_{co}^j = t_{co}^{j-1} + \frac{t_w^j - t_w^{j-1}}{d(t)} \qquad (4)$$

L'association entre signaux élémentaires et marques d'écriture est discutée plus loin.

#### 3.1.2 Cas d'un signal non-voisé :

Dans le cas d'une zone non-voisée, les marques d'écriture et les temps de correspondance sont placés de manière équidistante. Leur position s'obtient à l'aide de (eq. 3) et (eq. 4) avec $f(t) = f_0(t)$.

#### 3.1.3 Cas des transitoires d'attaque :

Les transitoires d'attaque (non-voisés) ne doivent subir aucune modification, leur dilatation étant perceptivement désagréable. L'ensemble des $s_i(t)$ correspondant à un transitoire d'attaque est donc recopié sans modification de leurs espacements temporels.

### 3.2 Détermination des signaux élémentaires utilisés

Il nous reste à déterminer quels signaux élémentaires seront recopiés aux positions des marques d'écriture. Ce choix dépend des caractéristiques locales du signal.

#### 3.2.1 Dans les zones voisées :

Les signaux $\tilde{s}_j(t)$ utilisés dans (eq. 2) sont déterminés à partir des temps de correspondance $t_{co}^j$ calculés en (eq. 4). Nous distinguons ici les méthodes de TD-PSOLA, TDI-PSOLA et FDI-PSOLA selon l'association qui est faite entre marques d'écriture et temps de correspondance.

- Dans une approche simple (TD-PSOLA) nous prenons le $s_i(t)$ associé à la marque de lecture $t_r^i$ la plus proche de $t_{co}^j$.

  Le signal $\tilde{s}(t)$ est alors formé de la somme des $\tilde{s}_j(t) = s_i(t_r^i - t_w^j)$

  Cette méthode présente le désavantage d'introduire des discontinuités dans le signal de synthèse. En effet, pour des facteurs $d(t)$ importants, une même forme d'onde peut être recopiée plusieurs fois avant passage à la forme d'onde suivante.

- Une solution plus sophistiquée, inspirée de la concaténation entre diphones [CM88], consiste à prendre les signaux élémentaires associés aux marques $t_r^i$ et $t_r^{i+1}$ encadrant $t_{co}^j$ et à effectuer leur interpolation temporelle (TDI-PSOLA) ou fréquentielle (FDI-PSOLA).

  Le signal $\tilde{s}(t)$ est alors formé de la somme des signaux interpolés. Lors d'une dilatation du son le signal passe alors continuement d'une forme d'onde à une autre.

**Interpolation temporelle TDI-PSOLA**

L'interpolation temporelle s'effectue directement sur les formes d'onde $s_i(t)$ et $s_{i+1}(t)$ encadrant le temps de correspondance $t_{co}^j$.

$$\begin{cases} \tilde{s}_j(t) = (1-\alpha)s_i(t) + \alpha s_{i+1}(t) \\ \alpha = \dfrac{t_{co}^j - t_r^i}{t_r^{i+1} - t_r^i} \end{cases} \qquad (5)$$

**Interpolation fréquentielle FDI-PSOLA**

L'interpolation s'effectue sur les spectres d'amplitude et de phase des deux signaux encadrant $t_{co}^j$.

$\forall\, k \in [0, N]$

$$\begin{cases} \tilde{A}_j(k) = (1-\alpha)A_i(k) + \alpha A_{i+1}(k) \\ \tilde{\varphi}_j(k) = (1-\alpha)\varphi_i(k) + \alpha \varphi_{i+1}(k) \end{cases} \qquad (6)$$

L'interpolation des spectres de phase pose cependant un problème. La phase étant estimée en valeurs principales, on ne peut interpoler les spectres directement (exemple de l' interpolation de $\varphi_i(k) = -\pi$ avec $\varphi_{i+1}(k) = \pi$ donnant $\varphi = 0$). Il

est nécessaire d'ajouter une procédure de "déroulement" de la phase.

Le déroulement de la phase s'avère une manipulation tout aussi risquée puisque n'ayant de pertinence que dans les zones fréquentielles voisées. Toute erreur dans la détection du voisement peut être fatale.

Notre solution consiste à dérouler seulement le spectre de phase du deuxième signal élémentaire $\varphi_{i+1}(k)$ de manière à ce que $[\varphi_{i+1}(k) + 2n\pi - \varphi_i(k)] \leq \pi$.

#### 3.2.2 Dans les zones non-voisées :

Afin de permettre la dilatation de ces zones tout en évitant l'introduction d'une périodicité artificielle ("phasy-effect"), le signal élémentaire $s_i(t)$ déterminé par le $t_r^i$ le plus proche de $t_{co}^j$ est décalé aléatoirement autour de sa position. Nous appliquons de plus une inversion alternative de l'axe du temps de $s_i(t)$ lorsque celui-ci est répété successivement [CM88]. Cette méthode permet de conserver le spectre d'amplitude tout en brouillant le spectre de phase.

### 3.3 Overlap-Add

La synthèse du signal est alors effectuée par Superposition/Addition des $\tilde{s}_j(t)$. Nous utilisons la méthode de Griffin et Lim des moindres carrés [GL84] :

$$\tilde{s}(t) = \frac{\sum_j \tilde{s}_j(t)\tilde{h}_j(t - \tilde{t}_w(j))}{\sum_j \tilde{h}_j^2(t - \tilde{t}_w^j)} \qquad (7)$$

## 4 Observation des résultats

### 4.1 Comparaison TD/TDI/FDI-PSOLA

La méthode TD-PSOLA quoique très simple d'application introduit dans le signal des discontinuités qui n'apparaissent pas avec les méthodes utilisant l'interpolation. Cela se traduit perceptivement par une certaine "rugosité"du signal de synthèse.

L'expérience suivante est réalisée afin de permettre la comparaison des signaux de synthèse issus des deux méthodes d'interpolation avec le signal original.

- La durée et la hauteur des signaux de synthèse sont gardées identiques à celles du signal original.
- Les formes d'ondes utilisées pour la synthèse sont le résultat de l'interpolation des signaux élémentaires en posant $\alpha = 0.5$ (cas le plus critique).
- Les signaux sont comparés uniquement dans les zones voisées stables (la forme d'onde évolue très peu d'une période à l'autre).

D'un point de vue perceptif, l'interpolation fréquentielle préserve mieux la brillance du signal original comparée à l'interpolation temporelle. Ceci pondère quelque peu [CM88]. L'analyse de l'enveloppe spectrale du signal (calcul de l'enveloppe par filtre LPC d'ordre 40) sur l'ensemble d'une zone stationnaire du signal corrobore notre impression (fig. 2). Ceci peut s'expliquer en considérant l'interpolation temporelle en terme de filtrage passe-bas (somme de deux signaux élémentaires quasi-identiques mais dont le marquage $t_r^i$ diffère par rapport au signal élémentaire respectif).

L'observation des spectres du signal original et des signaux interpolés (fig. 3) nous montre que les deux interpolations donnent des résultats quasi-équivalents en basse fréquence. Aux fréquences élevées ($> 4000Hz$), les spectres des signaux interpolés diffèrent. Les raisons de cette divergence ne sont pas les mêmes pour les deux méthodes. En effet si l'influence d'une différence de marquage se traduit par un filtrage en TDI-PSOLA, en FDI-PSOLA celle-ci se répercute sur le spectre de phase. La rotation du spectre de phase entraînée rend leur interpolation difficile.

Les interpolations fréquentielles et temporelles sont donc l'une comme l'autre extrêmement sensibles à la position des marques de lecture.

### 4.2 Dilatation dans les zones non-voisées

Nous comparons notre méthode comprenant le décalage aléatoire des signaux élémentaires à la méthode utilisant seulement l'inversion alternative de l'axe du temps. Nous constatons une nette diminution de l'effet de périodicité artificielle. Notre méthode permet des dilatations du signal jusqu'à un facteur 3, sans apparition d'artefacts.

## 5 Application de la méthode PSOLA pour la transformation des sons musicaux

La méthode PSOLA est généralement utilisée pour la synthèse de la voix parlée. Nous l'avons appliquée pour la modification de sons de voix chantée avec des résultats très probants.

Un de nos buts est de l'appliquer aux sons d'instruments de musique.

Nous l'avons appliquée à des sons de trompettes. Les résultats sont là aussi très probants. Il s'agit de sons monophoniques présentant une forme d'onde

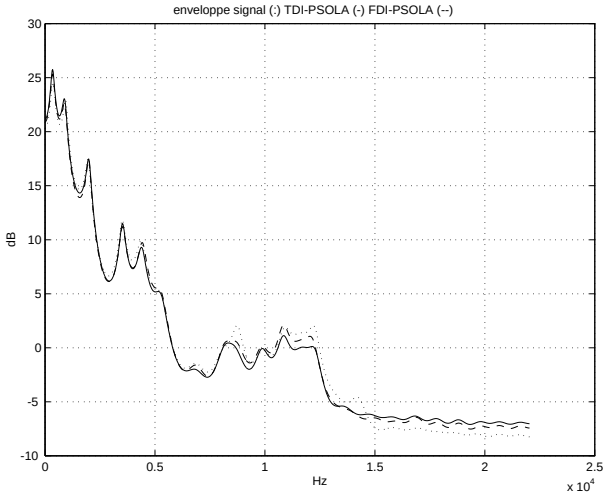


Figure 2: Comparaison des enveloppes spectrales du signal original, du signal traité en TDI-PSOLA et FDI-PSOLA (LPC ordre=40)

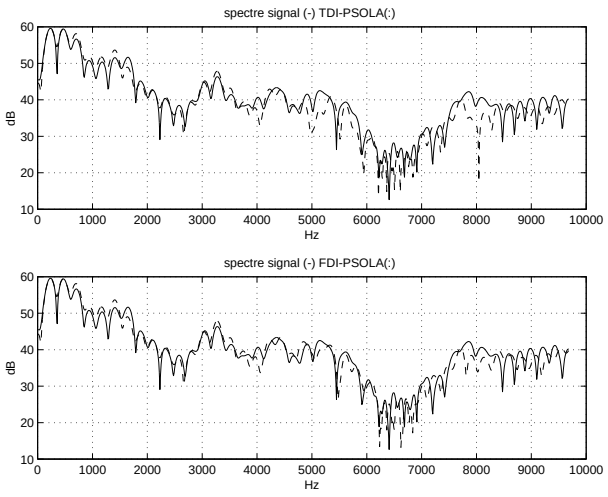


Figure 3: Comparaison des spectres du signal original, du signal traité en TDI-PSOLA et FDI-PSOLA

concentrée temporellement par rapport à sa période. Nous avons ensuite appliquée la méthode à des sons de violons et de pianos. Les résultats sont moins bons (rugosité), les formes d'ondes étant moins bien localisées temporellement, et le signal présentant une certaine inharmonicité.

Les conditions favorables à l'application de la méthode PSOLA semblent donc être: la nécessité d'un son monophonique, périodique ou pseudo-périodique, et de forme d'onde concentrée temporellement.

Des solutions devront être trouvées pour la synthèse des sons ne présentant pas ces caractéristiques.

# Conclusion

Nous avons exposé une procédure de marquage automatique s'adaptant aux caractéristiques du signal, en vue de son utilisation pour la synthèse des sons musicaux par la méthode PSOLA.

Deux méthodes visant à améliorer la qualité de la synthèse PSOLA dans les zones voisées ont été étudiées : la première utilise l'interpolation temporelle des signaux élémentaires, la seconde leur interpolation fréquentielle. Une analyse du signal de synthèse résultant de ces méthodes, nous montre que la seconde préserve mieux les caractéristiques du signal aux fréquences élevées. Une méthode permettant la dilatation des zones non-voisées du signal a également été proposée.

L'application de la méthode PSOLA pour la synthèse des sons musicaux nous a permis de déterminer les caractéristiques du signal favorables à l'application de la méthode.

Des exemples sonores seront donnés lors de la présentation de l'article.

# References

[CM88] F. Charpentier and E. Moulines, *Text-To-Speech Algorithms Based on FFT Synthesis*, ICASSP, 1988.

[CS86] F. Charpentier and M. Stella, *Diphone Synthesis Using an Overlap-Add Technique for Speech Waveforms Concatenation*, ICASSP, 1986.

[DR91] B. Doval and X. Rodet, *Estimation of Fundamental Frequency of Musical Sound Signals*, ICASSP, 1991.

[GL84] D. Griffin and J. Lim, *Signal Estimation from Modified Short-Time Fourier Transform*, IEEE Trans. ASSP, vol. ASSP-32, 1984, pp. 236–243.

[GL85] D. Griffin and J. Lim, *A New Model-Based Speech Analysis/Synthesis System*, ICASSP, 1985.

[Kor97] Reinier Kortekaas, *Physiological and psychoacoustical correlates of perceiving natural and modified speech*, Ph.D. thesis, TU Eindhoven, 1997.

[MC90] E. Moulines and F. Charpentier, *Pitch-Synchronous Waveform Processing Techniques for Text-To-Speech Synthesis using Diphones*, Speech Communication (1990).

[RDP87] X. Rodet, P. Depalle, and G. Poirot, *Speech Analysis and Synthesis Methods Based on Spectral Envelopes and Voiced/Unvoiced Functions*, European Conference on Speech Technology, 1987.