Non-stationary Analysis/Synthesis using Spectrum Peak Shape Distortion, Phase and Reassignment

Geoffroy Peeters, Xavier Rodet
Ircam - Centre Georges-Pompidou, Analysis/Synthesis Team,
1, pl. Igor Stravinsky, 75004 Paris, France
Geoffroy.Peeters@ircam.fr, Xavier.Rodet@ircam.fr

Abstract

A new Analysis/Synthesis method, named SINOLA, based on "sinusoidal additive" and "OLA/PSOLA" synthesis, is proposed. It allows high quality transformation of both stationnary and non-stationnary parts of a signal. Time-frequency characterization and synthesis parameters estimation is done by a novel method based on spectrum peak shape distortions and time-frequency phase evolutions.

Introduction

Speech and musical sound transformation plays an essential role in many applications today such as movie production (post-synchronization), musical studio effects (pitch-shifting, time-warping), Text-to-Speech, prosody matching and so on. Depending on the required quality and on the allowed complexity, several methods can be used , starting from the simplest, elementary resampling by changing the speed of reading from a circular read/write-buffer, to the most complex, the creation of an elaborate model of the source signal. At first, transformation of the signal can be obtained through a blind process, this is the case, for example, with the "phase-vocoder". But for better results, one can apply an Analysis/Synthesis (A/S) method. The analysis stage allows the extraction of the parameters necessary for an accurate transformation of the signal. These parameters will then be changed according to the modification desired and then used to synthesize the transformed signal.

In this paper, we propose SINOLA, a new sound transformation method which uses two different A/S methods : the "sinusoidal additive" and the "OLA/PSOLA" methods. Each of these methods is apropriate for parts of the signal having different characteristics.

1 The SINOLA model

"Sinusoidal additive" A/S consists of decomposing a signal into a sum of sinusoidal components with parameters varying slowly over time. This method is extremely accurate for signals which can be considered as a sum of sinusoids with stationary parameters in a window of 3 to 4 pitch periods. It allows high quality and extended sound transformation thanks to a complete control of sinusoidal parameters. However, it is not appropriate for transitory, non periodic pulses and random components, which are difficult to represent by slowly varying sinusoids.

On the other hand, Time-Domain Overlap-Add (TD-OLA) and TD-Pitch-Synchronous OLA (TD-PSOLA, which is important for periodic, i.e. harmonic sounds), are well adapted for non-stationary or non-sinusoidal components and require shorter windows.

In SINOLA, the sinusoidal additive A/S is used to model the stationary sinusoidal components while the OLA/PSOLA method is used to process attacks, transients, non periodic pulses and random components (see Figure 1 bottom).

SIN: Sinusoidal additive A/S model [5]

$$s(t) = \sum_l A_l(t) \cdot sin\left(\phi_l(0) + \int_0^t \omega_l(t)dt\right)$$

where $A_l(t)$, $\omega_l(t)$ and $\phi_l(0)$ are the amplitude, frequency and initial phase of the l^{th} frequency component of the signal. Usually $A_l(t)$ and $\omega_l(t)$ are supposed to be low-pass signals and are therefore considered constant during a short analysis frame. At the synthesis stage, these parameters are interpolated between adjacent frames in order to avoid signal discontinuities. In section 2.3 we show

how parameter variations can be included and evaluated in the analysis stage.

OLA: TD-OLA/TD-PSOLA method [3]

As opposed to sinusoidal additive A/S, OLA and PSOLA do not use a model. This can be viewed as a drawback since possibilities for sound modification are limited. But it can also be viewed as an advantage since the whole signal frame is taken into account, not only the stationary sinusoidal part. The OLA method consists of decomposing the signal into overlapping frames while PSOLA constrains these frames to be positioned in a pitch-synchronous way at the analysis and at the synthesis stage. A general formulation is:

$$\begin{cases} s_i(t) & = s(t) \cdot h_{L_i}(t - t_i) \\ s_i(t) & \rightarrow \tilde{s}_i(t) \\ \tilde{s}(t) & = \sum_j \tilde{s}_i(t - (t_j - t_i)) \end{cases}$$

where

- $s_i(t)$ is the i^{th} frame obtained by windowing the signal with a function $h_{L_i}(t)$ defined during a duration L_i and centered around time t_i,
- $\tilde{s}_i(t)$ is the modified i^{th} frame,
- $\tilde{s}(t)$ is the synthesis signal constructed by overlap-adding the successive frames positioned at the t_j.

In the case of PSOLA the t_i are positioned in a pitch-synchronous way, L_i is equal to 2 local pitch period and the positions of the t_j determine the pitch periods of the synthesis signal. The OLA/PSOLA method is depicted in Table 1 for each type of signal.

Frequency Shifted OLA / PSOLA

We introduce the Frequency Shifted OLA (FS-OLA) and Frequency Shifted PSOLA (FS-PSOLA) in order to allow low-cost spectral modifications of the sound and this, independently of the pitch and time modifications. As opposed to FD-PSOLA [3] which is based on spectrum resampling, FS-OLA and FS-PSOLA are based on spectrum shifting:

$$x(t) \cdot e^{j\Omega t} \rightleftharpoons X(\omega - \Omega) \qquad (1)$$

Unfortunatelly when (1) is applied without cares to an harmonic signal, the signal becomes inharmonic. However if (1) is applied to each fundamental waveform (FW) $s_i(t)$ separately, (1) results only in the shifting of the spectral envelope but does not change the

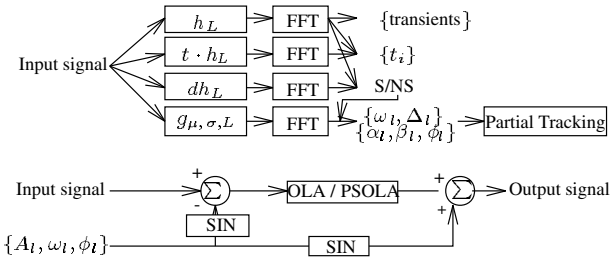

Figure 1: SINOLA : (top) Analysis stage (bottom) Synthesis stage

harmonicity properties. This is because one sole FW does not have any notion of pitch. Therefore $s_i(t)$ can be used as a rough approximation of the spectral envelope. The process is the following :

$$\tilde{s}_i(t) = \Re\left\{[s_a(t) \cdot h_{L_i}(t - t_i)]e^{-j\Omega t}\right\} \qquad (2)$$

where $s_a(t)$ denotes the analytic signal corresponding to $s(t)$, L_i is the size of $s_i(t)$ and Ω is the frequency shift factor. $\tilde{s}_i(t)$ is then processed by the PSOLA method giving the required pitch. In FS-OLA, the same frequency shifting is applied but this time without any consideration about pitch and harmonicity.

2 Parameter Estimation

Three types of information are needed for SINOLA and are retrieved simultaneously using the Short Time Fourier Transform (STFT) of the signal (see Figure 1 top):

1. a time-frequency characterization of the signal for its decomposition into transients, sinusoidal and non-sinusoidal components (see 2.1, 2.2),
2. the time-varying frequency, amplitude and phase of the sinusoidal components (see 2.3),
3. the pitch-synchronous markers in the case of PSOLA (see 2.4).

2.1 Transient detection

Transients are detected using cross-entropy measurement derived from the Kullback-Leibler distance [2]:

$$D_{KL}(t) = \sum_{\omega_k} F\left(\frac{A(t_i, \omega_k)}{A(t_{i+1}, \omega_k)}\right) \qquad (3)$$

Table 1: OLA - PSOLA method for different types of signals

Type	**transient**	**random**	**random (+ periodic part)**	**periodic**
Method	OLA	OLA	OLA-PSOLA	PSOLA
t_i	= transient positions	$t_{i+1} - t_i = T_0(t)$ [1] + random component	= t_i of periodic part + random component	$t_{i+1} - t_i$ = original signal pitch period (see 2.4)
t_j	= transient positions	$t_{j+1} - t_j = \overline{T_0(t)}$	= t_j of periodic part	$t_{j+1} - t_j$ = synthesis signal pitch period
$\tilde{s}_i(t)$	$s_i(t)$	alternate time reversing + morphing between $s_i(t)$ and $s_{i+1}(t)$	alternate time reversing + morphing between $s_i(t)$ and $s_{i+1}(t)$	morphing between $s_i(t)$ and $s_{i+1}(t)$

where $F(x) = x - \log(x) - 1$, and $A(t, \omega_k)$ is the amplitude of the STFT at time t and frequency ω_k.

2.2 S/NS signal characterization

The Sinusoidal versus Non-sinusoidal (S/NS) signal characterization consists of measuring how well a part of the time/frequency plane can be represented by a sinusoidal model. It is therefore strongly dependent on the assumptions defining the sinusoidal model: local stationarity or non-stationarity of the sinusoidal parameters. Numerous methods have been proposed for S/NS characterization (see [8] for a review) but most of them use this stationarity assumption.

In [6] we have proposed a method, called the **"Phase Derived Sinusoidality Measure"** (PDSM), which allows measurement of sinusoidality without a stationary frequency assumption. For this, PDSM compares a temporal model of the evolution of measured frequencies and a temporal model of the corresponding phase derivative. For a specific frequency, if the models are close (according to a distance measure) this band can be represented by a sinusoidal model. We give here a low-cost method to compute it using **frequency "reassignment"** [1] which can be written (using band-pass convention):

$$\begin{aligned} \omega_r(t,\omega) &= \frac{\partial}{\partial t}\phi^{\mathrm{BP}}(t,\omega) \\ \omega_r(t,\omega_k) &= \omega_k - \Im\left\{\frac{\mathrm{STFT}^{\mathrm{BP}}_{dh}(x)}{\mathrm{STFT}^{\mathrm{BP}}_{h}(x)}\right\} \end{aligned} \quad (4)$$

The first formulation of $\omega_r(t,\omega)$ is the **instantaneous frequency** definition which is often used in order to obtain precise frequencies from a Discrete Fourier Transform. The second formulation gives a low cost method to compute it. It also expresses the correction to apply to the discrete frequency ω_k in order to obtain the exact frequency. The distance given by PDSM can be shown to be similar to this correction.

2.3 Complex Short-Time Spectrum Distortion measure

In classical A/S methods, parameters are often estimated from short-time spectra. The signal is usually assumed to be stationary during the analysis window and, thus, the spectrum is assumed to have peaks at the frequencies of the sinusoidal components. Unfortunately, the signal is rarely stationary during the analysis window: amplitude and frequency modulation of signal components distort the shape of the assumed spectral peaks, therefore inducing incorrect parameter estimation. Previous studies have shown the importance of spectrum distortion induced by these variations and have proposed partial solutions (neural network, signal normalization [6]), or analytical formulation [4]. We propose here a complete parameter estimation method taking into account amplitude and frequency modulation.

The **signal model** is a sum of sinusoids with linear variation of amplitude ($\alpha_l + \beta_l t$) and of frequency ($\omega_l + 2\Delta_l t$). $\phi_{0,l}$ is the initial phase and l is the peak index. For t in the i^{th}

[1] $\overline{T_0(t)}$ means "average pitch period of neighboring periodic regions"

frame centered on t_i, (we note $\tau_i = t - t_i$):

$$s(t) = \sum_l (\alpha_{l,i} + \beta_{l,i}\tau_i)\cos(\phi_{0,l,i} + \omega_{l,i}\tau_i + \Delta_{l,i}\tau_i^2)$$

The **Short Time Complex Spectrum** is estimated using a truncated gaussian window $g_{\mu,\sigma,L}(t)$ where μ and σ are the mean and standard deviation of the gaussian function and L is the size of the truncation (L must be greater than 9σ in order to reduce the truncation effect). The Distortion is measured by fitting a second order polynomial around each log-amplitude spectrum peak ($\mathrm{P}_{\log}(\omega) = a_{\log}\omega^2 + b_{\log}\omega + c_{\log}$) and around each corresponding unwrapped phase spectrum region ($\mathrm{P}_\phi(\omega) = a_\phi\omega^2 + b_\phi\omega + c_\phi$). For a specific peak index l, parameters are given by:

$$\begin{cases} \omega_l \simeq & -\frac{1}{2}\frac{b_{\log}a_{\log} + a_\phi b_\phi}{G} \\ \Delta_l \simeq & -\frac{1}{4}\frac{a_\phi}{G} \\ \frac{\beta_l}{\alpha_l} \simeq & \frac{1}{2}\frac{b_\phi a_{\log} - b_{\log}a_\phi}{G} \\ log(\alpha_l) \simeq & \mathrm{P}_{\log}(\omega_{\max}) + \frac{1}{4}\log(D_l) \\ & -\frac{\beta_l^2}{\alpha_l^2}\frac{2\Delta_l^2\sigma^6}{D_l} \\ \phi_{0,l} \simeq & \mathrm{P}_\phi(\omega_{\max}) - \frac{1}{2}\mathrm{atan}(2\Delta_l\sigma^2) \\ & +\frac{\beta_l^2}{\alpha_l^2}\frac{2\Delta_l\sigma^4}{D_l}(1+2\Delta_l^2\sigma^4) \end{cases} \quad (5)$$

where $G = a_{\log}^2 + a_\phi^2$, $D_l = 1 + 4\Delta_l^2\sigma^4$ and $\omega_{\max} = -b_{\log}/(2a_{\log})$

It is easy to show that usual sinusoidal estimators of frequency, amplitude and phase ($\omega_{\max}$, $\mathrm{P}_{\log}(\omega_{\max})$ and $\mathrm{P}_\phi(\omega_{\max})$) have a bias proportional to the frequency and amplitude modulation and to the size of the window (see [7] for details).

Partial Tracking with time-varying parameters

Once the sinusoidal parameters are estimated, the peaks of adjacent analysis frames are connected to form frequency tracks. This is called "Partial Tracking". Usual partial tracking methods consider three successive frames in order to construct a track. Since the time derivatives of parameters are part of our model, it suffices to consider only two frames (t_i, t_{i+1}) together. For each couple of peaks $(m(t_i), n(t_{i+1}))$, a track-score θ is computed. In a frequency band, the couple that leads to the maximum score (if this score is above a certain threshold) is chosen. If the maximum score is below the threshold, there is a birth, a death or no track in this band.

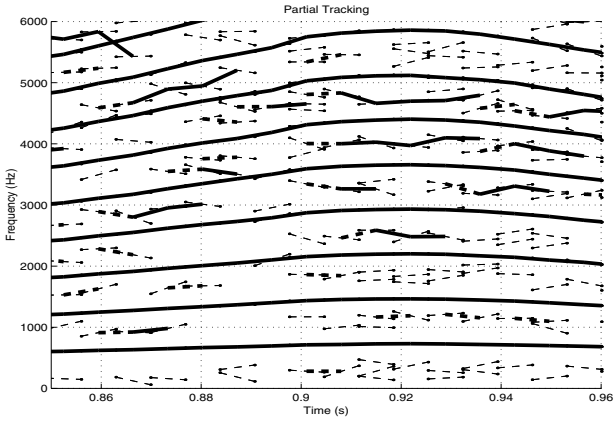

Figure 2: Partial Tracking method: frequency and frequency slope estimations (thin dashed lines), partial births (thick dashed lines), partials (thick lines), signal: female singing voice, window size: 14 ms, analysis step: 7 ms

$$\theta(m,n) = \exp\left(-\frac{c_f^2(m,n)}{\sigma_f^2} - \frac{c_a^2(m,n)}{\sigma_a^2}\right) \quad (6)$$

where $c_f^2(m,n)$ and $c_a^2(m,n)$ are the maximum curvature [2] of 3rd order polynomials with the following boundary conditions : for frequency $\{\omega_{m,i}; \Delta\omega_{m,i}; \omega_{n,i+1}; \Delta\omega_{n,i+1}\}$, for amplitude $\{a_{m,i}; \Delta a_{m,i}; a_{n,i+1}; \Delta a_{n,i+1}\}$. σ_f^2 and σ_a^2 are model parameters. Results obtained with (6) are shown in Figure 2.

2.4 PSOLA markers positioning

PSOLA markers (noted t_i) have to be placed in a pitch synchronous way, i.e. the distance between two markers must be equal to the local pitch period. Moreover, because of the windowing applied in the PSOLA method, the markers must be close to the local maxima of signal energy. In speech processing, Glottal Closure Instants (GCI) detection methods are used in order to place PSOLA markers [9]. But for musical signals, GCI methods

[2] second order derivative

are not relevant. This is why other methods, which use phase spectrum information, have been proposed. But then, we cannot guarantee that markers will be close to local maxima of energy. In order to fulfill both periodicity and energy conditions we propose here a new method based on **group delay**. The method uses a weighted sum of frequency component group delays. The weighting is made according to component amplitudes. Let us define:

$$f(t) = t + \frac{\sum_{\omega_k} \mathrm{Gd}(t,\omega_k)\, A(t,\omega_k)}{A(t,\omega_k)} \qquad (7)$$

where $\mathrm{Gd}(t,\omega_k)$ is the group delay of frequency ω_k for a window centered at time t. $\mathrm{Gd}(t,\omega_k)$ can be computed in an efficient way using **time "reassignment"** [1] which can be written (using band-pass convention):

$$\begin{aligned} t_r(t,\omega) &= t - \frac{\partial}{\partial\omega}\phi^{\mathrm{BP}}(t,\omega) \\ t_r(t,\omega_k) &= t + \Re\left\{ \frac{\mathrm{STFT}^{\mathrm{BP}}_{(s-t)h}(x)}{\mathrm{STFT}^{\mathrm{BP}}_{h}(x)} \right\} \end{aligned} \qquad (8)$$

Marker positions are then given by the local minima of the time derivative of $f(t)$: $\partial f(t)/\partial t$ (special care has to be taken as $f(t)$ is not injective). Because of the windowing applied before computing Gd, a confidence measure of $f(t)$ must be computed for each t. It is given by an amplitude weighted standard deviation (in ω_k) of the $\mathrm{Gd}(t,\omega_k)$. Large std values mean small confidence while small std values mean large confidence. Results obtained with this new method are shown in Figure 3.

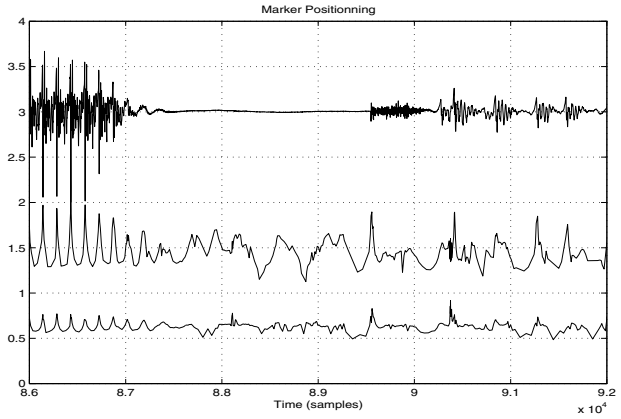

Figure 3: PSOLA markers positioning: signal (top), confidence measure (middle), inverse of the derivative of $f(t)$ (bottom), signal: male speech voice, window size: 20 ms, analysis step: 1 ms

Conclusion

From spectrum analysis SINOLA derives all the information necessary for high quality sound processing such as time warping, pitch shifting, spectrum dilatation and so on. Because of its dual processing (SIN + OLA), it preserves the inherent local characteristics of the signal (sinusoidal, random-noise, attacks-transients) and allows easy and natural modifications of the signal. Examples of the sound quality obtained with this method will be given during the presentation of this paper.

References

[1] F. Auger and P. Flandrin. Improving the Readibility of Time-Frequency and Time-Scale Representations by the Reassignment Method. *IEEE Trans. Signal Processing*, 43(5):1068–1089, 1995.

[2] M. Basseville. Distance Measures for Signal Processing and Pattern Recognition. *Signal Processing*, 18:349–369, 1989.

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

[4] J. Marques and L. Almeida. A Background for Sinusoid Based Representation of Voiced Speech. In *ICASSP*, Tokyo, 1986.

[5] R. McAulay and T. Quatieri. Speech Analysis/Synthesis based on a Sinusoidal Representation. *IEEE Trans. Acoust. Speech Signal Process*, 34(4):744–754, 1986.

[6] G. Peeters and X. Rodet. Sinusoidal versus Non-Sinusoidal Signal Characterisation. In *COST-G6 DAFX*, Barcelona, 1998.

[7] G. Peeters and X. Rodet. SINOLA : A New Analysis/Synthesis Method using Spectrum Peak Shape Distortion, Phase and Reassigned Spectrum. In *ICMC*, Peking, 1999.

[8] G. Richard and C. d'Alessandro. Analysis/Synthesis and Modification of the Speech Aperiodic Component. *Speech Communication*, (19):221–244, 1996.

[9] H. Strube. Determination of the Instant of Glottal Closures from the Speech Wave. *J. Acoust. Soc. Am.*, 56(5):1625–1629, 1974.