SINOLA: A New Analysis/Synthesis Method using Spectrum Peak Shape Distortion, Phase and Reassigned Spectrum

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

In this paper we present a new Analysis/Synthesis method named SINOLA, which benefits from both sinusoidal additive model and OLA/PSOLA method, and which allows adequate processing according to the inherent local characteristics of the signal. All the parameters of the models are derived at the same time from spectrum analysis. We propose an analytical formulation of a Complex Short-Time Spectrum Distortion measure, which allows the retrieval of precise sinusoidal parameters as well as their slopes. A new partial tracking method is proposed which benefits from these informations. Reassigned Spectrum is used in both time and frequency in order to characterize the signal and to position the PSOLA markers.

Introduction

Sinusoidal additive Analysis/Synthesis (A/S) is extremely accurate for signals which can be considered as a sum of sinusoids with stationary parameters in a window of 3 to 4 fundamental periods. On the other side, 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. We present a new A/S method, named SINOLA, which benefits from both the sinusoidal additive A/S and OLA/PSOLA method.

1 The SINOLA model

In SINOLA, the sinusoidal additive model is used for the stationary sinusoidal components while OLA method is used to process attacks, transients, non or nearly periodic pulses and random components.

- SIN: Sinusoidal additive A/S model [6]

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

  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 in 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:

$$\begin{cases} s_i(t) & = s(t) \cdot h_{2L_i}(t - t_i) \\ s_i(t) & \rightarrow \tilde{s}_i(t) \\ \tilde{s}(t) & = \sum_j \tilde{s}_i(t - (t_j - t_i)) \end{cases}$$

  where

  - $s_i(t)$ is the i^{th} frame obtained by windowing the signal with a function $h_{2L_i}(t)$ defined on a duration $2L_i$ 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 the local fundamental period and the positions of the t_j determine the fundamental periods of the synthesis signal. The OLA/PSOLA method is depicted in Table 1 for each type of signal.

2 Parameter Estimation

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

1. a time-frequency characterization of the signal for its decomposition into transients, sinusoidal and non-sinusoidal components (see 2.1, 2.2, 2.3.2),

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

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

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]:

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

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 Sinusoidal versus Non-sinusoidal (S/NS) signal characterization

The S/NS signal characterization consists in 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 [7] we have proposed a method, called the "Phase Derived Sinusoidality Measure" (PDSM), which allows to measure the sinusoidality coefficient without a stationary frequency assumption. PDSM was based on the following considerations:

- for the main-lobe of a sinusoidal component, the frequency derived from the complex spectrum and the frequency derived from the evolution of the corresponding phase spectrum are the same
- when parameter stationarity is not assumed, we cannot derive a sinusoidality measure from an instantaneous measurement only, but through the continuity of the parameters along time.

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

Therefore PDSM compares a temporal model of the evolution of measured frequencies and a temporal model of the corresponding phase derivative. But the measurements used in [7] to create the models were biased (see 2.3.1), because taken from a stationary model. In section 2.2.1, we show how the bias in frequency can be avoided by bypassing the use of a model. In section 2.3, we propose a new model which takes into account modulation of amplitude and frequency.

2.2.1 PDSM using frequency "reassignment"

"Reassignment" [1] has been proposed to improve time-frequency representations. In usual time/frequency representations, the values obtained when decomposing the signal on the time/frequency atoms are assigned to the geometrical center of the cells (center of the analysis window and bins of the FFT). In [1] it is proposed to assign each value to the center of gravity of the cell's energy. Frequency reassignment can be written [1] (using band-pass convention):

$$\begin{aligned} \omega_r(t,\omega) &= \Re\left\{\frac{\int \xi X(\xi) H^*(\xi-\omega) e^{j\xi t} d\xi}{STFT_h^{BP}(x)}\right\} \\ &= \frac{\partial}{\partial t}\phi^{BP}(t,\omega) \\ \omega_r(t,\omega_k) &= \omega_k - \Im\left\{\frac{STFT_{dh}^{BP}(x)}{STFT_h^{BP}(x)}\right\} \end{aligned} \quad (2)$$

The second 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 third formulation 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, but using (2) we do not face the frequency bias cited above. The third formulation also provides a low cost method to compute the instantaneous frequency and to measure the sinusoidality.

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 on 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 on 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 [5], signal normalization [7]), 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} frame centered on t_i, (we note $\tau_i = t - t_i$):

$$s(t) = \sum_l (\alpha_{l,i} + \beta_{l,i}\tau_i)\cos(\phi_{0,l,i} + \omega_{l,i}\tau_i + \Delta_{l,i}{\tau_i}^2) \tag{3}$$

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 ($a_{\log}\omega^2 + b_{\log}\omega + c_{\log}$) and around each corresponding unwrapped phase spectrum region ($a_\phi\omega^2 + b_\phi\omega + c_\phi$). For a specific peak index, parameters are given by:

$$\begin{cases} \omega_l & \simeq -\dfrac{1}{2}\dfrac{b_{\log}}{a_{\log}} - \dfrac{\beta_l}{\alpha_l}2\Delta_l\sigma^2 \\ \Delta_l & \simeq \dfrac{-1 \pm \sqrt{1 - 16a_\phi^2/\sigma^4}}{8a_\phi} \\ \dfrac{\beta_l}{\alpha_l} & \simeq 2\omega_l\Delta_l\sigma^2 - \dfrac{D_l}{\sigma^2}b_\phi \end{cases} \tag{4}$$

where $D_l = 1 + 4\Delta_l^2\sigma^4$

2.3.1 Bias of usual sinusoidal estimators

From 3 and 4 it is easy to show that

- the frequency of the maximum of the log-amplitude spectrum (noted ω_{max} and usually considered as the frequency position of the sinusoidal component) is in fact at $\omega_l + \frac{\beta_l}{\alpha l}2\Delta_l\sigma^2$. Therefore usual frequency estimators have a bias proportional to the amplitude modulation, to the frequency modulation and to the length of the analysis window.
- a similar bias is found for the log-amplitude of the spectrum at ω_{max} which is equal to $\log(\alpha_l) - \frac{1}{4}\log(D_l) + \frac{\beta_l^2}{\alpha_l^2}\frac{2\Delta_l^2\sigma^6}{D_l}$ instead of $\log(\alpha_l)$
- a similar bias is found for the phase of the spectrum at ω_{max} which is equal to $\phi_{0,l} + \frac{1}{2}\text{atan}(2\Delta_l\sigma^2) - \frac{\beta_l^2}{\alpha_l^2}\frac{2\,\Delta_l\sigma^4}{D_l}(2\Delta_l^2\sigma^4 + 1)$ instead of $\phi_{0,l}$

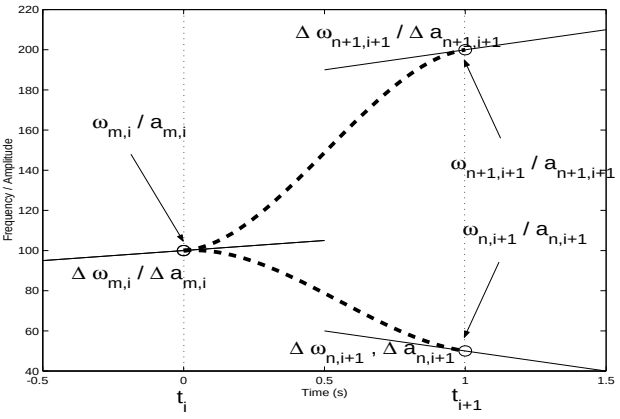

Figure 1: Curvature computing for couples of peaks (m,n) and (m,n+1)

2.3.2 Sinusoidality measure and Partial Tracking with time-varying parameter

Extending the sinusoidal model with linear variations renders S/NS estimation more difficult. As suggested in 2.2, information about sinusoidality can only be given by the continuity of the model parameters along time. This can be evaluated by a partial tracking method. 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 together. For each couple of peaks (m,n) (see Figure 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.

$$\theta(m,n) = \exp\left(-\frac{c_f^2(m,n)}{\sigma_f^2} - \frac{c_a^2(m,n)}{\sigma_a^2}\right) \tag{5}$$

where $c_f^2(m,n)$ and $c_a^2(m,n)$ are the maximum curvature [2] of the 3rd order polynomials with the following boundary conditions (see Figure 1): 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 (5) 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 fundamental 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]. These GCI occur pitch-synchronously and are close to the local maxima of energy. For musical signals, GCI methods are not relevant.

[2] second order derivative

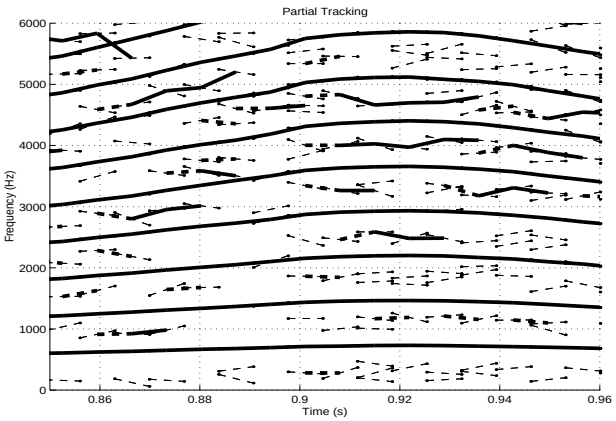

Figure 2: Partial Tracking method: frequency and frequency slope estimations (thin dashed lines), partial births (thick dashed lines), partials (thick lines), signal: female singing voice, window size: 14 ms, analysis step: 7 ms

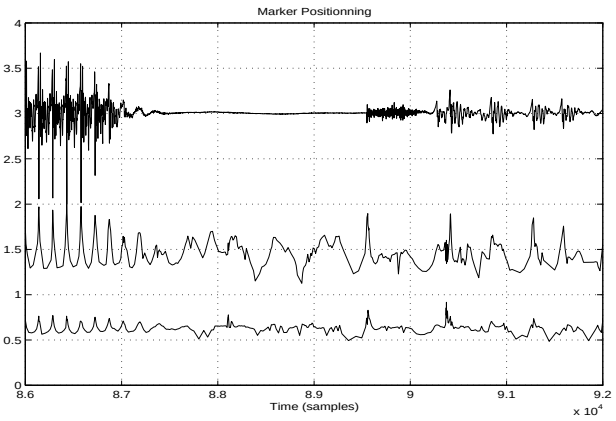

Figure 3: PSOLA markers positioning: signal (top), confidence measure (middle), inverse of the derivative of $f(t)$ (bottom), signal: male speech voice, window size: 20 ms, analysis step: 1 ms

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:

$$f(t) = t + \frac{\sum_{\omega_k} \mathrm{Gd}(t,\omega_k) A(t,\omega_k)}{A(t,\omega_k)} \qquad (6)$$

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):

$$\begin{aligned} t_r(t,\omega) &= t + \Re\left\{ \frac{\int (s-t)x(s)h^*(t-s)e^{j\omega(t-s)}ds}{STFT_h^{BP}(x)} \right\} \\ &= t - \frac{\partial}{\partial\omega}\phi^{BP}(t,\omega) \\ t_r(t,\omega_k) &= t + \Re\left\{ \frac{STFT^{BP}_{(s-t)h}(x)}{STFT_h^{BP}(x)} \right\} \end{aligned} \qquad (7)$$

where we recognize, in the second formulation, the group delay definition. As explained in the following, this relates the new PSOLA marker positioning method to time reassignment. The third formulation gives a method for computing the group delay at low cost.

Marker positions are then given by the local maxima of the inverse of the derivative of $f(t)$ (special care has to be taken considering that $f(t)$ is not injective).

$$t_i = \max_t \left(1 / \frac{\partial f(t)}{\partial t} \right) \qquad (8)$$

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.

Conclusion

SINOLA derives from spectrum analysis 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] P. Masri. *Computer Modeling of Sound for Transformation and Synthesis of Musical Signals*. PhD thesis, University of Bristol, 1996.

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

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

[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.