Enhancing the Tracking of Partials for the Sinusoidal Modeling of Polyphonic Sounds

Mathieu Lagrange, Sylvain Marchand, *Member, IEEE*, and Jean-Bernard Rault

***Abstract*—This paper addresses the problem of tracking partials, i.e., determining the evolution over time of the parameters of a given number of sinusoids with respect to the analyzed audio stream. We first show that the minimal frequency difference heuristic generally used to identify continuities between local maxima of successive short-time spectra can be successfully generalized using the linear prediction formalism to handle modulated sounds such as musical tones with vibrato. The spectral properties of the evolutions in time of the parameters of the partials are next studied to ensure that the parameters of the partials effectively satisfy the slow time-varying constraint of the sinusoidal model. These two improvements are combined in a new algorithm designed for the sinusoidal modeling of polyphonic sounds. The comparative tests show that onsets/offsets of sinusoids as well as closely spaced sinusoids are better identified and stochastic components are better avoided.**

***Index Terms*—Partial-tracking algorithms, polyphonic audio analysis, sinusoidal modeling.**

I. Introduction

A MONOPHONIC sound produced by the vocal tract or a musical instrument may be decomposed into a stationary pseudoperiodic part, often named the deterministic part of the signal (voiced speech signal, sustain and release phases of tones produced by resonant instruments) and a stochastic part (turbulences, unvoiced speech signals). The deterministic part can be conveniently decomposed into *partials*. Each partial is usually corresponding to a mode of vibration of the producing sound system and is modeled as a sinusoid with given amplitude, phase, and frequency.

Observing that the distinction between deterministic and stochastic processes may not be necessary for speech signal modification and reconstruction purposes, McAulay and Quatieri developed in [1] an analysis/synthesis system with applications to time-scale, pitch-scale modifications, and mid-rate speech coding. Voiced signals are represented as sums of sinusoids with frequencies nearly harmonically related, and unvoiced signals are represented as sums of sinusoids sufficiently close in frequency so that the Karhunen–Loève expansion [2] constraint is satisfied.

Manuscript received July 29, 2005; revised December 7, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Malcolm Slaney.

M. Lagrange is with the Computer Science Department, University of Victoria, Victoria, BC V8W 2Y2, Canada (e-mail: lagrange@uvic.ca).

S. Marchand is with LaBRI (Computer Science Laboratory), CNRS–University of Bordeaux 1, Talence cedex F-33405, France (e-mail: sylvain.marchand@labri.fr).

J.-B. Rault is with France Telecom R&D, Cesson-Sévigné cedex F-33512, France (e-mail: jeanbernard.rault@orange-ftgroup.com).

Digital Object Identifier 10.1109/TASL.2007.896654

Although this system achieves good signal reconstruction quality even in case of polyphonic audio signals, this system is not intended to identify the underlying structure of the pseudoperiodic part of the analyzed sound. As it will be demonstrated in the paper, this underlying structure can be better modeled with a relevant design of a part of the system called the partial-tracking (PT) algorithm.

Some PT algorithms were proposed in the literature to better identify the partials of the sound with harmonic assumptions [3], [4]. Additionally, some post-processing methods were proposed again with harmonic assumptions to overcome the problem of partials with close or crossing frequencies [5]–[7] that occurs in polyphonic recordings.

Since many pseudoperiodic sounds are not harmonic, more generic PT algorithms were proposed by considering a statistical framework as in [8] and in our previous work introduced in [9]. However, our experience is that the relatively loose relationship between the problem and its mathematical formulation leads to a difficult parameterization of this type of algorithms. Their relatively high complexity can also be an issue for specific applications. The proposed approach distinguishes from previous ones by relying solely on the physical properties of the sinusoidal model. The estimated partials are ensured to respect constraints of the model and can therefore be safely used as a front-end representation for auditory scene analysis applications, such as low bit-rate coding [10], [11] or audio content indexing [12].

Without any assumption about the sources that compose the mixture, a perfect identification of each partial of each source can generally not be achieved. For example, partials of unison notes or notes with musical pitch relationship can hardly be separated due to time/frequency resolution issues, leading to a contaminated representation of the spectral content of the sound. We will show that the proposed PT algorithm is useful to locally overcome these contamination problems by assuming that the uncontaminated portion of the spectrum reflects fairly perceptively important parts of the analyzed sound. Furthermore, the sinusoidal modeling of this uncontaminated part can also be used to recover contaminated parts of wider range using post-processing methods as proposed in [7] and [13].

This paper is organized as follows. The estimation of the parameters of the partials at discrete time locations is first described in Section II. The resulting short-term sinusoidal (STS) model is used by PT algorithms to determine the continuous evolutions of the parameters of the partials. After a presentation of the long-term sinusoidal (LTS) model and a review of several existing PT algorithms in Section III, the general structure of the proposed PT algorithm is introduced in Section IV.

The performance of this algorithm relies on the precise prediction of the evolutions of the frequency and amplitude parameters of the partials in untracked STS frames and the selection of a continuation with parameters close to the predicted ones that comply with the slow time-varying constraint of the LTS model. These two key elements were previously addressed, respectively, in [14] and in [15]. We advance in this work by extensively developing theoretical and implementation issues in Sections V and VI of this paper.

The gain of using and combining these prediction and selection methods in an enhanced PT framework is next studied. Experiments are first presented in Section VII using some test-case studies, each reflecting a particular issue in the sinusoidal modeling of polyphonic sounds. The performances are then evaluated using an original methodology in Section VIII that aims at evaluating PT algorithms independently of any other elements of the analysis/synthesis chain.

II. Short—Term Sinusoidal Analysis

Sinusoidal modeling aims at representing a sound signal as a sum of sinusoids having given amplitudes, frequencies, and phases. It is rooted in Fourier's theorem, which states that any periodic function can be modeled as a sum of sinusoids at various amplitudes and harmonic frequencies. Since considering these parameters as constant through the whole signal duration is not perceptually relevant, a first approach segments the signal into small successive frames. The size of these—often overlapping—frames N as well as the hop size H are determined according to the local stationarity of the signal. The discrete signal $x^k(n)$ at frame index k is then modeled as follows:

$$x^k(n) = \sum_{l=1}^{L^k} a_l^k \cos\left(\frac{2\pi}{F_s} f_l^k \cdot n + \phi_l^k\right) \qquad (1)$$

where F_s is the sampling frequency, and ϕ_l^k is the phase at the beginning of the frame of the lth component of L^k sine waves, f_l^k and a_l^k are, respectively, the frequency in Hertz and the amplitude considered as constant within the frame.

A set of sinusoidal parameters $\mathcal{S}^k = \{p_1^k, \ldots, p_{L^k}^k\}$ is used to model each frame k. The system parameters of this STS model $\mathcal{S}^k$ are the L^k triplets $p_l^k = (f_l^k, a_l^k, \phi_l^k)$, often called *peaks*. These parameters can be efficiently estimated by picking some local maxima from a short-term Fourier transform (STFT) using spectral techniques detailed next.

A. Time/Frequency Analysis

To estimate each set of peaks $\mathcal{S}^k$, the spectrum X^k is computed using a discrete Fourier transform (DFT) operated on the windowed samples of frame k. For a robust estimation of the phase [16], in the case of the use of a periodic Hann window with an even N, the weighted samples can be circularly shifted by $N/2$ samples before the computation of the DFT (zero-phase windowing).

The number of samples necessary for the computation of the DFT is constrained by the spectral structure of the analyzed sound. Since the frequencies of the sinusoidal components of an harmonic monophonic sound are separated by the fundamental frequency, the size of the DFT can be adapted according to a pitch estimate [1]. The frequency resolution is then sufficient to separate harmonics and the time resolution is close to optimal. On the contrary, the frequency distribution of the sinusoidal components of polyphonic sounds is not known in advance. The size of the window should then be set arbitrarily. If the frequency resolution is too small, two sinusoidal components may lay in the same frequency bin and may be misdetected. On the other hand, the loss of temporal resolution and short-term stationarity may lead to poor—averaged—estimates.

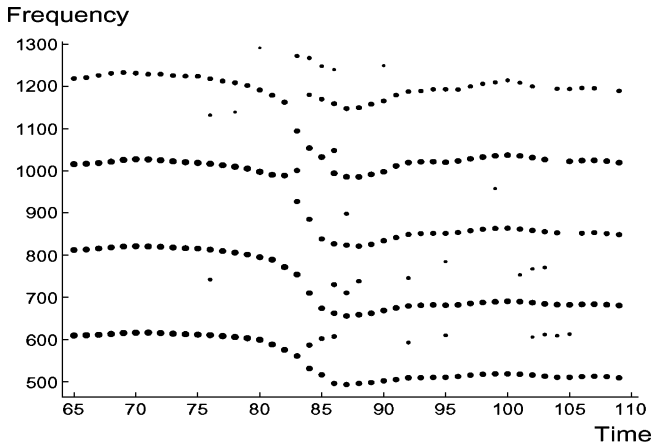

Fig. 1. Spectral peaks (spectrum local maxima) of a singing voice sampled at 44 100 Hz and analyzed with a hop size of $H = 512$ samples and a window size of $N = 2048$. Since the window size is four times larger than the hop size, some spurious peaks appear with modulated partials (around frame 85).

In order to reduce interpretation problems due to the bad temporal resolution, the hop size can be set significantly shorter than the window length. This method is far from perfect though because even if the spectrum is estimated at a given rate, the number of samples used to estimate the spectrum is significantly larger. This leads to a temporal "smearing" of the representation as shown in Fig. 1.

During our experiments, the STFT of a signal sampled at $F_s = 44\,100$ Hz is done with a window length of $N = 2048$ points since a shorter window length is not convenient for the analysis of polyphonic sounds, and a larger one gives very poor estimates in case of modulations. The parameters of the underlying sine waves are evolving with time and the analysis of these evolutions will be a key factor for the enhancements proposed in this article. The hop size is set to $H = 512$ points to provide a sufficiently dense sampling of these modulations (≈86 Hz). The window size and the hop size are then respectively of 46.4 and 11.6 ms.

B. Peak Parameters Estimation

Next, the parameters of the elements of $\mathcal{S}^k$ as well as its cardinal are estimated given X^k, the complex values of the DFT spectrum. The frame index k is omitted in the remainder for clarity sake. Supposing that the absolute difference between the frequencies of two sinusoidal components is at least the width of the spectrum main lobe of the window (F_s/N if the rectangular window is used), each component gives rise to a local maximum in the power spectrum, located at the DFT index m_l so that $(m_l - 0.5)F_s/N \leq f_l \leq (m_l + 0.5)F_s/N$.

The value of the index m_l, the magnitude, and the phase of the spectrum bin give, respectively, rough estimates of the frequency, amplitude, and phase of p_l

$$\hat{f}_l = m_l \frac{F_s}{N} \qquad \hat{a}_l = \frac{2}{N}\left|X(m_l)\right| \qquad \hat{\phi}_l = \angle X(m_l). \quad (2)$$

Since we consider real signals, the power spectrum value is multiplied by 2 to estimate the amplitude and due to the shifting of the samples, ϕ_l is the estimated phase at the center of the window.

To increase the precision of the frequency estimate, we can consider the relationship between the amplitude of the signal and the one of its derivative

$$\hat{f}_l = \frac{F_s}{\pi} \arcsin\left(\frac{1}{2F_s}\frac{|X_{\_}(m_l)|}{|X(m_l)|}\right) \quad (3)$$

where $X_{\_}$ denotes the spectrum of difference $x(n+1)-x(n)$, approximating the first derivative of x, see [17], [18] for further explanations. This estimator relies on a precise estimation of the amplitude of the sinusoidal signal and the one of its derivative. The analysis window should be chosen to achieve good asymptotic side lobe attenuation because these amplitudes are estimated with the DFT bins of the spectra $X_{\_}$ and X. The periodic Hann window has proven to give the best results [19] and thus is used in all our experiments.

During the analysis of natural sounds, bin contamination or noise may lead to incoherent estimates. If the frequency $\hat{f}_l$ of a local maximum located at DFT bin m_l is closer to the frequency of another DFT bin, the local maximum should have been located at this bin. Therefore, a local maximum with an estimated frequency that does not satisfy the following condition is discarded: $|\hat{f}_l N/F_s - m_l| \leq 0.5$.

Next, considering that the power spectrum of a sinusoid is the shifted power spectrum of the window, the increase of frequency precision can be used to estimate more precisely the amplitude:

$$\hat{a}_l = 2\frac{|X(m_l)|}{\left|W_H(\hat{f}_l - m_l F_s/N)\right|} \quad (4)$$

where $W_H(f)$ is the frequency response of the Hann window, f being the frequency in Hertz.

The frequency and amplitude estimators of (3) and (4) improve the rough estimates provided by the DFT given in (2). As asserted by comparison methodologies, these estimators are very precise under the idealized assumption of stationarity required by the DFT [18], [20].

However, the problems inherent to the use of the STFT remain in the analysis of polyphonic recordings. The lack of frequency resolution leads to bin contamination if two sinusoidal components have their frequencies too close. In such a case, at least one peak is missing and the second one is corrupted. The lack of temporal resolution and the presence of noise lead to the appearance of noisy peaks that do not correspond to sinusoids. By considering the evolutions of the parameters of the sinusoids over numerous frames, the long-term sinusoidal model described in the next section can be considered to address these issues.

III. Long-Term Sinusoidal Analysis

For stationary pseudoperiodic sounds, the correlation between parameters of peaks of successive frames can be exploited. A "long-term" sinusoidal (LTS) model can be applied, where amplitudes and frequencies parameters continuously evolve slowly with time, controlling a set of pseudosinusoidal oscillators commonly called *partials*. The audio signal s can be calculated from the additive parameters using (5) and (6), where L is the number of partials and the functions F_l, A_l, and Φ_l are the instantaneous frequency, amplitude, and phase of the lth partial, respectively. The L triplets $P_l = (F_l(t), A_l(t), \Phi_l(t))$ are the parameters of the LTS model noted $\mathcal{L}$

$$s(t) = \sum_{l=1}^{L} A_l(t)\cos\left(\Phi_l(t)\right) \quad (5)$$

$$\Phi_l(t) = \Phi_l(0) + 2\pi \int_0^t F_l(u)du. \quad (6)$$

For natural sounds, the continuous functions F_l, A_l, and Φ_l are unknown. Alternatively, a STS representation of the sound is used to identify the values of these functions at discrete time locations. In the considered LTS model, we assume that the control signals corresponding to the evolutions of the frequency and amplitude parameters of the partials are deterministic—thus predictable—and that they are slow time-varying and more precisely inaudible (to avoid modulations, for the perceptive coherence of the model). Thus, we have two constraints on the model parameters: predictability and inaudibility.

Once the partials are carefully extracted from the STS representation so that these two constraints are met, a partial may be represented by a triplet of discrete time signals

$$P_l = \left(F_l(n), A_l(n), \Phi_l(n)\right), \qquad n \in [b_l, d_l] \quad (7)$$

where b_l and d_l denote, respectively, the frame indices of "birth" and "death" of the partial. The continuous functions F_l, A_l, and Φ_l may be approximated from the parameters of the successive peaks of the partial P_l using interpolation schemes presented in [1] and [21].

As discussed above, the STS representation has drawbacks, some peaks may be missing due to the detection of only one peak during the crossing of two underlying sinusoids or the rejection of peaks with incoherent parameters. If a partial is decided to be prolongated at a given frame where no peak issued from the STS frame can be used, a "virtual" peak with interpolated parameters can be used.

A partial can be regarded as a series of successive peaks which may be interpolated or issued from the STS representation:

$$P_l = (p^{b_l}, p^{b_l+1}, \ldots, \hat{p}^{b_l+m}, \ldots, p^{d_l}) \quad (8)$$

where $\hat{p}$ denotes an interpolated peak. The main issue is then to determine which elements of the STS model (the peaks) belong to which elements of the LTS model (the partials). Partial-tracking (PT) algorithms achieve such a task mostly in a

streaming manner. The peaks of $\mathcal{S}_{n+1}$ are used for the continuation of $\mathcal{L}_n$, the set of partials tracked until frame n. This allocation process, detailed in the next section, is repeated iteratively until no untracked frame $\mathcal{S}_m$ remains.

IV. Proposed PT Algorithm

The time evolution of a musical tone can generally be decomposed in three steps: attack, sustain, and release. During the attack that typically lasts for up to 40 ms [22], the partials that compose this tone are small and the evolutions of their parameters are generally chaotic. Therefore, once a partial is born, it stays in a "young" state for about 120 ms and then becomes "mature" until its death. As it will be detailed further, the behavior of these two types of partials will be different in the three steps that compose the proposed PT algorithm.

Given a set of partials tracked until frame n, the first step determines which partial of $\mathcal{L}_n$ should seek for continuation with the highest priority. Partials are sorted in decreasing order according to the s_l criterion, so that the partials having the highest amplitude and the mature—most reliable—ones can select their continuations first

$$s_l = \begin{cases} A_l(n)/K_a, & \text{if } P_l \text{ is mature} \\ -\left|f_i^{n+1} - F_l(n)\right|/K_f, & \text{otherwise} \end{cases} \tag{9}$$

where $A_l(n)$ and $F_l(n)$ are the amplitude and the frequency values of the partial P_l at the last tracked frame indexed n, and K_a and K_f are normalizing constants. The amplitude $A_l(n)$ is always positive so that K_a and K_f can be safely set to 1 in this article, since only the relative order—and not the absolute values—is important here. The frequency f_i^{n+1} is the frequency of a peak verifying the following condition:

$$\left|f_i^{n+1} - F_l(n)\right| \leq \left|f_j^{n+1} - F_l(n)\right|, \qquad \forall j \neq i. \tag{10}$$

Next, the continuation of the partials is searched out in decreasing s_l order. The evolutions of the frequency and amplitude parameters of a partial P_l are predicted and peaks of $\mathcal{S}_{n+1}$ with parameters close to the predicted ones are considered. One of these peaks is selected for the continuation of the partial according to a given relevance criterion. If satisfied, this peak is removed from $\mathcal{S}_{n+1}$ and inserted in P_l.

If the partial is young, the continuation peak is selected according to (10). This peak is effectively used for the continuation of the partial if the absolute frequency difference between this peak and the last inserted peak is below a given threshold Δ_f. An interpolated peak is used otherwise.

Once a partial is mature, a prediction of the frequency evolution of the partial in the next L_T frames is computed using the prediction module described in the next section. Some peaks are chosen in these frames so that the frequency difference between the frequency of the peak and the interpolated one is below a threshold Δ_f, see Fig. 2(a). An original smoothness criterion introduced in Section VI is then used to select one of all possible trajectories that go through peaks of STS frames (dots) or predicted ones (diamonds), see Fig. 2(b). The peak used to prolongate the partial is the first peak of this selected trajectory.

This prediction/selection process is iterated until no untreated partial remains. Finally, the remaining peaks of $\mathcal{S}_{n+1}$ are used to initiate new partials. A young partial is labeled dead if interpolated peaks are successively inserted in a young partial for a given time (around 50 ms in our experiments). For mature partials, a specific maximal time when successive insertion of interpolated peaks are allowed (I_l) is assigned to each mature partial P_l and this time may change from frame to frame. If a peak from an STS frame is inserted, this number is incremented by the duration of the hop size and decremented otherwise. In the experiments, I_l is initialized to 50 ms and cannot exceed 500 ms. If the number of interpolated peaks successively inserted is above I_l, the partial is labeled dead. Those dead partials are removed from the tracking process, and the interpolated peaks lastly inserted in these partials are removed.

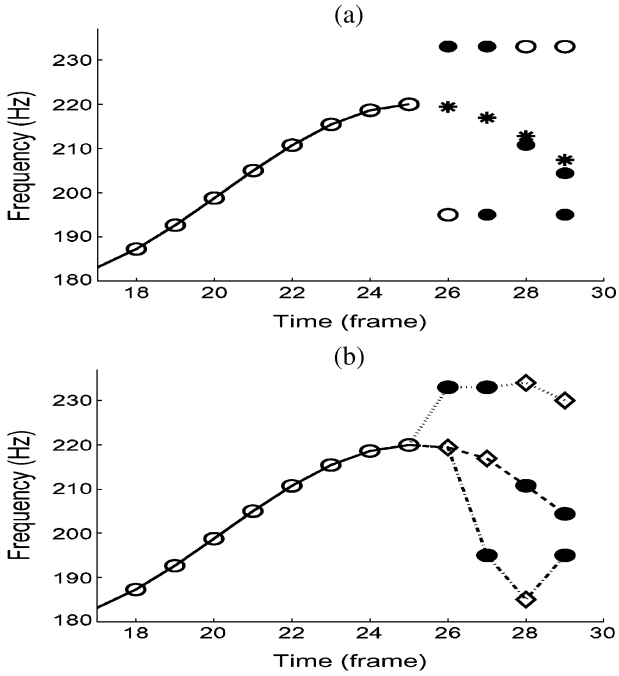

Fig. 2. Selecting peak candidates in the future frames and exploring possible trajectories. On top, the predicted frequencies are plotted with stars. Some STS peaks are chosen so that the frequency difference between the frequency of the peak and the predicted one is below Δ_f. At bottom, possible trajectories that go through these selected measured peaks (dots) and interpolated ones (diamonds) are tested. The first peak of the chosen trajectory is added to the partial.

The prediction and selection steps of mature partials are crucial and will be extensively detailed in the two next sections.

V. Predicting the Evolutions of the Partials

In the McAulay–Quatieri (MQ) algorithm [1], a constant predictor is implicitly used, meaning that the predicted frequency is the frequency of the last inserted peak

$$\hat{F}(n+d) = F(n) \tag{11}$$

where d is the distance between the predicted sample and last observed sample. Assuming that the STS representation is of high quality, we can consider that the best evolution for frequency is the constant one. However, in degraded STS representation as in Fig. 1, a better prediction is crucial to identify the correct continuation of a partial among several noisy peaks.

An improved predictor is used in the HMM algorithm [8], by considering the slope between the two last inserted peaks

$$\hat{F}(n+d) = 2F(n) - F(n-d) \tag{12}$$

where d is often set to 1. We propose to further improve the prediction capability by considering a more complex predictor suitable for the modeling of a wide variety of natural modulations. The evolution of partials in the time/frequency and time/amplitude planes can be constant, exponentially increasing or decreasing (portamento in the time/frequency plane and fade in/out in the time/amplitude plane) or sinusoidal (vibrato in the time/frequency plane and tremolo in the time/amplitude plane).

It is proposed in [23] to model the evolutions of the partials of instrumental sounds of the brass family by means of Kalman filtering using pre-extracted statistical informations. In order to gain generality, we showed in [24] that these evolutions can be modeled by an autoregressive (AR) model. The linear prediction (LP) is then used to predict the evolutions of the parameters of partials in future frames. The current sample $x(n)$ is approximated by a linear combination of past samples of the input signal

$$\hat{x}(n) = \sum_{k=1}^{K} a(k)x(n-k). \tag{13}$$

Given N_{AR} successive past samples considered as observations, the $a(k)$ coefficients are calculated using a specific error-minimization method.

We have shown in [24] that the Burg method must be chosen against other, more commonly used methods such as the autocorrelation or the covariance methods [25], [26]. It only requires $N_{\mathrm{AR}} > 2K$ and the minimum phase property is ensured, leading to stable filters.

The prediction step of the PT algorithm is processed as follows. Given the last N_{AR} frequency samples of a mature partial P_m tracked until frame n, the prediction coefficients $a(k)$ are calculated using the Burg method. The predicted frequencies $\hat{F}_m(n+1), \ldots, \hat{F}_m(n+L_T)$ are then obtained by successive filtering iterations of (13) using the $a(k)$ coefficients. The same process is applied for the computation of the predicted amplitudes $\hat{A}_m(n+1), \ldots, \hat{A}_m(n+L_T)$. Considering that the parameters of the partials are locally stationary, these prediction coefficients are used to estimate the parameters of the interpolated peaks, represented with diamonds in Fig. 2(b). During the experiments presented in this paper, L_T is set to 6.

In order to demonstrate the capability of this predictor and to determine its best parameterization, some experiments were conducted. The testing material is the frequencies of some partials of a saxophone tone with vibrato from the Iowa database [27]. These frequency samples are identified using the STS analysis module described in Section II with a DFT size adapted to the pitch of the tone and tracked correctly using the MQ algorithm.

For each predictor, a given number of past samples of a frequency trajectory are used to predict next samples from the adjacent one ($d = 1$) to the more distant one ($d = 4$). We consider the mean error (expressed in Hertz) obtained by considering the whole frequency trajectory of all the partials of the saxophone tone. The error obtained for each partial is normalized by the harmonic rank of the considered partial prior to summation. The maximal error is also considered because it indicates the robustness of the predictor.

TABLE I
MEAN (AND MAXIMAL) PREDICTION ERRORS OF THE CONSTANT, LINEAR, AND LP PREDICTORS WHILE PREDICTING THE FREQUENCY EVOLUTION OF PARTIALS OF A SAXOPHONE TONE WITH VIBRATO FOR DIFFERENT VALUES OF d (THE DISTANCE IN FRAME INDICES BETWEEN LAST OBSERVATION AND THE PREDICTED VALUE). THE ORDER OF THE LP PREDICTOR GROWS FROM TOP TO BOTTOM AND, FOR EACH SQUARE, THE NUMBER OF SAMPLES CONSIDERED IS 16 AND 32. THE PREDICTION ERRORS OF THE LP PREDICTOR ARE LOWER THAN THOSE OF THE BEST SIMPLE PREDICTOR, AND THE IMPROVEMENT IS GETTING MORE AND MORE SIGNIFICANT WHEN d IS INCREASING

d:	1	2	3	4
Constant	0.35 (0.9)	0.69 (1.8)	1.01 (2.5)	1.31 (3.1)
Linear	0.16 (0.8)	0.42 (1.4)	0.76 (2.4)	1.18 (3.7)
LP order 2	0.16 (0.6) 0.16 (0.7)	0.41 (1.3) 0.41 (1.3)	0.72 (2.2) 0.72 (2.1)	1.09 (3.5) 1.06 (3.3)
LP order 4	0.14 (0.6) 0.13 (0.6)	0.35 (1.3) 0.32 (1.2)	0.62 (2.0) 0.52 (1.7)	0.92 (3.3) 0.77 (2.5)
LP order 6	0.14 (0.6) 0.13 (0.7)	0.34 (1.4) 0.29 (1.1)	0.57 (2.2) 0.46 (1.7)	0.83 (3.1) 0.65 (2.3)
LP order 8	0.14 (0.6) 0.12 (0.6)	0.34 (1.3) 0.28 (1.1)	0.57 (2.1) 0.43 (1.4)	0.81 (3.0) 0.58 (1.9)

The results of these experiments are summarized in Table I. The constant and linear predictors of (11) and (12) are first considered. As far as the mean error is considered, the linear predictor achieves better performance than the constant one when d is small but this improvement decreases when d grows. Considering the maximal error, the robustness improvement is not as significant.

The performance of the LP predictor depends on the choice of a relevant number of observations N_{AR} and model order K. The number of observations should be large enough to extract the signal periodicity, and short enough not to be too constrained by the past evolution. Since we want to handle natural vibrato with a frequency about 4 Hz and the frequency and amplitude trajectories are sampled at ≈86 Hz, we need at least 20 samples to get the period of the vibrato. Since we want to model exponentially increasing or decreasing evolutions (portamento) and sinusoidal evolutions (vibrato), K should not be below 2. In practice, the order should be set at a higher value [24] because observations suffer from imprecision of the estimation of the spectral parameters as shown by the experimental results summarized at the bottom of Table I. In these experiments, the LP predictor is considered with an increasing order, from 2 to 8 and for each square and the number of samples considered are 16 and 32. It shows that if the number of samples is close to 20, the LP predictor reduces the mean error by a factor of 2 and the maximal error by a factor of 1.3 over the constant predictor.

The Burg method considers an error minimization on a finite support, thus the model order should not be greater than half the number of observations. In the experiments reported in the last section of the paper, the prediction of the frequency and amplitude of each partial is done using a maximum of 20 observations if available and a model order $K = 8$. Otherwise, all observations and a model order of half the number of observations are considered.

We have shown that the continuation of modulated partials in the next frames can then be identified more precisely using the

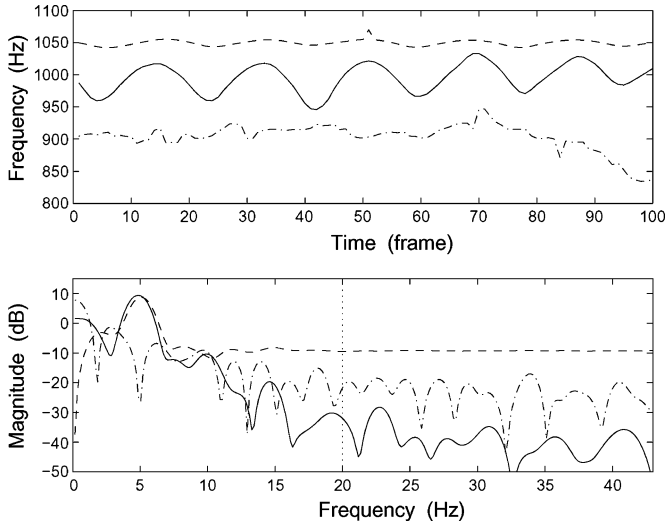

Fig. 3. Three evolutions of the frequency of partials tracked with the MQ algorithm and their corresponding magnitude spectra computed with a Hann-windowed DFT. From top to bottom, a harmonic of a saxophone tone with a local burst around frame 50, a well-tracked harmonic vibrato, and a partial tracked by error from a white noise signal. Only the well-tracked partial has a low HFC.

LP predictor by selecting peaks with parameters close to the predicted ones. The next problem to address is to determine which of the trajectories that go through these peaks is most satisfying the constraints of the LTS model expressed in Section III.

As we proposed in a previous work [14], one can consider the peak whose frequency is the closest to the predicted one and effectively prolongate the partial with this peak if the absolute difference between its frequency and the predicted one is below a given threshold. This simple smoothness criterion was not found satisfying because the higher frequency partials are more modulated than the lowest ones. The threshold should therefore be adaptive which cannot be safely done without any assumption about relationships between partials of the same source such as harmonicity. A more robust selection method is considered in the next section.

VI. Selecting the Continuation of Partials

The definition of the LTS model given in Section III states that the frequency and the amplitude of a partial must evolve slowly with time. From a perceptual point of view, we can consider that these parameters evolve slowly with time if they do not show noticeable energy level in frequency bands upper than 20 Hz. Otherwise, the induced distortion can be heard and the extracted representation becomes no longer relevant because it does not follow perception anymore.

We then propose to study spectral properties of possible continuations of partials to detect if they satisfy the constraints of the LTS model. In a first attempt, slow time-varying evolutions can be discriminated from the others by considering the power of a Hann-windowed DFT spectrum of the evolutions of the frequency of the partials. As shown in Fig. 3, only the well-tracked partials have a high-frequency content (HFC) around -30 dB. Noisy evolutions, local burst, and change of harmonic rank induce a higher HFC.

Such a spectral analysis can only be used at a post processing stage because the number of samples required to compute the DFT with a sufficient frequency resolution is too consequent. Furthermore, the removal of wrong partials after the tracking process may lead to an incomplete sinusoidal representation because the partials with a local discontinuity will be removed erroneously.

Thus, the HFC estimation must be integrated within the tracking process itself to determine whether the use of a given continuation will lead to audible distortions or not. The HFC estimation method should then be as responsive as possible. We use low-delay elliptic infinite response (IIR) high-pass filters to estimate the HFC. The high-pass filtered version of the frequencies of partial P_m is

$$\tilde{F}_m(n) = \sum_{l=0}^{L} d(l)x(l) - \sum_{l=1}^{L} c(l)y(l) \tag{14}$$

where $x(l)$ and $y(l)$ are the memories of the filter and $c(l)$ and $d(l)$ are, respectively, related to the poles and the zeros of the IIR filter. They are mainly determined by the desired cutting frequency and the order of the filters which depend on the frame rate. For frequency and amplitude parameters sampled at $\approx$86 Hz, order-4 filters having normalized cutting frequency of 0.25 are convenient. The following coefficients are used in the experiments:

$$c(0,\ldots,L) = (1, 0.7358, 1.0762, 0.5540, 0.2346)$$
$$d(0,\ldots,L) = (0.06, -0.2274, 0.335, -0.2274, 0.06).$$

At the beginning of a partial, two filters are respectively dedicated to the estimation of the HFC in the evolutions of frequency and amplitude. The memories of these filters $x(l)$ and $y(l)$ are first set to 0 and updated as follows:

$$x(l) = F_m(n-l) - F_m(b) \tag{15}$$
$$y(l) = \tilde{F}_m(n-l) \tag{16}$$

each time a peak p^n is inserted, b being the birth index of the partial. An efficient implementation of this high-pass filter is done using IIR order-2 cells [28].

As can be seen on Fig. 4, the output of the proposed high-pass filter is quite responsive. The insertion of a peak with parameters inducing noticeable HFC in the evolutions of the parameters can be detected very rapidly, with a response delay around two to three samples.

The problem to address now is the definition of a metric that considers the HFC both in the frequency and the amplitude evolutions to determine the best continuation. Considering that a partial is correctly tracked, its frequency and amplitude are slow time-varying, so as the trajectory composed of predicted peaks plotted with stars in Fig. 2(a). Moreover, the frequency or the amplitude of a trajectory made up with peaks of STS frames will have more HFC than the predicted trajectory mostly made of predicted—thus virtual—peaks.

A small HFC difference between these two types of trajectories may be due to measurement imprecision of the STS representation or a smooth change of dynamic. In this case, the noninterpolated trajectory should be used for continuation. On contrary, a larger HFC difference indicates that the noninterpolated trajectory contains spurious peaks or peaks of another partial and should therefore be avoided.

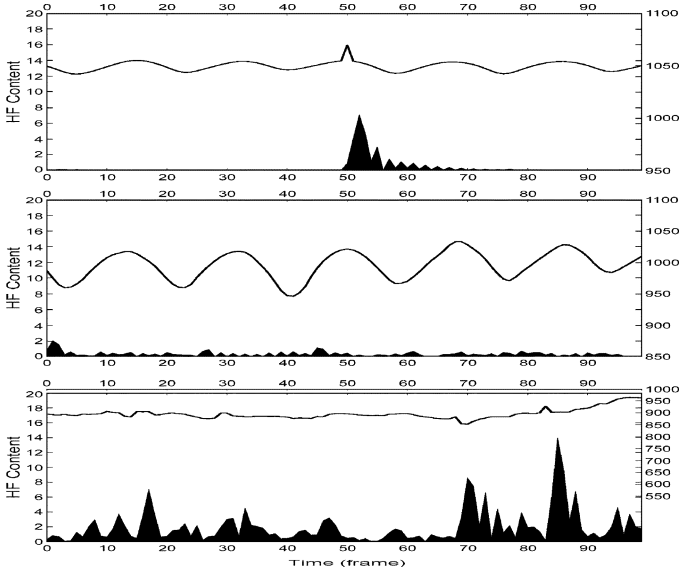

Fig. 4. Output of the high-pass filter (plain) given three different evolutions of the frequency parameter of the partials (line). The discontinuities are noticeable with the output of the high-pass filter with a small delay.

The chosen trajectory should then contain the highest number of peaks of STS frames possible while maintaining a small HFC both in frequency and amplitude. To identify this trajectory, we propose to use an empirically defined cost function associated to each trajectory. This function considers the HFC both in frequency and amplitude and is divided by a factor $\Gamma \in]0,1]$ each time an interpolated peak is used to reflect the constraints cited above:

$$\Pi_T = \left(\frac{1}{\Gamma}\right)^{N_T} \cdot \frac{\sum_{l=1}^{L_T} \left|\tilde{a}_T(l)\right|^2}{K'_a} \cdot \frac{\sum_{l=1}^{L_T} \left|\tilde{f}_T(l)\right|^2}{K'_f} \qquad (17)$$

where $\tilde{a}_T(l)$ and $\tilde{f}_T(l)$ are, respectively, the high-frequency filtered amplitude and frequency of the lth peak of trajectory T of length L_T. This filtering is done using memories of the filters associated to the current partial. N_T is the number of interpolated peaks in the trajectory, K'_a and K'_f are normalizing constants. The choice of the best trajectory leads to constraints on the relative order between costs and not on their absolute values, so that K'_a and K'_f can be safely set to 1 in this paper.

VII. Experiments

To study the properties of the proposed algorithm, we use the following methodology. A signal $s(n)$ is synthesized from a LTS source $\mathcal{L}$ and a perturbation $p(n)$ is added which is either a Gaussian white noise or an artificial LTS source. A STS representation $\hat{\mathcal{S}}$ is extracted using the method described in Section II. The evaluated PT algorithm is used to estimate $\hat{\mathcal{L}}$ from $\hat{\mathcal{S}}$. This LTS representation is then synthesized to obtain $\hat{s}(n)$. The closeness of $\mathcal{L}$ and $\hat{\mathcal{L}}$ is evaluated according to the reconstruction signal-to-noise ratio (R-SNR) versus the degradation signal-to-noise ratio (D-SNR) defined as

$$\text{R}-\text{SNR} = 10\log_{10}\left(\frac{\sum_{n=0}^{N-1} s^2(n)}{\sum_{n=0}^{N-1}\left(s(n)-\hat{s}(n)\right)^2}\right) \qquad (18)$$

$$\text{D}-\text{SNR} = 10\log_{10}\left(\frac{\sum_{n=0}^{N-1} s^2(n)}{\sum_{n=0}^{N-1} p^2(n)}\right). \qquad (19)$$

Three PT tracking algorithms are compared using these metrics. The first is the MQ algorithm with a Δ_f of 80 Hz, used as a reference. The second one, called the LP algorithm [14], only uses the prediction module presented in Section V. Every peak of the next untracked STS frame whose distance between its frequency and the predicted one is below a Δ_f of 40 Hz are selected. The one with the amplitude closer to the predicted one is then chosen for continuing the partial. The third PT algorithm is the proposed tracking method, called the HFC algorithm, with Γ set to 0.9 and trajectory length $L_T = 6$. All the partials shorter than 100 ms are discarded.

The criteria introduced in [15] are used. Let $\mathcal{L}$ be a reference LTS representation and $\hat{\mathcal{L}}$ be the LTS representation computed with the tested PT algorithm from a synthesized version of $\mathcal{L}$. First, $\hat{\mathcal{L}}$ has to be *efficient*, meaning that a partial of $\mathcal{L}$ should be represented with only one partial of $\hat{\mathcal{L}}$. In case of polyphonic recording, $\hat{\mathcal{L}}$ should also be *precise*, meaning that a partial of $\hat{\mathcal{L}}$ represents only one partial of $\mathcal{L}$. Moreover, the PT algorithm should be able to discriminate between deterministic and stochastic components.

In the following experiments, audio inputs of increasing complexity are considered, each modulated by a vibrato since this kind of modulation is a worst-case scenario as far as tracking is concerned.

A. Deterministic/Stochastic Separation

Efficiency and discriminating capabilities of the three algorithms are evaluated using a synthetic constant-amplitude vibrato tone of 2-kHz base frequency, with a vibrato depth and rate of, respectively, 50 and 4 Hz, mixed with a white noise of increasing level.

In a first experiment, to evaluate the efficiency, only the partial having the highest mean amplitude was synthesized to compute the R-SNR. At D-SNR below -7 dB, the MQ algorithm produces partials that are a combination of noisy peaks and tonal peaks so that the tones are split into several partials. The LP method and the HFC method are both able to track correctly the tone with vibrato and thus perform similarly, as can be seen on Fig. 5(a). In the second experiment, to evaluate the discriminating capability of the two algorithms, all retained partials that lay in the [1900, 2100] Hz band are synthesized to compute the R-SNR. As shown in [14], the LP method provides a significant improvement over the MQ method. Compared to the LP method, the HFC method achieves an additional improvement of the same magnitude, see Fig. 5(b).

B. Management of Polyphony

The problem of crossing partials arises when dealing with a mixture of nonstationary sounds. The tracking algorithm has to be able to identify the evolutions of the partials and to interpolate missing spectral data. In order to test the management of crossing, a natural A 440-Hz saxophone tone is corrupted by a synthetic constant-amplitude sinusoid beginning 500 ms later and whose frequency is increasing linearly from 200 Hz to 4 kHz. Only the extracted partials starting before 500 ms

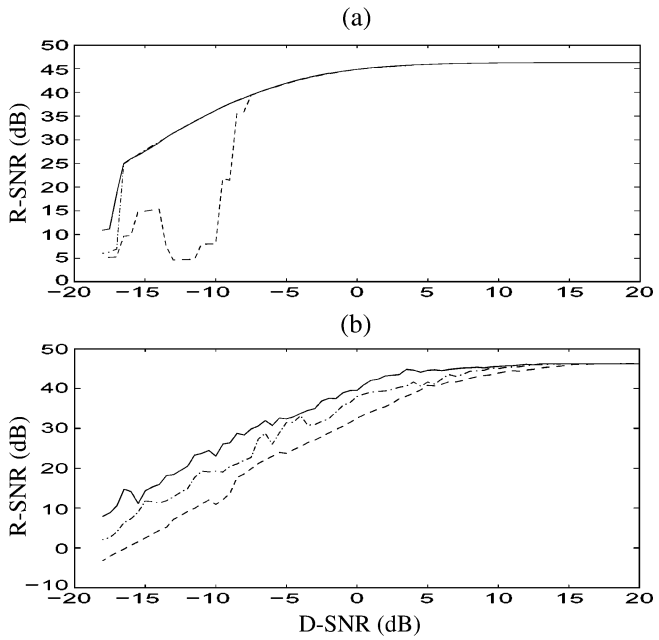

Fig. 5. Evaluation of the efficiency (top) and discrimination capabilities (bottom) of the three methods: MQ method (dashed line), LP method (dotted line), and the HFC method (solid line) using a synthetic vibrato tone embedded in white noise.

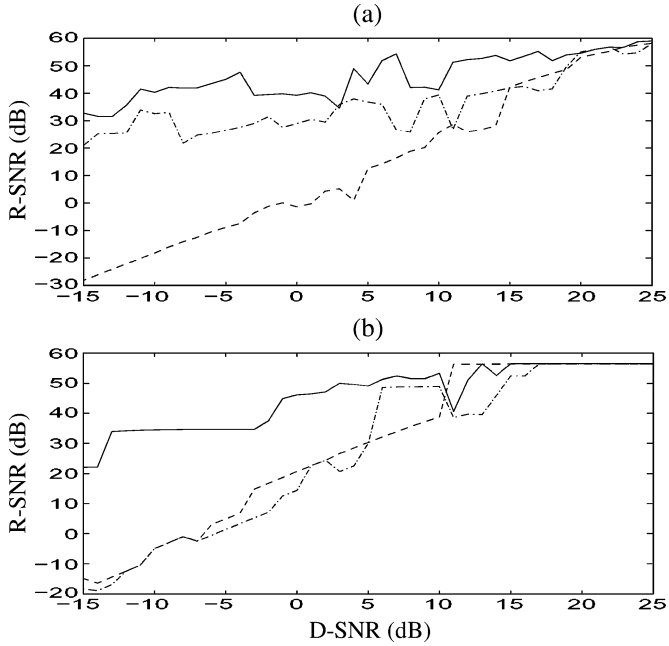

Fig. 6. Evaluation of the closely spaced sinusoids (top) and crossing management (bottom) capabilities of the three methods: the MQ method (dash-dotted line), the LP method (dotted line), and the HFC method (solid line).

were synthesized to compute the R-SNR. Having a model of the evolutions of the parameters leads to an easier management of crossing partials, by being more selective and by having a better interpolation capability. Furthermore, the presented algorithm sorts the partials in decreasing amplitude, so that the partial with the lower degradation is processed first. This reduces the probability of handling the crossing incorrectly, leading to better results as can be seen in Fig. 6(a).

The time/frequency analysis of polyphonic sounds requires a high frequency resolution, but the tradeoff between time and frequency in a musical context leads to the use of analysis windows of reasonable lengths. Pitch relation between harmonic tones leads to DFT bin contamination and closely spaced sinusoids in most natural cases. To evaluate the management of the closely spaced sinusoids, a natural saxophone tone with vibrato is mixed with a set of synthetic constant-frequency and constant-amplitude sinusoids harmonically related, beginning 20 frames later. The fundamental frequency of this synthetic set is the same than the one of the saxophone tone, but all the frequencies within this set have been shifted by 70 Hz towards the low frequencies in order to obtain the same DFT bin contamination for all the harmonics of the original source. Only the extracted partials starting before frame 20 were synthesized to compute the R-SNR. When the synthetic tone begins, the spectral informations are blurred, and some noisy peaks are present between the two close harmonics. The LP method is unable to avoid bad links and performs as the MQ method does, whereas the HFC one performs quite well even at high SNR levels as can be seen on Fig. 6(b).

C. Readability of the LTS Representation

In applications such as indexing or source separation of stationary pseudoperiodic sounds, a good LTS representation should provide a higher level of description, useful to identify sources or to robustly detect informations such as note onset/offset, or pitch. In this experiment, we have chosen as input a three-tone violin sequence. The hissing of the bow leads to many noisy peaks and since the three tones are played *legato*, the transitions can hardly be identified.

In order to robustly detect the note onset/offset, a partial should belong to only one source, and in order to detect the pitch and to identify the sources, the partials should show clear time/frequency and time/amplitude evolutions in order to be able to cluster partials using common variation cues [29]. As can be seen on Fig. 7, the LP method better identifies the vibrato than the MQ method does, but the representation is not satisfying because many partials belong to more than one source. The proposed method (HFC) shows better results in time separation and the vibrato of the second tone is also clearer.

VIII. Evaluation

The experiments presented in the preceding section show the properties of the different algorithms while used in a complete analysis/synthesis chain. In contrast, the evaluation methodology introduced here aims at evaluating PT algorithms solely, i.e., the degradation should be added at the STS level and the evaluation criteria should be defined at the LTS level.

Typical perturbations of the STS representation due to the addition of noise or other sources in the polyphonic case are, respectively, the addition of noisy peaks, the degradation of the precision of the parameters of the peaks, or even the removal of relevant peaks.

From a given STS representation $\mathcal{S}$ of only one partial, such degradations are simulated by, respectively, adding peaks with random parameters, randomizing the parameters of randomly

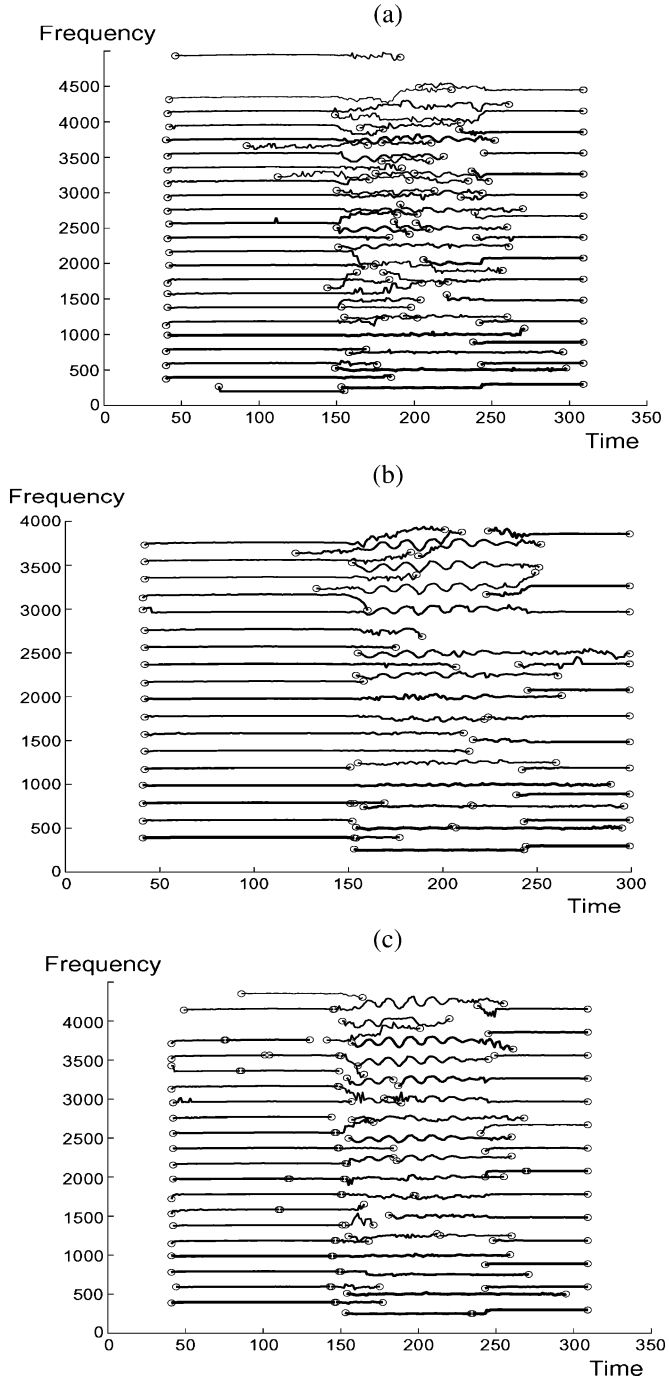

Fig. 7. Partials extracted from three successive violin tones by (a) the MQ method, (b) the LP method, and (c) the HFC one. The partials are represented by solid lines, starting and ending with circles matching the birth and the death of the partials. The proposed method better identifies the vibrato of the second tone as well as the onset/offset of each partial.

selected peaks of $\mathcal{S}$, or removing peaks from $\mathcal{S}$. The strength of the degradation is expressed as the ratio between added, modified, or removed peaks versus the size of $\mathcal{S}$. The randomized parameters are set to be in the same range as those of peaks of $\mathcal{S}$, i.e., the frequency is randomly chosen between the minimal and maximal values of the frequencies in $\mathcal{S}$ set, respectively noted $f_{\min}$ and $f_{\max}$. The amplitude is chosen similarly. From this degraded STS representation, a set of partials $\hat{\mathcal{L}}$ is extracted using a PT algorithm.

TABLE II
PERFORMANCE OF THE FOLLOWING PT ALGORITHMS: MQ (M), LP (L) AND HFC (H) VERSUS INCREASING DEGRADATION

Addition of Peaks									
	Efficiency			Completeness			FP (AP)		
	M	L	H	M	L	H	M	L	H
20 %	56	98	99	99	99	99	98 (98)	99 (99)	99 (99)
50 %	32	51	63	98	99	99	82 (87)	90 (93)	92 (96)
80 %	15	34	35	98	99	99	61 (72)	62 (78)	70 (82)

Randomization of the Frequency Parameter									
	Efficiency			Completeness			FP (AP)		
	M	L	H	M	L	H	M	L	H
20 %	61	63	98	99	99	99	98 (91)	99 (91)	99 (99)
50 %	47	53	51	98	99	99	90 (82)	91 (82)	98 (85)
80 %	28	48	47	98	98	98	88 (79)	89 (80)	92 (83)

Removal of Peaks							
	Efficiency			Completeness			FP (AP)
	M	L	H	M	L	H	M L H
10 %	99	99	99	98	99	99	100 (100)
30 %	53	97	88	92	99	98	100 (100)
40 %	32	82	61	88	96	93	100 (100)

The performance is next evaluated using some criteria defined in the previous section. To evaluate the efficiency, we set the first criterion as the inverse of the number of partials in $\hat{\mathcal{L}}$ or 0 if $\hat{\mathcal{L}}$ is empty. To evaluate the completeness, the second criterion is defined as the number of frames where there is at least an active partial in $\hat{\mathcal{L}}$. The precision is evaluated by means of the frequency and amplitude errors defined as

$$\mathrm{FP} = (f_{\max} - f_{\min}) N_T / \sum_{i=0}^{N_T-1} \sum_{k=1}^{\mathrm{Card}\hat{\mathcal{L}}} \left| F(i) - \hat{F}_k(i) \right| e_k(i)$$

where N_T is the number of frames, $F(i)$ is the frequency of the original partial at frame T_i, and $e_k(i)$ is equal to 1 if the partial k exists at frame T_i and 0 otherwise. The amplitude precision (AP) is defined similarly.

Those criteria are evaluated using a large set of partials extracted using the MQ algorithm from monophonic individual tones of every musical instruments of the Iowa database [27]. The mean results expressed in percentages are presented in Table II. The results show that the use of the LP method provides a significant improvement over the MQ method. Compared to the LP method, the HFC method achieves most of the time an additional improvement of the same magnitude in terms of precision by successfully discarding partials with noisy evolutions.

IX. CONCLUSION

A new partial tracking algorithm dedicated to the analysis of polyphonic sounds has been proposed. The linear prediction of the parameters of the partials is used to select more precisely the continuation of partials and to reliably interpolate this continuation if necessary. Next, a perceptually defined smoothness criterion is used to ensure that the prolongated partial satisfies the slow time-varying constraint of the LTS model. The combination of these two improvements allows the proposed PT algorithm to extract more reliably the pseudoperiodic part of polyphonic sounds.

REFERENCES

[1] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 4, pp. 744–754, Aug. 1986.

[2] H. Van Trees, *Detection, Estimation and Modulation Theory, Part I*. New York: Wiley, 1968, ch. 3.

[3] X. Serra, "Musical sound modeling with sinusoids plus noise," in *Musical Signal Processing*, ser. Studies on New Music Research. Lisse, The Netherlands: Swets & Zeitlinger, 1997, pp. 91–122.

[4] R. C. Maher and J. W. Beauchamp, "Fundamental frequency estimation of musical signals using a two-way mismatch procedure," *J. Acoust. Soc. Amer.*, vol. 95, no. 4, pp. 2254–2263, 1994.

[5] T. F. Quatieri and R. G. Danisewicz, "An approach to co-channel talker interference suppression using a sinusoidal model for speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 1, pp. 56–69, Jan. 1990.

[6] R. C. Maher, "Evaluation of a method for separating digitized duet signals," *J. Audio Eng. Soc.*, vol. 38, no. 12, pp. 956–979, 1990.

[7] H. Viste and G. Evangelista, "A method for separation of overlapping partials based on similarity of temporal envelopes in multichannel mixtures," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 1051–1061, 2006.

[8] P. Depalle, G. Garcia, and X. Rodet, "Tracking of partials for additive sound synthesis using hidden Markov model," in *IEEE ICASSP*, Apr. 1993, vol. 1, pp. 225–228.

[9] M. Lagrange, S. Marchand, and J.-B. Rault, "Partial tracking based on future trajectories exploration," in *116th Convention Audio Eng. Soc.*, Berlin, Germany, May 2004, Audio Eng. Soc. (AES), Preprint 6046 (10 pp.).

[10] H. Purnhagen and N. Meine, "HILN-The MPEG-4 parametric audio coding tools," in *IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2000, vol. 3, pp. 201–204.

[11] B. den Brinker, E. Schuijers, and W. Oomen, "Parametric coding for high-quality audio," in *112th Convention Audio Eng. Soc.*, May 2002, Audio Eng. Soc. (AES), preprint 5554.

[12] P. Fernandez-Cid and J. Casajus-Quiros, "Multi-pitch estimation for polyphonic musical signals," in *IEEE ICASSP*, Apr. 1998, pp. 3565–3568.

[13] M. Lagrange and S. Marchand, "Long interpolation of audio signals using linear prediction in sinusoidal modeling," *J. Audio Eng. Soc.*, vol. 53, no. 10, pp. 892–905, Oct. 2005.

[14] M. Lagrange, S. Marchand, and J.-B. Rault, "Using linear prediction to enhance the tracking of partials," in *IEEE ICASSP*, May 2004, vol. 4, pp. 241–244.

[15] M. Lagrange, S. Marchand, and J.-B. Rault, "Tracking partials for the sinusoidal modeling of polyphonic sounds," in *IEEE ICASSP*, Mar. 2005, vol. 3, pp. 229–232.

[16] T. F. Quatieri and R. J. McAulay, "Audio signal processing based on sinusoidal analysis/synthesis," in *Applications of DSP to Audio and Acoustics*. Norwell, MA: Kluwer, 1998, pp. 343–416.

[17] M. Desainte-Catherine and S. Marchand, "High precision Fourier analysis of sounds using signal derivatives," *J. Audio Eng. Soc.*, vol. 48, no. 7/8, pp. 654–667, Jul./Aug. 2000.

[18] F. Keiler and S. Marchand, "Survey on extraction of sinusoids in stationary sounds," in *Proc. Digital Audio Effects (DAFx) Conf.*, Hamburg, Germany, Sep. 2002, pp. 51–58, Univ. Federal Armed Forces.

[19] S. Marchand, "Sound models for computer music (analysis, transformation, synthesis)," Ph.D. dissertation, Univ. Bordeaux 1, LaBRI, Talence, France, Dec. 2000.

[20] S. Marchand and M. Lagrange, "On the equivalence of phase-based methods for the estimation of instantaneous frequency," in *Proc. 14th Eur. Conf. Signal Process. EURASIP*, 2006.

[21] L. Girin, S. Marchand, J. di Martino, A. Röbel, and G. Peeters, "Comparing the order of a polynomial phase model for the synthesis of quasi-harmonic audio signals," in *IEEE WASPAA*, New Paltz, NY, Oct. 2003.

[22] J. W. Gordon, "Perception of attack transients in muscial tones," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1984.

[23] A. Sterian and G. H. Wakefield, "A model-based approach to partial tracking for musical transcription," in *Proc. SPIE*, San Diego, CA, 1998, vol. 3461, pp. 171–182.

[24] M. Lagrange, S. Marchand, M. Raspaud, and J.-B. Rault, "Enhanced partial tracking using linear prediction," in *Proc. Digital Audio Effects (DAFx) Conf.*, Queen Mary, Univ. London, London, U.K., Sep. 2003, pp. 141–146, .

[25] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Nov. 1975.

[26] S. M. Kay, "Autoregressive spectral estimation: Methods," in *Modern Spectral Estimation*, ser. Signal Processing Series. Englewood Cliffs, NJ: Prentice-Hall, 1988, pp. 228–231.

[27] The Iowa Music Instrument Samples. [Online]. Available: http://theremin.music.uiowa.edu

[28] L. B. Jackson, *Digital Filters and Signal Processing*. Boston, MA: Kluwer, 1996, ch. 11.

[29] M. Lagrange, "A new dissimilarity metric for the clustering of partials using the common variation cue," in *Proc. Int. Comput. Music Conf. (ICMC)*, Barcelona, Spain, Sep. 2005, Int. Comput. Music Assoc. (ICMA).

Mathieu Lagrange was born in Caen, France, in 1978. He received the M.Sc. degree in computer science from the University of Rennes 1, Rennes, France, in 2000. He obtained a post-graduate diploma focusing on spectral sound synthesis at the University of Bordeaux 1, Talence, France. He carried out research on sound analysis and coding at France Telecom Laboratories in partnership with the LaBRI (Computer Science Laboratory) where he received the Ph.D. degree in 2004.

He is currently a Research Assistant with the Computer Science Department, University of Victoria, Victoria, BC, Canada. His research focuses on structured modeling of audio signals applied to the indexing, browsing, and retrieval of multimedia.

Sylvain Marchand (M'07) was born in Pessac, France, in 1972. He received the M.Sc. degree in algorithmics and the Ph.D. degree, while carrying out research in computer music and sound modeling, from the University of Bordeaux 1, Talence, France, in 1996 and 2000, respectively.

He was appointed Associate Professor at the LaBRI (Computer Science Laboratory), University of Bordeaux 1, in 2001. He is particularly involved in spectral sound analysis, transformation, and synthesis. He is a member of Studio de Création et de Recherche en Informatique et Musique Electroacoustique (SCRIME), University of Bordeaux 1.

Jean-Bernard Rault received the Ph.D. degree in signal processing and telecommunications from the University of Rennes, Rennes, France, in 1987.

He joined the CCETT, Rennes, to collaborate in the European project Eureka 147 (DAB) in the area of digital audio compression. From 1990 to 1992, he spent two years at Thomson-LER, where he was involved in multicarrier digital modulation studies. Form 1993 to 2004, he was a France Telecom representative at ISO/MPEG and, as such, participated in the development of the MPEG Audio coding standards (cf. mp3). He was also involved in several European projects (MoMuSys, Cinenet, Nadib, Song, Ardor) to contribute in audio-related work packages. Since mid-2005, he has been focusing mainly on audio segmentation and identification.