TRACKING PARTIALS FOR THE SINUSOIDAL MODELING OF POLYPHONIC SOUNDS

Mathieu Lagrange†*, Sylvain Marchand*†*, and Jean-Bernard Rault*‡

†SCRIME – LaBRI, Université Bordeaux 1
351, cours de la Libération,
F-33405 Talence cedex, France
firstname.name@labri.fr

‡France Telecom R&D
4, rue du Clos Courtel, BP 59
F-35512 Cesson Sevigné cedex, France
jeanbernard.rault@rd.francetelecom.com

ABSTRACT

This paper proposes to further improve the tracking of partials in a polyphonic context. Spectral characteristics of the controlling parameters (amplitude and frequency) are taken into account to ensure that these parameters evolve slowly with time. The resulting algorithm better tracks closely-spaced sinusoids and is able to avoid most of the spectral data belonging to noise. As a consequence, the proposed algorithm extracts a more meaningful sinusoidal representation from polyphonic recordings.

1. INTRODUCTION

The sinusoidal model presented in Section 2 provides a high-quality representation of pseudo-stationary sounds. Therefore, this model is widely used for many musical audio processing purposes such as musical source separation, transcription or coding. One of the most challenging part of the analysis chain is known as partial tracking. In [1], we proposed to enhance the partial tracking algorithm proposed in [2] by means of Linear Prediction (LP). This leads to better performance for modulated sounds such as notes with vibrato or tremolo. Furthermore, the interpolation capability of this algorithm is useful for interpolating missing parts of partials in case of crossing sinusoids. In this article, we propose to further improve this algorithm by considering the spectral properties of the evolutions of the controlling parameters of the partials in order to find the best continuation for each partial trajectory.

After a short introduction about sinusoidal modeling, the constraint on the evolutions of the parameters of the partials that will be exploited in this article is presented in Section 2. This constraint leads to a criterion presented in Section 3 to be used during the tracking process in an algorithm described in Section 4. The performances of this new partial-tracking algorithm are finally described in Section 5.

2. SINUSOIDAL MODEL

The sinusoidal model proposes to represent sounds as sums of sinusoids – so-called partials – controlled by parameters that evolve slowly with time. The audio signal s can be calculated from the controlling parameters using Equations 1 and 2, where P is the number of partials and the functions f_p, a_p, and ϕ_p are the instantaneous frequency, amplitude, and phase of the p-th partial, respectively. The P pairs (f_p, a_p) – so-called peaks – are the parameters of the additive model and represent points in the frequency-amplitude plane at time t.

$$s(t) = \sum_{p=1}^{P} a_p(t)\ \cos(\phi_p(t)) \quad (1)$$

$$\phi_p(t) = \phi_p(0) + 2\pi \int_0^t f_p(u)\ du \quad (2)$$

Very low bit-rate audio coding [3] can be achieved using sinusoidal modeling since the parameters controlling the oscillators are slow time-varying and thus can be encoded very efficiently. More recently, this model has been used for musical transcription [4] and source separation [5].

Partial tracking aims at selecting peaks and linking them together from frame to frame in order to form partials. The first tracking algorithm was proposed by McAulay and Quatieri in [2] for the sinusoidal modeling of monophonic speech. This algorithm is based on the assumption that partials composing a voiced signal have stationary frequency evolutions. Frequency differences between peaks of immediate successive frames are considered in order to form partials. A maximal frequency difference threshold Δ_f is set between the frequency of the last inserted peak of the partial and the frequency of a peak candidate for linking.

The performance of the algorithm relies on an unambiguous short-term sinusoidal representation: the peaks identified must belong to sinusoidal components and a peak must be identified every time a sinusoidal component is present. During the analysis of polyphonic sounds, the time / frequency trade-off leads to an ambiguous sinusoidal short-term representation as explained in [1]: some peaks belong to stochastic components and some peaks are missing.

An algorithm proposed in [6] enhances the previous algorithm by considering dynamic programming. A cost function based on amplitude and frequency deviations between the two last inserted peaks in the partial and the peak candidate is globally (for all partials) minimized to determine the linking of all peaks at a frame. This algorithm is very innovative because it allows to better avoid noisy peaks and handles crossing of sinusoids. Unfortunately, the global minimization is time-consuming and the problem of missing peaks is not handled.

We propose to further improve the avoidance of noisy peaks by considering an original cost function based on the frequency analysis of the evolution of the parameters of the partials. The parameters of the partials should not have a stochastic behavior (indicating that the partial models some stochastic component), nor having abrupt changes (indicating that the partial models some transient component or jumps across several sinusoidal components) such as those plotted with dashed / dotted lines on top of Figure 1, because these two behaviors do not satisfy the slow time-varying constraint.

The challenge is then to define a sinusoidal analysis procedure that is able to extract partials having parameters that are slow time-varying and to reject the others. A first approach is to restrict the range of possible variations as in [2]. Yet, there can be heavily modulated evolutions that are slow time-varying. Alternatively, we propose to study spectral properties of the evolutions of the parameters of the partials in order to decide if they are slow time-varying or not. Indeed, if the evolutions of the partials show noticeable energy levels in frequencies bands upper than 20 Hz, then the induced distortion can be heard. The extracted representation

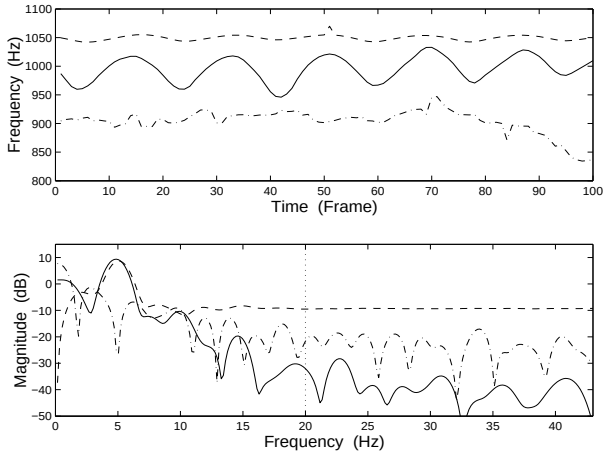

Fig. 1. Three evolutions of the frequency of partials extracted using a nearest frequency neighbor approach (on top) and their corresponding spectra (at bottom). From top to bottom, an harmonic of a saxophone tone with a synthetic local burst around frame 50 (dashed line), a well-tracked harmonic with vibrato (solid line), and a partial wrongly extracted from a white noise signal (dash-dotted line). The spectral estimates of the two unwanted evolutions contain significant energy in the high frequencies.

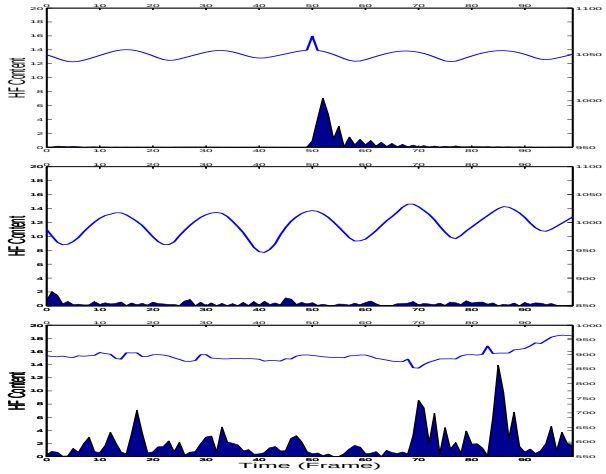

Fig. 2. Output of the high-pass filter (plain) given three different evolutions of the frequency parameter of the partials (line).

becomes no longer relevant because it does not follow perception anymore.

3. HIGH-FREQUENCY CONTENT ESTIMATION

As shown at bottom of Figure 1, it is possible to discriminate slow time-varying evolutions from the others (noisy partials, local burst in the evolution, or change of harmonics rank) by considering spectral estimates of the evolutions. Therefore, it is possible to identify partials that belong to the model by considering the high frequency content (HFC) of the evolutions of their frequencies and their amplitudes. Unfortunately, removing wrong partials after the tracking process may lead to an incomplete sinusoidal representation. The "noisy" partials will be removed but also the partials with a local discontinuity.

To extract partials that conform to the model, the HFC estimation must be integrated within the tracking process itself. To decide whether a peak candidate should be the continuation of a partial, the HFC induced by the insertion of this peak is estimated. Several spectral methods for the estimation of the HFC have been tested, and the use of low-delay elliptic high-pass filters gave the most relevant results. Parameters such as the cutting frequency and the order of the filters depend on the frame rate. For frequency and amplitude parameters sampled at ≈ 86 Hz, order-4 filters having normalized cutting frequency of 0.5 are convenient. An efficient implementation is done using IIR order-2 cells with the following coefficients:

$$\begin{pmatrix} 1 & 0.2274 & 0 & 1 & -0.2346 & 0 \\ 1 & 0.1673 & -0.0137 & 1 & -1.2898 & 0.4076 \\ 1 & -0.3951 & 0.0201 & 1 & -2.0762 & 1.0762 \end{pmatrix}$$

As can be seen on Figure 2, the output of the high-pass filter is quite responsive so that the insertion of peak with parameters inducing noticeable HFC in the evolutions of the parameters can be detected very rapidly. In practice, each time a peak is inserted, the amplitude and the frequency parameters of this peak (minus the amplitude / frequency of the first inserted peak) are used to update memories of the two filters dedicated to the estimation of the HFC in the evolutions of amplitude / frequency parameters of the partial. While exploring possible continuations, these memories are used as-is.

4. PROPOSED ALGORITHM

Since the filters used for HFC estimation require at least a few observed samples to be effective, a peak selection process that considers HFC cannot be used from the very beginning of the partial. Therefore, partials can be in two states during the tracking process: the "young" state and the "mature" one, whether the number of inserted peaks is below or upper a given threshold N_s. If the partial is young, the selection is similar to the one proposed in [2]. It selects the peak having the frequency closest to the frequency of the last inserted peak so that the difference between these two frequencies is below a Δ_f threshold. If the partial is mature, the estimation of the HFC can be exploited, the selection strategy described in the next section is used. Once each partial has selected its best continuation in the next frame, the partials are sorted using a criterion presented in the second part of this section, so that the most reliable partials can extend themselves first in the next frame.

4.1. Exploring Possible Trajectories

Once a partial is mature, the HFC of the evolutions of the frequency and the amplitude of possible trajectories in future frames can be exploited. Small trajectories in futures frames of length $N_f > 1$ are considered for two reasons. First, the high-pass filters used for HFC estimation have a response delay. Second, considering several frames in the future is valuable to avoid local discontinuities.

A prediction of the frequency evolution of the partial in the next frames is computed using Linear Prediction (LP), see Figure 3.1 and [1] for further details. Two peaks per frame are chosen so that the frequency difference between the frequency of the peak and the interpolated one is below Δ_f, see Figure 3.2. Considering local stationarity, the LP coefficients used for the prediction are then used to compute the parameters of the interpolated peaks. All possibles trajectories that go through the measured peaks (dots) or predicted ones (diamonds) are tested.

The chosen trajectory should then contain the highest number of extracted peaks possible while maintaining a small HFC in both frequency and amplitude. To each trajectory is associated a cost function that considers the HFC both in frequency and amplitude. Additionally, the cost function is divided by a factor $\Gamma \in]0, 1]$ each time an interpolated peak is used:

$$\pi_t = \left(\frac{1}{\Gamma}\right)^{n_t} \cdot \frac{\sum_{i=0}^{N_f} |\tilde{a}_i^k|^2}{K_a} \cdot \frac{\sum_{i=0}^{N_f} |\tilde{f}_i^k|^2}{K_f} \quad (3)$$

where $\tilde{a}_i^k$ and $\tilde{f}_i^k$ are, respectively, the high-frequency filtered amplitude and frequency of the k-th peak in the frame i. 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 number t, 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 article. The selected peak is then the first peak of the trajectory with the smallest π_t.

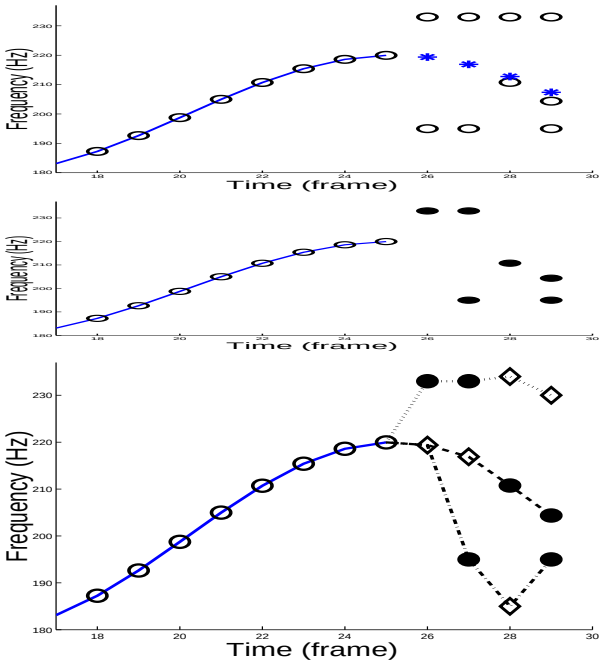

Fig. 3. Selecting peak candidates in the future frames and exploring possible trajectories. On top, the predicted frequencies using linear prediction are plotted with stars. Two peaks per frame are chosen so that the frequency difference between the frequency of the peak and the predicted one is below Δ_f (plotted in the middle). At bottom, possible trajectories that go through these selected measured peaks (dots) and interpolated ones (diamonds) are tested. One is chosen according to a cost function π_t that penalizes HFC both in frequency and amplitude and the use of interpolated peaks (see Equation 3). The first peak of this trajectory is added to the partial.

4.2. Partials Prioritization

Since a peak extracted from the Short-Time Fourier Transform (STFT) should be allocated to only one partial, concurrency between partials must be taken into account. Once the partials have selected their best continuation, they are sorted in decreasing order, according to the s_p criterion defined in Equation 4, so that the partials having the highest amplitude and the mature – most reliable – ones can select their next peak first. Once the partials are scheduled, each partial extends itself using its selected peak if this peak is still available.

$$s_p = \begin{cases} a_l & \text{if } p \text{ is mature} \\ \frac{-|f'-f_l|}{\Delta_f} & \text{otherwise} \end{cases} \quad (4)$$

Where a_l and f_l are the amplitude and the frequency of the last inserted peak, and f' is the frequency of the peak selected for continuation.

5. RESULTS

This section compares three tracking algorithms: the first is the McAulay-Quatieri (MAQ) algorithm with a Δ_f of 80 Hz. The second one is presented in [1]. All peaks whose distance between its frequency and the predicted one is below a Δ_f of 40 Hz are selected and the one with the amplitude closer to the predicted one is chosen for continuing the partial. The third one is the proposed tracking method with Γ set to 0.9, N_f to 6 and $N_s = 20$.

For the three methods, only 4 successive interpolated peaks are allowed and all the partials having less than 10 extracted peaks are discarded. Concerning the STFT, the frame size is 2048 samples and the hop size is 360 samples at a 44100-Hz sampling frequency. The synthesis algorithm used is the one described in [2], based on the linear interpolation of the amplitude and the maximally-smooth cubic interpolation of the phase.

5.1. Deterministic / Stochastic Separation

Efficiency and discriminating capabilities of the two 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. The degradation is evaluated with the Degradation SNR (D-SNR) defined as the noise energy to original signal energy ratio. The quality of the tracking algorithm is then measured with the Reconstruction SNR (R-SNR), defined as the error (original signal - synthesized signal) energy to the original signal energy ratio. In the 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 MAQ algorithm produces partials that are a mix of noisy peaks and tonal peaks so that the tones are split into several partials. The LP method and the proposed method are both able to track correctly the tone with vibrato and thus perform similarly. 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 [1], the LP method provides a significant improvement over the MAQ method. Compared to the LP method, the HFC method achieves an additional improvement of the same magnitude.

5.2. Management of Polyphony

The problem of crossing partials arises when dealing with a mixture of non-stationary 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 20 frames later and whose frequency is increasing linearly from 200 Hz to 4 kHz. Only the extracted partials starting before frame 20 were synthesized to

D-SNR (dB)	-15	0	15	-15	0	15
MAQ	-30	0	20	-15	15	55
LP	20	25	35	-15	10	50
HFC	30	40	50	20	40	55

Table 1. Performances of the three tracking algorithms evaluated with the R-SNR for the crossing sinusoids test (left) and the closely-spaced sinusoids test (right).

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 algorithms sort the partials in decreasing amplitude, so that the partial with the lower degradation is processed first. It reduces the probability of handling the crossing incorrectly, see left part of Table 1.

The time / frequency analysis of polyphonic sounds requires a high frequency resolution, but the trade-off between time and frequency leads to the use of analysis windows of reasonable length. Pitch relation between harmonic notes leads to FFT 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 FFT 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.

The right part of Table 1 shows the advantages of the analysis of the HFC of possible evolutions in several future frames. 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 MAQ method does, whereas the proposed one performs quite good even at high SNR levels.

5.3. Readability of the Sinusoidal Representation

In applications such as indexing or source separation of stationary pseudo-periodic sounds, a good partial representation should provides a higher level of description, useful to detect robustly high-level informations such as note onset / offset, pitch detection and source identification.

In order to easily detect the note onset / offset, one would like to have a good time separation, meaning that 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. As explained in [1], the LP method better identifies the vibrato than the MAQ method does, but the representation is not precise because many partials belong to more than one source. As can be seen on Figure 4, the proposed method shows better results in time separation and the vibrato of the second tone is also clearer.

6. CONCLUSION

In this article, we propose to replace heuristics in frequency distance by the analysis of the high frequency contents of possible evolutions of partials. This new approach further improves the quality of the sinusoidal representation of polyphonic recordings. In particular, onsets and offsets of partials are better identified and close sinusoids can be better tracked, two important requirements for source identification and sources separation based on sinusoidal modeling.

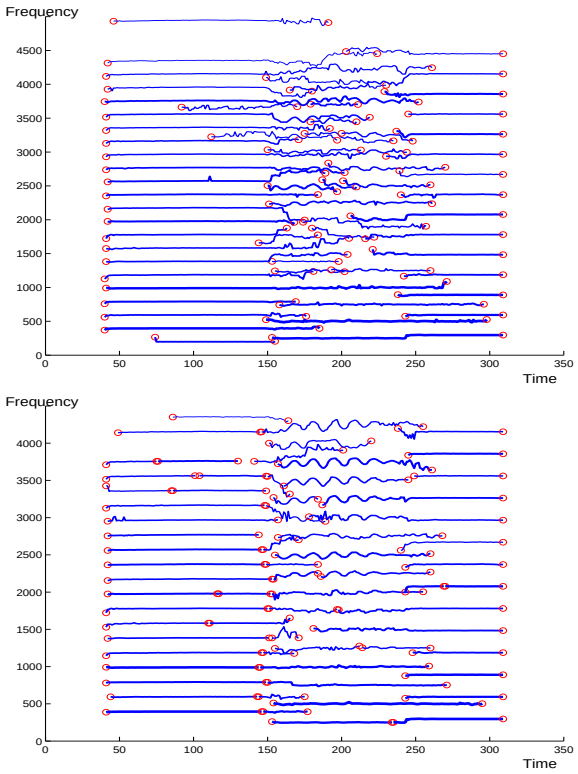

Fig. 4. Partials extracted from three successive violin tones by the MAQ method (top) and the proposed one (bottom) – see [1] for a comparison with the LP method. The partials are represented by solid lines, starting and ending with circles matching the birth and the death of the partials.

7. REFERENCES

[1] Mathieu Lagrange, Sylvain Marchand, and Jean-Bernard Rault, "Using Linear Prediction to Enhance the Tracking of Partials," in *IEEE ICASSP*, may 2004, vol. 4, pp. 241–244.

[2] Robert J. McAulay and Thomas F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[3] Bert den Brinker, Erik Schuijers, and Werner Oomen, "Parametric Coding for High-Quality Audio," in *112th Convention of the AES*. Audio Engineering Society (AES), May 2002.

[4] Pablo Fernandez-Cid and Javier Casajus-Quiros, "Multi-Pitch Estimation for Polyphonic Musical Signals," in *IEEE ICASSP*, April 1998, pp. 3565–3568.

[5] Tuomas Virtanen and Anssi Klapuri, "Separation of Harmonic Sound Sources Using Sinusoidal Modeling," in *IEEE ICASSP*, April 2000, vol. 2, pp. 765–768.

[6] Philippe Depalle, Guillermo Garcia, and Xavier Rodet, "Tracking of Partials for Additive Sound Synthesis Using Hidden Markov Models," in *IEEE ICASSP*, April 1993, vol. 1, pp. 225–228.