Using Linear Prediction to Enhance the Tracking of Partials

M. Lagrange, Sylvain Marchand, J.-B. Rault

USING LINEAR PREDICTION TO ENHANCE THE TRACKING OF PARTIALS

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

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

‡SCRIME – LaBRI, Université Bordeaux 1
351, cours de la Libération,
F-33405 Talence cedex, France
sylvain.marchand@labri.fr

ABSTRACT

In this article, we present an enhanced algorithm, of low complexity, for the tracking of partials in the context of sinusoidal modeling. By considering the past evolution of each partial in the time/frequency and time/amplitude planes to predict its future evolutions, this algorithm allows a better discrimination between sinusoidal and noisy components and an easier cancellation of sudden changes in the evolutions of the partials.

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 parts of the analysis chain presented in Figure 1 is known as partial tracking.

Given a discrete time/frequency representation, a partial tracker should be able to extract continuous informations by linking frequency components of successive frames. To achieve such a task, most tracking methods [1, 2, 3] use heuristics such as the distance in frequency, amplitude, and phase between two successive frequency components. While such heuristics are successful with monophonic sounds, they generally fail when extended to polyphonic sounds because of spectral degradations "blurring" the time/frequency representation. In such cases, a model for the evolutions of the partials is clearly needed. Each partial should have an evolution which is slow time-varying and predictable, and for specific purposes the model can be more rigorous [4]. In Section 3 we present a tracking method that exploits these constraints. It is based on the well-known linear prediction model presented in Section 4, but this time applied to the spectral parameters of the sounds. Its implementation is detailed in Section 5. Discussion about the problem of the validation and comparison of the different tracking methods, as well as results, follow in Section 6.

2. SINUSOIDAL MODELING

Additive synthesis is the original spectrum modeling technique. 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. For stationary pseudo-periodic sounds, these amplitudes and frequencies continuously evolve slowly with time, controlling a set of pseudo-sinusoidal oscillators commonly called *partials*. The audio signal s can be calculated from the additive 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) are the parameters of the additive model and represent points in the frequency/amplitude plane at time t. This representation is used in many analysis/synthesis programs such as Lemur [5], SMS [3], or InSpect [6].

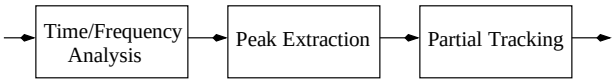

Fig. 1. Sinusoidal analysis procedure.

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

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

2.1. Analysis Procedure

Sinusoidal analysis is generally made of three steps as shown in Figure 1. Instantaneous parameters of partials are estimated by picking some local maxima commonly called *peaks* from a short-term time/frequency analysis. Partials are then formed by tracking peaks over time, from frame to frame.

In our system, spectral estimation is done using a short-time Fourier transform. The instantaneous frequency, amplitude, and phase of the peaks are estimated using the derivative method [7]. For various reasons, some peaks can be missing or erroneous. The peak-picking process can remove some erroneous ones [8], but is unable to recover missing spectral information. In order to form partials of reasonable length, the partial-tracking step should be able to interpolate missing peaks.

2.2. Constraints on the Evolutions of Partial Parameters

We assume that the partials have a minimum length and that their evolutions in frequency and amplitude are predictable, since repeated "birth"/"death" and sudden changes in the evolutions of the partials will generate noisy "clicks" at the resynthesis stage and offer a poor representation of the spectral contents.

3. ALGORITHM OVERVIEW

The first partial tracking algorithm was introduced in [1] by McAulay and Quatieri in the field of the sinusoidal modeling of the voice. The algorithm is based on the assumption that the partials composing a voiced signal have stationary frequencies. Given a set of partials ending at frame k, it is proposed to consider frequency differences below a given threshold Δ_f between the last inserted peak of each partial and the peaks of frames $k+n_i+1$. The n_i

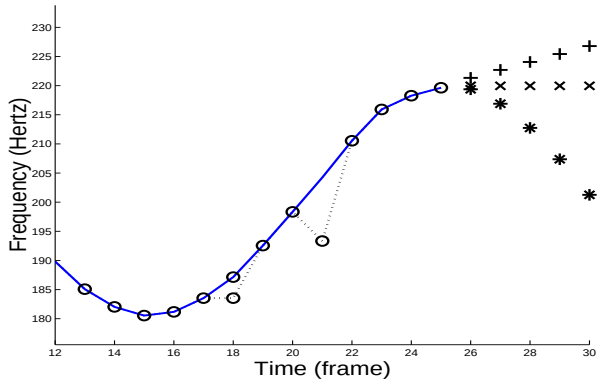

Fig. 2. Prediction capability of constant (x), linear (+) and LP (*) predictors on a synthetic vibrato. Peaks are represented with circles (o). The already-tracked partial using the MAQ algorithm and the proposed method are plotted, respectively, with dotted and solid lines.

parameter is the number of interpolated peaks allowed to be used at once. If no peak satisfying this constraint can be found the partial is "dead" and removed from the tracking process. If a peak remains unlinked at frame $k+1$, a partial is "born". Considering $n_i = 0$ gives the original McAulay and Quatieri (MAQ) algorithm.

To select the next peak, the frequency and amplitude evolutions of the partials are considered as constant (see Figure 2). Yet, the frequencies and amplitudes of the partials are not stationary, but their evolutions are often predictable. A better approximation has been proposed in [2] by considering the continuity of the slope. However, without a model for the evolutions of the partials it seems impossible to overcome spectral artifacts. It is proposed in [4] 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 [9] that linear prediction (LP) can be used to model and predict the frequency and amplitude evolutions with parameters computed from the past evolutions of the partials. Therefore, we propose to use both the predicted frequency and amplitude to select peak candidates in the next frames and to interpolate missing peaks.

Using linear prediction has two advantages that can be seen in Figure 2. The predicted frequency is closest to the one of the next peak to be linked, so that the algorithm is more precise (we now choose the right peak at frame 18). Since the prediction error is lower than with using a constant predictor, the algorithm can be much more selective by using a lower Δ_f. We can then discard the peak at frame 21 and use an interpolated one instead.

4. LINEAR PREDICTION

In the linear prediction (LP) model, also known as the autoregressive (AR) model, the current sample $x(n)$ is approximated by a linear combination of k past samples of the input signal. We are then looking for a vector a of k coefficients, k being the order of the LP model. Provided that the a vector is estimated, the predicted value $\hat{x}$ is computed simply by FIR filtering of the k past samples with the coefficients using Equation 3:

$$\hat{x}(n) = \sum_{i=1}^{k} a_i\, x(n-i) \tag{3}$$

The challenge with linear prediction modeling is to choose the model order k, the number of samples and type of method to estimate coefficients that suit specific needs.

For frequency and amplitude evolutions, since we want to be able to model exponentially increasing or decreasing evolutions (portamento) and sinusoidal evolutions (vibrato), the order of the LP model should not be below 2. Experimental testing showed that a model order in the $[2,8]$ range is suitable.

The number of samples used has to be large enough to be able to extract the signal periodicity, and short enough not to be too constrained by the past evolution. The short-term analysis module uses a sliding time/frequency transform with a hop size of 512 samples on sound signals sampled at CD quality (44.1 kHz). This means that the frequency and amplitude trajectories are sampled at ≈ 86 Hz. Since we want to handle natural vibrato with a frequency about 4 Hz, we need at least 20 samples to get the period of the vibrato. Experimental testing showed that for most cases a number of samples in the $[4,32]$ range is suitable.

On one hand spectral data suffer from imprecision, so the method has to be resistant to noise. On the other hand, the evolutions of the partials in frequency and amplitude are sampled at a low sampling rate, so the method to estimate the coefficients has to be reactive. Among the three methods tested in [9], the Burg method was the most satisfactory. It minimizes the average of the power of the forward e^f and backward e^b errors calculated using Equations 4 and 5, it then leads to stable filter coefficients (see [10] for details). Moreover, the minimization is done on a finite support and the method requires only a few samples.

$$e_k^f(n) = \frac{1}{(N-k)} \sum_{n=k}^{N-1} |x(n) + \sum_{i=1}^{k} a(i)x(n-i)|^2 \tag{4}$$

$$e_k^b(n) = \frac{1}{(N-k)} \sum_{n=0}^{N-1-k} |x(n-k) + \sum_{i=1}^{k} a(i)x(n+k+i)|^2 \tag{5}$$

5. ENHANCED ALGORITHM

Since linear prediction requires at least a few samples to be effective, all starting partials then begin using a tracking mode similar to the one used in the MAQ algorithm for a fixed number of frames n_s. During these first frames, these partials can be considered as "young" (the others being called "mature"). We propose to process the partials in decreasing amplitude order and the mature – most reliable – ones are processed first in order to reduce clicks and thus improve resynthesis quality.

At each frame, each mature partial selects and links a peak that is close to the prediction, if any. Defining a distance between peaks that combines frequency and amplitude is not straightforward because these two parameters do not have the same physical dimensions. Instead, we propose a two-step selection algorithm. First, the partial selects all peak candidates in the next frame whose frequency distance to the prediction is below a given threshold Δ_f. Second, it keeps from these peaks the one whose amplitude is closest to the prediction.

If no peak can be found, the predicted frequency and amplitude are used to create a new peak. Its phase is then interpolated using the maximally smooth cubic interpolation detailed in [1]. In case of musical sound modulations, corrupted tremolo and vibrato are well resynthesized.

If two partials cross, a peak corruption occurs in the crossing region (see Figure 3). In this case, interpolated peaks will be used for the partial having the lower amplitude since there is no peak satisfying the prediction constraints. When this partial links to

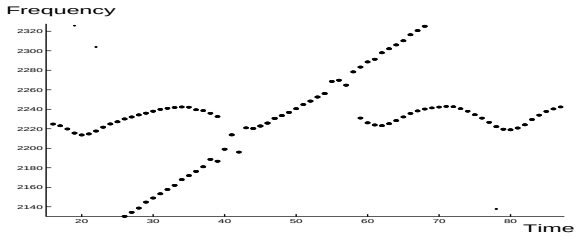

Fig. 3. Peak extraction of the crossing of an harmonic of a saxophone tone and a synthetic frequency-ramped sinusoid sound.

an existing peak, the crossing is detected (at frame 58 in Figure 3). If the amplitude difference between the two partials is below a given threshold, the partial with the highest amplitude will also use interpolated peaks in the crossing region because the extracted peaks in this region are considered as too corrupted to be reliable.

6. RESULTS

Due to the numerous possible applications of a partial tracking module, it seems hard to give an exhaustive evaluation of partial-tracking algorithms by a straightforward quality measure. As a consequence, we left this task for further research and give an intuitive comparison of the MAQ algorithm and the proposed method using several criteria.

We can roughly set two requirements concerning the results of a tracking algorithm: the resynthesis quality concerning coding applications and the identification of onset/offset and evolutions of partials concerning indexing applications. Both requirements are needed for source separation. The resynthesis quality will be discussed in the first part using a synthetic tone. Crossing partials management is studied in a second part. The identification of onset/offset and evolutions of partials is then studied in a third part using natural violin tones.

6.1. Resynthesis Quality

In the processing chains of hybrid sound models such as sinusoids-+noise or sinusoids+transients+noise, the sinusoidal components are first extracted and then synthesized to be subtracted from the original signal to give a residual that will be considered as a random process (noise). Ideally, the partials should efficiently represent all sinusoidal components and only them. Otherwise, if the partial set represents only a few of the sinusoidal components, there will be some sinusoidal components left in the residual. If the partial-tracking algorithm has no discrimination capability, it will wrongly identify partials in noise and the random process will be modeled as slow-time varying sinusoids, forcing the residual module to handle those synthesized sinusoids that were not present in the original sound.

Efficiency and discrimination capabilities of the two algorithms are tested using a synthetic constant-amplitude 4-Hz vibrato sinusoid with frequency ranges from 1950 to 2050 Hz, embedded in a white noise of growing level. Concerning the parameterization of the tracking algorithms, n_i was set to 4. For the MAQ algorithms, we used the smallest Δ_f such that the vibrato could be tracked, that is 20 Hz. For the proposed method, Δ_f was set to 12 Hz, the prediction order to 6, the prediction length to 20 frames, and n_s to 10. All partials whose length was below 15 peaks were discarded. The peak extraction method proposed in [7] and the original partial

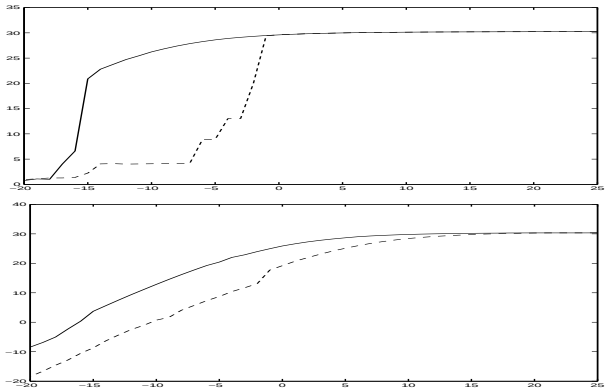

Fig. 4. Evaluation of the efficiency and discrimination capabilities of the two methods: proposed method (solid line) and MAQ method (dashed line) using a synthetic vibrato tone embedded in white noise. The plot shows reconstruction-SNR versus degradation-SNR on the dB scale. To evaluate the efficiency, on the top, only the partial with the highest amplitude is synthesized. At the bottom, to evaluate the discrimination capability, all extracted partials are synthesized to compute the reconstruction-SNR.

resynthesis method introduced in [1] were used for the two compared methods. The quality of the tracking algorithm is measured with the reconstruction-SNR in function of the degradation-SNR.

In the first experiment, in order to evaluate the efficiency, only the partial having the highest mean amplitude was synthesized to compute the reconstruction-SNR. At degradation-SNR below -7 dB (see Figure 4.1), the MAQ algorithm produces partials that are a mix of noisy and tonal peaks so that the tone is split into several partials. Whereas the proposed method, with its smaller Δ_f, is able to track correctly the tone with vibrato by not choosing noisy peaks. The slow decay is due to errors in the estimation of the spectral peaks. In the second experiment, to evaluate the discriminating capability of the two algorithms, all retained partials having frequencies in the $[1900, 2100]$ Hz band are synthesized to compute the degradation-SNR (see Figure 4.2). Because the proposed algorithm allows the use of a lower Δ_f, it has a better discrimination capability.

6.2. Crossing of Partials

The problem of crossing partials arises when we have to deal with a mixture of non-stationary sounds. The tracking has to be able to identify the evolutions of the partials and to interpolate correctly missing spectral data (see Figure 3). In order to test the management of crossing, a natural A-440 Hz saxophone tone was 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 compute the reconstruction-SNR. Results for the MAQ method using a Δ_f of 80 Hz and the proposed method using a Δ_f of 25 Hz are plotted in Figure 5. 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 schedules 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.

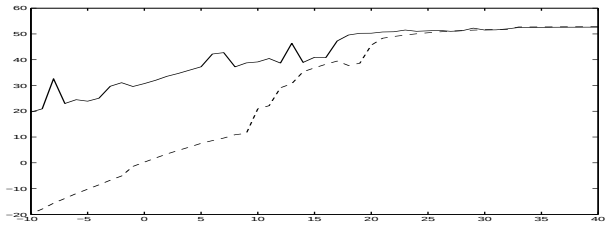

Fig. 5. Evaluation of the crossing management capability of the two methods: the proposed method (solid line) and the MAQ algorithm (dashed line). A natural A-440 Hz saxophone tone is corrupted by a synthetic constant-amplitude sinusoid beginning 20 frames later and whose frequency is increasing linearly from 200 Hz to 4 kHz. The plot shows reconstruction-SNR versus degradation-SNR on the dB scale. Only the extracted partials starting before frame 20 were synthesized to compute the reconstruction-SNR.

6.3. Identification of Partials

In applications such as indexing or source separation of stationary pseudo-periodic sounds, a good representation of the partials provides a higher level of description, useful to extract 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 can be seen in Figure 6.1, the partial set extracted by the MAQ algorithm is not satisfactory. A lot of partials belong to two or three tones and it would be very difficult to detect the vibrato frequency of the second tone. The proposed method shows better results in time separation and the vibrato of the second tone is clearer (see Figure 6.2). However the algorithm described in Section 5 is not perfect, since some partials belong to two tones and the vibrato is still not perfectly handled in noisy conditions.

7. CONCLUSION

In this article, we proposed to use linear prediction to replace the classic stationary assumption to better track and interpolate partials in the context of sinusoidal modeling. While the presented implementation has still to be compared with tracking algorithm explicitly dedicated to musical recording analysis [2, 4], this new approach leads to promising results in terms of resynthesis and representation quality, two important requirements for an approach of source separation from mono recording based on sinusoidal modeling.

8. REFERENCES

[1] 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.

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

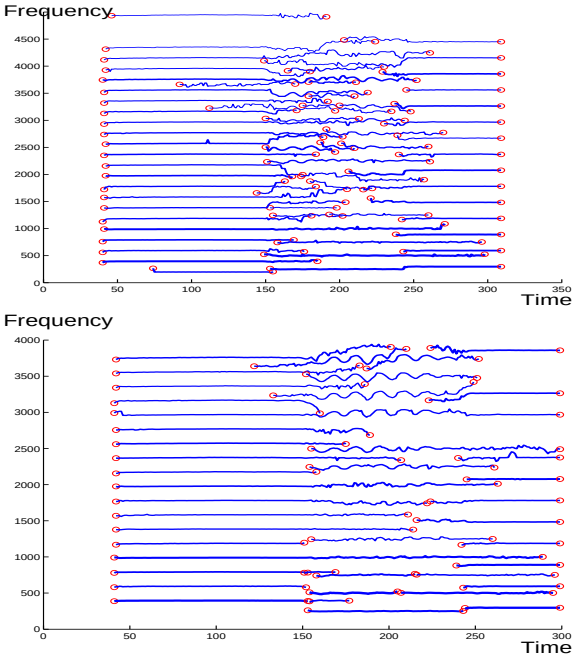

Fig. 6. Partials extracted from three successive violin tones (the second one with vibrato) by the MAQ algorithm (top) and the proposed method (bottom). The partials are represented by solid lines, starting and ending with circles matching their birth and death.

[3] Xavier Serra, *Musical Signal Processing*, chapter Musical Sound Modeling with Sinusoids plus Noise, pp. 91–122, Studies on New Music Research. Swets & Zeitlinger, Lisse, the Netherlands, 1997.

[4] Andrew Sterian and Gregory H. Wakefield, "A model-based approach to partial tracking for musical transcription," SPIE annual meeting, San Diego, California, 1998.

[5] Kelly Fitz and Lippold Haken, "Sinusoidal Modeling and Manipulation Using Lemur," *Computer Music Journal*, vol. 20, no. 4, pp. 44–59, Winter 1996.

[6] Sylvain Marchand and Robert Strandh, "InSpect and ReSpect: Spectral Modeling, Analysis and Real-Time Synthesis Software Tools for Researchers and Composers," in *Proc. ICMC*, Beijing, China, October 1999, ICMA, pp. 341–344.

[7] Myriam Desainte-Catherine and Sylvain Marchand, "High Precision Fourier Analysis of Sounds Using Signal Derivatives," *JAES*, vol. 48, no. 7/8, pp. 654–667, July/August 2000.

[8] Mathieu Lagrange, Sylvain Marchand, and Jean-Bernard Rault, "Sinusoidal Parameter Extraction and Component Selection in a Non Stationary Model," in *Proc. DAFx*, September 2002, pp. 59–64.

[9] Mathieu Lagrange, Sylvain Marchand, Martin Raspaud, and Jean-Bernard Rault, "Enhanced Partial Tracking using Linear Prediction," in *Proc. DAFx*, September 2003, pp. 141–146.

[10] Steven M. Kay, *Modern Spectral Estimation*, chapter Autoregressive Spectral Estimation : Methods, pp. 228–231, Signal Processing Series. Prentice Hall, 1988.