RECURSIVE LEAST-SQUARES ESTIMATION OF THE EVOLUTION OF PARTIALS IN SINUSOIDAL ANALYSIS

Leonardo de O. Nunes, Ricardo Merched, Luiz W. P. Biscainho

Universidade Federal do Rio de Janeiro
LPS - DEL/Poli & PEE/COPPE
Caixa Postal 68504 - 21941-972 Rio de Janeiro, RJ - Brasil

ABSTRACT

Classic methods for sinusoidal analysis rely on partial tracking, a technique where successive sets of spectral peaks of an audio signal must be properly associated in time. The resulting tracks describe, in terms of amplitude and frequency, the continuous evolution of the so-called partials which, combined, model the complex sounds emitted by a given instrument. A well-known challenge in this context is preserving amplitude and frequency coherence in the tracking mechanism, specially in cases where failure in peak detection may occur, or perhaps in the event of crossing partials. This paper presents a new decision-directed recursive least-squares (RLS) estimation method for frequency and amplitude tracking in sinusoidal analysis. Different performance measurements show that the proposed deterministic algorithm outperforms some procedures currently found in the literature.

***Index Terms*—** Sinusoidal modelling, Partial tracking, Linear prediction, Adaptive filtering

1. INTRODUCTION

Audio signals are usually termed musical for their inherent *tonal* characteristics. Loosely speaking, one could label *tonal* those sounds exhibiting an identifiable pitch (in the sense of perceived frequency), to which a musical note could possibly be assigned. For such well-behaved sounds, it is possible to assume a certain degree of time-periodicity, so that the resulting spectrum of music signals tends to present spectral peaks at the frequencies of each individual signal component. Obviously, in practice we are limited to a short-time description of sinusoidal partials, which gives rise to an approach usually referred to as sinusoidal modelling [1].

Sinusoidal analysis of audio signals can be linked to a variety of applications, such as automatic transcription of music, music information retrieval, audio compression, or recording restoration. Often associated to a synthesis procedure, the sinusoidal analysis comprises not only detection, but also identification and organization of spectral peaks along the time axis. As a consequence, the tonal part of the signal can be described by a set of continuous spectral lines, so-called tracks. Each line is in turn represented by a vector sequence, consisting of amplitude-frequency pairs obtained from time-contiguous spectral peaks within a determined time frame.

Besides handling errors in the detection of spectrum maxima, an efficient tracking mechanism must preserve coherence even in exceptional scenarios, as for instance, the ones characterized by *tremoli*, *vibrati*, *glissandi*, or the not uncommon occurrence of crossing lines. Moreover, a robust method for partial tracking normally includes a sequence of heuristics in order to tackle these issues [2], which can be further improved via standard linear prediction techniques [3, 4]. Other alternatives include the use of HMM [5].

The present work describes a new decision-directed recursive least-squares (RLS) method for estimating the successive amplitude-frequency pairs, under both decoupled and joint estimation scenarios. Based on different performance measurements, it is verified that the proposed deterministic algorithm can outperform both the classical procedure based on the McAulay & Quartieri (MQ) algorithm as well as a recent stochastic approach based on the Burg method.

The authors wish to thank CNPq and FAPERJ for supporting this work.

2. SINUSOIDAL MODELLING

Sinusoidal modelling describes a signal $x(n)$ as a sum of amplitude- and (phase- or) frequency-modulated sinusoids:

$$x(t) = \sum_{l=1}^{L} A_l(t) \sin \Phi_l(t), \tag{1}$$

$$\Phi_l(t) = \Phi_l(0) + \int_0^t \omega_l(u) du. \tag{2}$$

In this description, the continuous nature of the amplitude $A_l(t)$ and angular frequency $\omega_l(t)$ leads to a computationally intractable problem. In order to simplify the analysis, (1) is commonly replaced by a discrete model

$$x[n] = \sum_{l=1}^{L} A_l[n] \sin \Phi_l[n], \tag{3}$$

which can be further considered short-time stationary in amplitude and frequency. That is, for a given partial l, given that $A_l[n]$ and $\Phi_l[n]$ are low-pass narrow-bandwidth time series, one assumes that $A_l[n] \approx A_l$ and $\Phi_l[n] \approx \Omega_l n + \Phi_l[0]$, where A_l and Ω_l are constant values, during a time interval of N samples.

Figure 1 illustrates a simplified block diagram of a typical partial tracking based sinusoidal analysis system [2].

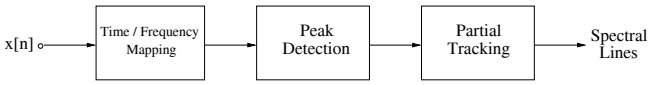

Fig. 1. The three steps of a sinusoidal analysis system.

The 'time / frequency mapping' performs a short-time Fourier transform (STFT) [6] analysis on the audio signal at predefined hops $H \leq N$. The spectrum of each time-windowed frame is then submitted to the 'peak detection' step, which identifies a set of local

maxima, each of them possibly corresponding to a stationary sinusoidal component of the signal [2]. Finally, a 'partial tracking' procedure is responsible for validating the extracted peaks which build, frame by frame, the model spectral lines.

2.1. The Fundamental Partial Tracking Algorithm

The most widely known algorithm for partial tracking is the so-called McAulay & Quatieri (MQ) algorithm, originally proposed in a speech analysis context [7] and independently developed for audio analysis [8].

The partial tracking algorithm is responsible for identifying the moments when individual partials emerge or vanish, as well as finding their best continuation. Of course, the primary criterium for track continuation is the frequency proximity of two peaks in consecutive frames. Since the number of detected peaks may vary from frame to frame, the underlying algorithm must be prepared to solve occasional conflicts. To this end, the MQ algorithm is further refined and extended in [8].

The term 'track' normally refers to a set of frequency, amplitude and, if necessary, phase information of time-contiguous, supposedly related spectral peaks. In the MQ algorithm, every peak is always associated to a track, which in turn can be cast into one of the following statuses: *emerging* and *evolving* (active tracks), or *vanishing* (inactive track).

Now, to each track i in frame k with corresponding peak frequency $f_i[k]$, the algorithm assigns the closest peak j, detected at frequency $f_j[k+1]$), such that $|f_j[k+1]-f_i[k]| \leq \Delta f$. If two tracks dispute the same peak, the closest one wins the dispute, whereas the losing track searches for the next closest peak. If a peak is not assigned to any pre-existing track, an associated *emerging* track is created. After E frames in the *emerging* status, a track changes to *evolving*, otherwise it is discarded. If a track does not find any peak in frame $k+1$, it is labelled as *vanishing*, and its current magnitude-frequency pair is copied to the next frame. If the track remains in this status during a sequence of S frames, then it is considered inactive, and hence extinguished from the frame it entered the vanishing status.

Note that a proper setup for the parameters $\{\Delta f, E, S\}$ is crucial for the performance of the tracking algorithm. The Δf parameter controls the maximum frequency variation allowed, and is usually frequency dependent; $\Delta f = 0.03 f_i(k)$ is a common choice, since it corresponds to a quarter-tone. The E parameter is responsible for removing short tracks, in case these are formed by wrongly identified peaks. On the other hand, the S parameter avoids track discontinuation as a consequence of missing peaks.

Recent results towards improving partial tracking capability have been reported in [4]. In the latter, a stochastic linear prediction method based on the Burg algorithm has shown to be notably useful, specially in situations where crossing spectral lines appear. We now propose a deterministic linear estimation procedure for partial tracking.

3. ADAPTIVE-FILTER SOLUTION

Adaptive filtering [9] is usually performed by a digital filter structure whose coefficients can be adjusted along the time according to some optimization criterium. This approach applies to any situation when the filtering action must adapt itself to a time-variant environment. A typical adaptive system can be seen in Figure 2.

The error signal $e[n]$ is the difference between the output signal $y[n] = \sum_{m=0}^{M} w_m[n]x[n-m]$ (for an FIR filter of order M) and

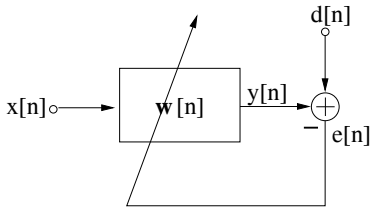

Fig. 2. A general adaptive filtering system.

a given desired signal $d[n]$. As the input signal $x[n]$ evolves, the filter coefficients $w_m[n]$ are sequentially updated in order to minimize some error-based function. In particular, choosing $d[n] = x[n+1]$ turns the system into a predictor which estimates the next signal sample by $\hat{x}[k+1] = y[k]$.

An important design concern is the choice of the optimization algorithm, which ultimately dictates solution biasing, convergence speed etc. The most common algorithms can be roughly grouped in two families: the stochastic least-squares, e.g. the Burg algorithm; and the exact deterministic least-squares, e.g. the recursive least-squares (RLS) algorithm.

3.1. Track Predictor & Builder

In this section, we propose a regularized recursive least-squares procedure for predicting the amplitude and frequency of each partial on a frame-by-frame basis. In this context, two prediction schemes are devised: a) a single predictor which jointly estimates the amplitude and frequency of the next peak; b) two independent predictors, for amplitude and frequency, respectively.

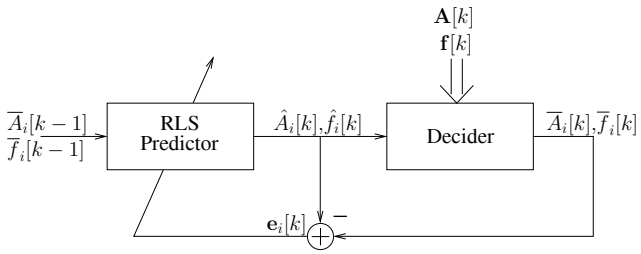

Fig. 3. Proposed prediction scheme for track i at frame k.

The proposed scheme can be seen in Figure 3. For a given track i, predicted values of magnitude ($\hat{A}_i[k]$) and frequency ($\hat{f}_i[k]$) help to choose the best track continuation, once the most prominent spectral peaks of the signal at frame k (stored in magnitude-vector $\mathbf{A}[k]$ and frequency-vector $\mathbf{f}[k]$) have been detected. A decision heuristics, to be defined further ahead, selects magnitude $\overline{A}_i[k]$ and frequency $\overline{f}_i[k]$ as valid values. These decisions are used as inputs to a J-th order predictor, which produces the best linear estimates $\hat{A}_i[k+1]$ and $\hat{f}_i[k+1]$ for frame $k+1$, and so on.

Defining the output vector $\mathbf{y}_i[k] = \left[\begin{array}{cc} \hat{A}_i[k+1] & \hat{f}_i[k+1] \end{array}\right]$ and the input vector $\mathbf{x}_i[k] = \left[\begin{array}{cc} \overline{\mathbf{A}}_i^T[k] & \overline{\mathbf{f}}_i^T[k] \end{array}\right]$, with

$$\overline{\mathbf{A}}_i[k] = [\overline{A}_i[k]\ \overline{A}_i[k-1]\ \cdots\ \overline{A}_i[k-(J-1)]]^T \quad (4)$$

$$\overline{\mathbf{f}}_i[k] = [\overline{f}_i[k]\ \overline{f}_i[k-1]\ \cdots\ \overline{f}_i[k-(J-1)]]^T, \quad (5)$$

one can write

$$\mathbf{y}_i[k] = \mathbf{x}_i[x]\mathbf{W}_i[k], \quad (6)$$

where $\mathbf{W}_i[k]$ is a $2J \times 2$ coefficient-matrix.

Given a judicious choice of $\alpha > 0$ and a forgetting factor $0 << \lambda \leq 1$, the exponentially-weighted regularized least-squares problem [10] seeks the matrix $\mathbf{W}_i[k]$ that minimizes

$$\lambda^{k+1}\mathbf{W}_i^T[k]\mathbf{\Pi}_J^{-1}\mathbf{W}_i[k] + \sum_{l=0}^{k}\lambda^{k-l}\left\|\mathbf{d}_i[l] - \mathbf{x}_i[l]\mathbf{W}_i[k]\right\|^2, \quad (7)$$

where $\mathbf{d}_i[k] = \left[\begin{array}{cc}\overline{A}_i[k+1] & \overline{f}_i[k+1]\end{array}\right]$ is the desired-signal vector, and $\mathbf{\Pi}_J^{-1} = \alpha^{-1}\mathbf{I}_J$.

The solution at frame k can be computed via the following recursions:

$$\begin{aligned}
\gamma_i[k] &= (1 + \lambda^{-1}\mathbf{x}_i[k]\mathbf{P}_i[k-1]\mathbf{x}_i^T[k])^{-1} && (8)\\
\mathbf{g}_i[k] &= \lambda^{-1}\mathbf{P}_i[k-1]\mathbf{x}_i^T[k]\gamma_i[k] && (9)\\
\mathbf{e}_i'[k] &= \mathbf{d}_i[k] - \mathbf{x}_i[k]\mathbf{W}_i[k-1] && (10)\\
\mathbf{W}_i[k] &= \mathbf{W}_i[k-1] + \mathbf{g}_i[k]\mathbf{e}_i'[k] && (11)\\
\mathbf{P}_i[k] &= \lambda^{-1}\mathbf{P}_i[k-1] - \mathbf{g}_i[k]\mathbf{g}_i^T[k]\gamma_i^{-1}[k] && (12)
\end{aligned}$$

The coefficient-matrix can be explicitly defined as

$$\mathbf{W}_i[k] = \left(\begin{array}{cc}\mathbf{w}_{i,AA} & \mathbf{w}_{i,fA}\\ \mathbf{w}_{i,Af} & \mathbf{w}_{i,ff}\end{array}\right), \quad (13)$$

where each vector $\mathbf{w}_{i,bc}$ conveys the effect of c over the prediction of b, each one given by either magnitude or frequency. This scheme considers both magnitude and frequency information into the prediction of each other. This is expected to work out in improving the performance of the predictor, considering that both parameters are ultimately controlled at the same time by the music performer.

However, depending on the type of sound source, or even on the noise contamination level, the signal amplitude may behave more unpredictably than the corresponding frequency, thus impairing the estimation of the latter. For these cases, an alternative uncoupled structure can be straightforwardly obtained by simplifying the structure of the solution to

$$\mathbf{W}_i[k] = \left(\begin{array}{cc}\mathbf{w}_{i,AA} & \mathbf{0}\\ \mathbf{0} & \mathbf{w}_{i,ff}\end{array}\right). \quad (14)$$

Observe that no cross-information between frequency and amplitude evolution of the partial is shared. As a consequence, the orders J_A and J_f of magnitude and of frequency estimators can be made different, if necessary.

3.2. Decision Heuristics

As it was mentioned earlier, a decision must be made on which amplitude-frequency pair is the most likely to belong to a track. This can be accomplished quite similarly to the strategy described in Section 2.1 and the method proposed in [4]. First, a group of candidate peaks is formed by the elements of $\mathbf{f}[k]$ and $\mathbf{A}[k]$, such that $|\hat{f}_i[k] - \mathbf{f}[k]| \leq \Delta f$. Then, the candidates are sorted according to a mixed metric consisting of a linear combination of amplitude and frequency square errors, given by the relative differences between their predicted and detected values. The peak showing the smallest distance to its predicted counterpart is then chosen as the one to continue the track. When a peak is not assigned to any track, a new track is created. During the first qJ frames, $q \geq 1$, the predicted results are simply discarded, while the samples are shifted in and the filter is trained. The treatment of these emerging as well as vanishing tracks follows the same lines described in Section 2.1.

In the next section, the performance of the proposed strategy is assessed under real conditions.

Table 1. Comparison of re-synthesized signals under PEAQ.

	MQ	Burg	Coupled RLS	Unc. RLS
PEAQ	-2.415	-1.910	-1.208	-0.975

4. COMPUTER SIMULATIONS

Assessing the performance of a partial tracking algorithm under practical circumstances is not trivial, since it constitutes part of a more general analysis system. A typical setup is used throughout this section: after segmenting the signal using a 1023-point Hann window, a time/frequency mapping is performed via 4096-point DFTs computed at 512-sample hops; frequency reassignment [11] is used to refine the instantaneous frequency estimation; finally, a peak detection stage employs a variable-threshold strategy.

The following experiments are based on real audio signals, and aim to compare the two versions of the proposed algorithm against the classical partial tracking method of Section 2.1 and the method based on the Burg predictor, described in [3]. The Burg estimator [9] minimizes a stochastic objective function composed by the sum of the mean-square forward- and backward- prediction errors. It can be alternatively implemented as a structurally minimum-phase lattice-form adaptive filter. Its overall complexity is $\mathcal{O}(J)$. It should be noticed that, although the conventional implementation of the RLS algorithm is $\mathcal{O}(J^2)$, its fast versions [9] can match the Burg complexity.

In the first simulation set, the music excerpt to be analyzed is the recording of a long violin *vibrato*, at a sample rate of 44.1 kHz. All methods employed $S = 20$, $V = 5$, and $\Delta f = 5$. For the predictor-based methods we have set $J = 4$, except for the uncoupled RLS, which used $J_A = 2$ and $J_F = 4$. These values follow approximately previous works' setups. Both RLS versions employed $\Pi = 2000$ and $\lambda = 0.98$, empirically chosen. Frequency prediction for every track is performed relatively to its first measured frequency value. In Fig. 4, four plots of a short segment of the signal allow one to compare the performance among the four methods. The MQ method (Fig. 4a) often misses track continuity. The Burg method (Fig. 4b) appears to be more robust when dealing with poorly detected peaks, but occasionally makes wrong decisions which severely disturb the tracks' behavior. The proposed RLS method (Fig. 4c) shows the most consistent performance along the signal. In particular, the joint predictor (Fig. 4d) turned out to be the only one capable of tracking some spectral lines.

In order to further evaluate the performance of the proposed method, a second set of experiments has been envisioned: a signal containing crossing harmonics, consisting of a clarinet *glissando* mixed with a violin *vibrato*, both extracted from real recordings, was formed. This test signal was analyzed under the four approaches above, using the same parameters of the first experiment. It was then re-synthesized from each sinusoidal representation and compared with its original version, this time making use of the so-called PEAQ (Perceptual Evaluation of Audio Quality) measure, defined by the ITU-R [12]. This measure has been originally conceived for evaluating high-quality coded signals, and should be judiciously employed in other contexts. Even so, in the present work it can provide some additional elements to compare the performance of the methods under analysis. The grades, ranging from 0 (for a perceptually identical signal) to -4 (the worst admissible grade), assigned to the signals can be found in Table 1. We can see that the RLS method outperforms the other two techniques as to the preservation of signal integrity. Moreover, this case exemplifies a situation where the

uncoupled estimation scheme yield better performance than the joint predictor.

5. CONCLUSION

This work presented a new approach for partial tracking, an essential block of any sinusoidal analysis system. The method is based on a decision-directed prediction of both magnitude and frequency via an RLS adaptive filter, on a frame-by-frame basis. The performance of the proposed strategy compares favorably to that of other methods found in the literature under real circumstances. A resynthesis experiment confirms the significant improvement achieved by the new method, under a formal objective measure of audio quality.

6. REFERENCES

[1] T. F. Quatieri and R. J. McAuley, "Audio signal processing based on sinusoidal analysis/synthesis," in *Applications of digital signal processing to audio and acoustics*, M Kahrs and K. Brandenburg, Eds., chapter 9. Kluwer, 1998.

[2] X. Serra and J. O. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[3] M. Lagrange, S. Marchand, M. Raspaud, and J. Rault, "Enhanced partial tracking using linear prediction," in *Proc. of the DAFx-03 - 6th International Conference on Digital Audio Effects*, 2003, Available at `http://www.dafx.de/`.

[4] M. Lagrange, S. Marchand, and J. Rault, "Using linear prediction to enhance the tracking of partials," in *Proc. of the ICASSP'04 - International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2004, vol. 4, pp. 241–244.

[5] P. Depalle, G. Garcia, and X. Rodet, "Tracking of partials for additive sound synthesis using hidden markov models," in *Proc. of the ICASSP'93 - International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1993, vol. 1, pp. 225–228.

[6] L. Cohen, *Time Frequency Analysis: Theory and Applications*, Prentice-Hall, 1994.

[7] T.F. Quatieri and R. J. McAulay, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[8] J. O. Smith and X. Serra, "PARSHL: An analysis/synthesis program for non-harmonic sounds based on sinusoidal representations," in *Proc. ICMC '87 - International Computer Music Conference*, 1987, pp. 290–297.

[9] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, 4th. edition, 2001.

[10] A. Sayed, *Fundamentals of Adaptive Filtering*, Wiley-IEEE, 2003.

[11] S. Hainsworth and M. D. Macleod, "Time frequency reassigment: A review and analysis," Technical Report CUED/F-INFENG/TR.458, Cambridge University, Engineering Dept., 2003.

[12] ITU-R, "Method for the objective measurement of perceived audio quality," Recommendation BS.1387, ITU, 1998.

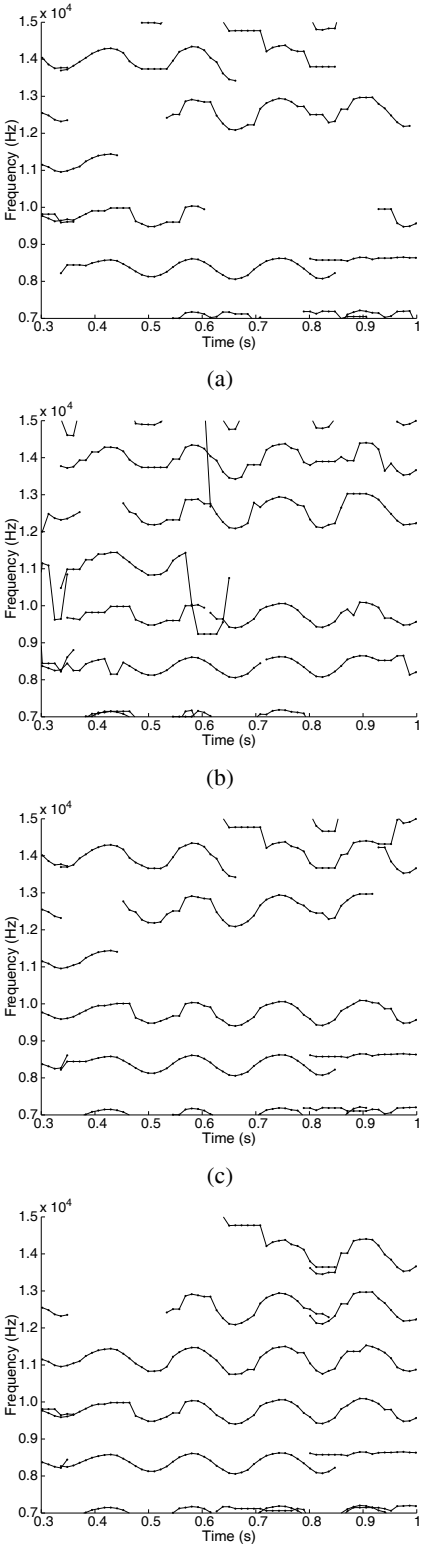

Fig. 4. Violin *vibrato* analysis: (a) Classic method; (b) Burg method; (c) uncoupled RLS; (d) coupled RLS.