FAST SEQUENTIAL LS ESTIMATION FOR SINUSOIDAL MODELING AND DECOMPOSITION OF AUDIO SIGNALS

*Bertrand DAVID, Roland BADEAU**

Ecole Nationale Supérieure des Télécommunications - Département TSI
46 rue Barrault - 75634 PARIS Cedex 13, France
bertrand.david@enst.fr

ABSTRACT

This work demonstrates a sequential Least Squares algorithm applied to the decomposition of sounds into sines-plus-residual models. For a given basis of r distinct frequency components, the algorithm derives recursively the Least Squares estimates of the associated amplitudes and phases. While a direct calculation achieves a $O(nr^2)$ complexity the main cost of our implementation is only of $4r$ multiplications *per* sample, whatever the length n of the analysis window. The technique is extended to basis of exponentially increasing or decreasing frequency components, which provides a fast and enhanced decomposition of rapidly varying segments of the sound. Finally, the proposed method is successfully applied to a real piano note.

1. INTRODUCTION

Numerous audio signal processing systems involve a decomposition stage to enable separate processing of the sinusoidal components and the stochastic part of sounds. Often partly owing to the early work of Serra [1], which has evolved into the referential SMS[1] framework, applications in the field of analysis/synthesis of musical sounds, as for instance Digital Audio Effects [3], have prospered in the past decade ; aside mentioning the successful fate of the Harmonic-plus-Noise Model (HNM) for speech processing (see [4] for instance). These techniques commonly compute spectral estimates (*e.g* frequency of partials or complex poles) and amplitude estimates of the sinusoidal part for subsequently obtaining the residual (*e.g.* the stochastic part) by subtraction. Both estimation problems have widely been studied, including Fourier based methods as in [1] and High Resolution methods [5] for the former ; spectral interpolation [6] and Least Squares (LS) approximation for the latter [7]. This last reference actually presents an in-depth survey of LS amplitude estimation. Different aspects of both problems are jointly discussed in [8].

If the motive which initially propelled this work forward was to obtain the less distorted mechanical residual noise left when removing the sinusoidal content of a piano tone, its applicability extents to any audible scene with a piecewise steady spectral content — *i.e.* the frequency of the partials remain constant over their duration while their amplitude may vary. It is thus a relevant approach for modeling a number of unsustained musical sounds[2].

*This work is supported by the *Groupe des Écoles de Télécommunications*, TAMTAM project and the *Agence Nationale de la Recherche*, under contract ANR-06-JCJC-0027, DESAM project.

[1]Spectral Modeling Synthesis, see [2] for an overview.

[2]The sound produced by a free vibrating, linear mechanical system can be decomposed into a sum of amplitude-modulated sinusoids

An adaptive algorithm is proposed for computing the slowly varying amplitudes of r spectral components, assuming their frequency is known in advance. The method operates at the sample scale, and both parts of the sound are derived at a $O(r)$ cost per sample. , allowing an accurate estimation together with a low latency processing. The novelty of the method is to account for the sinusoidal nature of the model and the associated rotational invariance property of the signal subspace to obtain a low-cost recursive computation.

The paper is organized as follows. The parametric model and the recursive algorithm for estimating and decomposing the signal components are presented in section 2. Section 3 is devoted to performance analysis, which is addressed both theoretically and via numerical simulations. An application to a real piano note is presented in section 4. Finally, the main conclusions of this study are summarized in section 5.

2. MODEL AND PRINCIPLE

2.1. Least Squares estimation of amplitudes

Let x denote a complex sequence, as for instance the analytic representation of an audio signal. It is decomposed as the sum of r complex exponentials and an additive noise w :

$$x(t) = \sum_{k=0}^{r-1} b_k e^{(-\delta_k + j2\pi\nu_k)t} + w(t), \tag{1}$$

where $t \in \mathbb{Z}$ is the discrete time index, δ_k and ν_k respectively being the damping factor and the frequency of the k^{th} partial, and b_k its associated complex amplitude. We assume in addition that the poles $z_k = e^{-\delta_k + j2\pi\nu_k}$, $k = 0, \ldots, r-1$ are distinct. The LS estimates $\hat{\mathbf{b}}$ of these amplitudes over the n-dimensional snapshot $[t, t+1, \ldots, t+n-1]$, with $n \geq r$, are the solution of $\mathbf{x}(t) = \mathbf{V}\mathbf{D}^t\mathbf{b} + \mathbf{w}(t)$which minimizes $\|\mathbf{w}(t)\|_2$; where

$$\begin{aligned} \mathbf{w}(t) &= \begin{bmatrix} w(t) & w(t+1) & \ldots & w(t+n-1) \end{bmatrix}^T, \\ \mathbf{x}(t) &= \begin{bmatrix} x(t) & \ldots & x(t+n-1) \end{bmatrix}^T, \\ \mathbf{b} &= \begin{bmatrix} b_0 & b_1 & \ldots & b_{r-1} \end{bmatrix}^T, \end{aligned}$$

$\mathbf{D} = \text{diag}\{z_0, \ldots, z_{r-1}\}$ and $\mathbf{V}$ is the Vandermonde matrix of the poles defined as:

$$\mathbf{V} = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ z_0 & z_1 & \ldots & z_{r-1} \\ z_0^2 & z_1^2 & \ldots & z_{r-1}^2 \\ \vdots & \vdots & \vdots & \vdots \\ z_0^{n-1} & z_1^{n-1} & \ldots & z_{r-1}^{n-1} \end{bmatrix}. \tag{2}$$

Since the z_k's are distinct, $\mathbf{V}$ is full rank, $\mathbf{V}^H\mathbf{V}$ and $\mathbf{D}^t$ non-singular, and $\hat{\mathbf{b}}$ is well known to satisfy the relation $\hat{\mathbf{x}}(t) = \mathbf{V}\mathbf{D}^t\hat{\mathbf{b}}$ where $\hat{\mathbf{x}}(t)$ is the projection of $\mathbf{x}(t)$ onto the range space of $\mathbf{V}$ (which is also the range space of $\mathbf{V}\mathbf{D}^t \forall t$). This leads to:

$$\hat{\mathbf{b}} = \mathbf{D}^{-t}(\mathbf{V}^H\mathbf{V})^{-1}\mathbf{V}^H\mathbf{x}(t) = \mathbf{D}^{-t}\mathbf{V}^\dagger\mathbf{x}(t), \quad (3)$$

where the superscript † denotes the pseudo-inverse. At each iteration, assuming that $\mathbf{V}^\dagger$ has been derived once and for all, the main cost of the operation described in (3) is the nr MACs[3] involved by the product $\mathbf{V}^\dagger\mathbf{x}(t)$, which achieves a $O(\tau nr)$ complexity when processing a τ-sample long fragment.

2.2. Recursive computation of amplitude estimates

In an adaptive context, the amplitudes are assumed to slowly vary, *e.g.* they become $b_k(t)$ and are considered as quasi-constant over the analysis window duration of n samples. We thus look for solving $\mathbf{d}(t) = \mathbf{V}^\dagger\mathbf{x}(t)$,where $\mathbf{d}(t) = \mathbf{D}^t\hat{\mathbf{b}}(t)$, recursively at each time step. Note that if this problem bears a resemblance to that of the adaptive estimation of FIR filters and its well known RLS (Recursive Least Squares [9]) solution, in our case the matrix $\mathbf{R} = \mathbf{V}^H\mathbf{V}$ does not depend on time[4]. The derivation below indeed uses the classical rewriting of the pseudo-inverse

$$\mathbf{d}(t) = \mathbf{R}^{-1}\mathbf{p}(t), \quad (4)$$

where

$$\mathbf{p}(t) = \mathbf{V}^H\mathbf{x}(t). \quad (5)$$

Let now define $\mathbf{V}_\downarrow$ (*resp.* $\mathbf{V}_\uparrow$) as the submatrix of $\mathbf{V}$ obtained by deleting its last (*resp.* first) row. Then we can rewrite either $\mathbf{V} = [\mathbf{u}\ \mathbf{V}_\uparrow^H]^H$ or $\mathbf{V} = [\mathbf{V}_\downarrow^H\ \mathbf{v}]^H$. Using the first identity while taking into account the rotational invariance property of the Vandermonde matrix $\mathbf{V}_\downarrow\mathbf{D} = \mathbf{V}_\uparrow$, (4) leads to:

$$\mathbf{p}(t) = \hat{x}_0(t)\mathbf{u} + \mathbf{S}\,\mathbf{D}\,\mathbf{d}(t), \quad (6)$$

where $\hat{x}_0(t)$ indicates the first coefficient of $\hat{\mathbf{x}}(t)$, equating the sum of the $\mathbf{d}(t)$ coefficients, and where $\mathbf{S} = \mathbf{V}_\uparrow^H\mathbf{V}_\downarrow$. The second form for the writing of the matrix $\mathbf{V}$, together with the Sherman-Morrison formula (also known as the matrix inversion lemma [10]), leads to

$$\mathbf{R}^{-1}\mathbf{D}^{-H} = \left(\mathbf{I}_r - \frac{{\mathbf{R_0}}^{-1}\mathbf{v}\mathbf{v}^H}{1+\mathbf{v}^H\mathbf{R}_0^{-1}\mathbf{v}}\right)\mathbf{S}^{-1}, \quad (7)$$

where $\mathbf{R}_0 = \mathbf{V}_\downarrow^H\mathbf{V}_\downarrow$ and $\mathbf{D}^{-H} = (\mathbf{D}^H)^{-1} = (\mathbf{D}^{-1})^H$. To obtain a recursion on $\mathbf{p}(t)$, the data vector is decomposed into $\mathbf{x}(t) = [x(t)\ \mathbf{x}_\uparrow(t)^T]^T$ and one sample ahead as $\mathbf{x}(t+1) = [\mathbf{x}_\uparrow(t)^T\ x(t+n)]^T$. The relation (5) then results in

$$\mathbf{p}(t+1) = \mathbf{D}^{-H}(\mathbf{p}(t) - x(t)\mathbf{u}) + x(t+n)\mathbf{v}. \quad (8)$$

Incorporating finally (6), (7) and (8) in (4) leads to

$$\mathbf{d}(t+1) = \mathbf{D}\mathbf{d}(t) - e_0(t)\mathbf{q}_1 + e_n(t)\mathbf{q}_2, \quad (9)$$

where $e_k(t) = x(t+k) - \hat{x}_k(t)$, $k = 0,\ldots,n$, $\mathbf{q}_1 = \mathbf{R}^{-1}\mathbf{D}^{-H}\mathbf{u}$, $\mathbf{q}_2 = \dfrac{{\mathbf{R_0}}^{-1}\mathbf{v}}{1+\mathbf{v}^H\mathbf{R}_0^{-1}\mathbf{v}} = \mathbf{R}^{-1}\mathbf{v}$. It can be noted that the predicted value of x at time $t+n$, taking into account the estimated amplitudes at time t, amounts to $\hat{x}_n(t) = \mathbf{v}^H\mathbf{D}^{t+1}\hat{\mathbf{b}}(t)$.

[3] Multiplication and ACcumulation

[4] Such recursive methods are sometimes referred to as *sequential least squares* in the literature.

Remark 1. The recursive computation of $\mathbf{d}(t)$ following equation (9) implies only $O(r)$ operations, if the fixed quantities $\mathbf{D}, \mathbf{q}_1$ and $\mathbf{q}_2$ have been computed *ab initio*. Besides, it is worth noting that the same recursive computation using the recursion on $\mathbf{p}(t)$ given by equation (8) incorporated in (4) without any further processing leads to a $O(r^2)$ complexity.

Remark 2. Equation (9) takes the form of an update of $\mathbf{d}(t)$ involving the updated value of the vector for constant amplitudes (first term) and correction terms related to past and future samples.

2.3. Recursive decomposition of the signal

$\hat{\mathbf{x}}(t)$ is the LS estimate of the sinusoidal component associated to the snapshot data vector $\mathbf{x}(t)$. At time t, an estimate of the sinusoidal signal is thus obtained by taking the first coefficient of $\hat{\mathbf{x}}(t)$, in short $x_s(t) = \hat{x}_0(t)$. The derivation below will emphasize the benefit of the obtained recursion on $\mathbf{d}(t)$ since it will achieve a $O(r)$ complexity for the recursive decomposition. This algorithm is simply obtained by noticing that $\hat{x}_0(t) = \mathbf{u}^H\mathbf{d}(t)$, leading to the pseudo-code given in table 1.

Table 1: Recursive sinusoidal+noise separation

Initialization :

$z_k \overset{\Delta}{\rightarrow} \mathbf{v_k} = \begin{bmatrix}1 & z_k & \ldots & z_k^{n-1}\end{bmatrix}^T$, $k = 0,\ldots,r-1$, $\mathbf{D} = \mathrm{diag}\{z_k\}$,

$\mathbf{V} = [\mathbf{v}_0\quad \mathbf{v}_1\quad \ldots\quad \mathbf{v}_{r-1}]$, $\mathbf{d}(0) = \mathbf{V}^\dagger\mathbf{x}(0)$,

$\mathbf{R} = \mathbf{V}^H\mathbf{V}$, $\mathbf{u} = \mathbf{1}_{r\times 1}$, $\mathbf{q}_1 = \mathbf{R}^{-1}\mathbf{D}^{-H}\mathbf{u}$, $\mathbf{q}_2 = \mathbf{R}^{-1}(\mathbf{D}^H)^{n-1}\mathbf{u}$,

$\hat{x}_0(0) = \mathbf{u}^H\mathbf{d}(0)$, $\hat{x}_n(0) = \mathbf{u}^H\mathbf{D}^n\mathbf{d}(0) \rightarrow \{e_0(0), e_n(0)\}$

For each time step do

Input sample : $x(t+n+1)$

main section	**Cost**	
$\mathbf{d}(t+1) = \mathbf{D}\mathbf{d}(t) - e_0(t)\mathbf{q}_1 + e_n(t)\mathbf{q}_2$	$3r$	MAC
$\hat{x}_0(t+1) = \mathbf{u}^H\mathbf{d}(t+1)$	r	Add
$\hat{x}_n(t+1) = \mathbf{u}^H\mathbf{D}^n\mathbf{d}(t+1)$	r	MAC
$e_0(t+1) = x(t+1) - \hat{x}_0(t+1)$	1	Add
$e_n(t+1) = x(t+n+1) - \hat{x}_n(t+1)$	1	Add

3. TUNING OF THE ALGORITHM PERFORMANCE

3.1. Variance of the estimates

Since we are dealing with non stationary signals, the question arises of choosing the snapshot length n. On one side it should be kept under a typical temporal value representing the duration where the signal parameters can be considered as roughly constant and on the other side the longer n, the more accurate the estimation of modes and particularly that of neighboring modes. Furthermore, in contrast to much frame-based processing, for instance those which use a Short Time Fourier Transform (STFT) framework, the computational burden in our case is decoupled from the accuracy and resolution issues since it does not depend on n.

In this section, the noise w is assumed to be white circular Gaussian, $\mathcal{N}(0, \sigma^2\mathbf{I})$. Thus, for a given set of r distincts poles z_k, $k = 0,\ldots,r-1$, the estimate of the amplitudes obtained by equation (3) is both that of the maximum likelihood for the considered sliding analysis window and minimum variance unbiased [11]. Following [11] the covariance matrix of the LS estimates $\mathbf{d}$ can be expressed as

$$\mathbf{C_d} = \sigma^2(\mathbf{V}^H\mathbf{V})^{-1} = \sigma^2\mathbf{R}^{-1}. \quad (10)$$

When the poles are on the unit circle, equation (10) indicates that the estimation becomes poor for frequencies close to each other; since the matrix $\mathbf{R}$ then tends to be singular. The calculation of the Cramer-Rao lower bound for parameter estimates or the analytical derivation of their covariance matrix is in fact a broadly investigated field (see for instance [7]) and we will here only stress the case of two components of respective frequencies ν_1 and ν_2. $\mathbf{R}$ is then 2×2 and this leads to the expression of the variance for the component amplitudes:

$$\operatorname{var} d_1 = \operatorname{var} d_2 = \frac{\sigma^2 n}{|\mathbf{R}|}, \tag{11}$$

where the determinant of $\mathbf{R}$ is given by

$$|\mathbf{R}| = n^2 - \frac{\sin^2(\pi n(\nu_1 - \nu_2))}{\sin^2(\pi(\nu_1 - \nu_2))}. \tag{12}$$

The interest of this last formula is to assess the asymptotic variances when $\delta = \nu_1 - \nu_2$ tends towards zero. By developing the expression (12) it follows:

$$\operatorname{var} d_1 \underset{\delta \to 0}{\sim} \frac{3\sigma^2}{\pi^2 \delta^2 n(n^2-1)}. \tag{13}$$

On the other hand, when n becomes large for a fixed δ, equations (11) and (12) show that the variance reduces to

$$\operatorname{var} d_1 \underset{n \to \infty}{\sim} \frac{\sigma^2}{n}. \tag{14}$$

The commutation between both asymptotic regimes occurs for (assuming $n \gg 1$) $\delta\, n = \sqrt{3/\pi^2} = 0.55$ and thus is interpretable in terms of resolution: the estimation accuracy of two spectrally close components becomes dubious when the frequency discrepancy is of order n^{-1}. The following section exemplifies the consequent summary of these comments: n has to be tuned as large as possible taking into account the characteristic temporal variation of the model parameters, since there is no additional cost associated to its growth.

3.2. A tricomponent exemple

Figure 1 represents the amplitude of three cisoids at 400, 430 and 2000 Hz, sampled at 8 kHz and which are respectively exponentially decaying, Hann window modulated and oscillating (tremolo musically compliant at 5 Hz) with finite duration. n has been chosen on purpose as low as 30 samples (3.7 ms) with an overall SNR of 20 dB (here $\sigma = 0.1$), so that the estimates' variances are clearly noticeable. The estimates are drawn in gray solid line while the true values are depicted in black. In dashed black are sketched the standard deviations of the estimates, the one for the first component (400 Hz) being derived following equation (13) and that for the third component (2000 Hz) following equation (14). This shows the usefulness of these asymptotic formulae for evaluating uncertainties in practical cases. Note that in this exemple, the basis[5] used is the Vandermonde matrix. This basis will be herein referred to as *steady* in contrast to those where at least one pole lies inside or outside the unitary circle. As a remark, it should be mentioned that, taking into account the steadiness of the basis, the obtained estimates are delayed by $(n-1)/2$ samples to be relocated in the middle of the analysis window. Finally, if n is increased to 200 samples (40 ms), the results are represented in figure 2. The variance for the first component (and thus for the second) is given by equation (11), reaching a standard deviation of 0.007. This does not explain the oscillations appearing for instance on the first decaying component. These are caused by the non stationarity of the component amplitudes over the observation duration and are thus interpretable as a deterministic bias.

[5] In most cases the signal does not actually globally fit a sines+noise model and the term of *basis* denotes here the linearly independant subset of vectors spanning the signal subspace *locally* defined by the model.

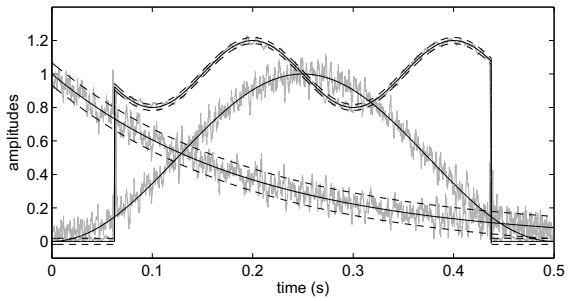

Figure 1: Estimation of 3 component amplitudes, $n = 30$.

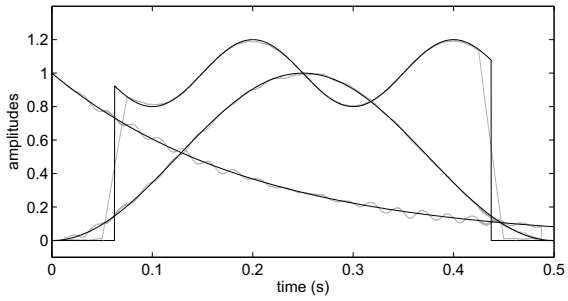

Figure 2: Estimation of 3 component amplitudes, $n = 200$.

3.3. Dealing with unsteady partials

Unsustained musical sounds often demonstrate decaying amplitudes after an attack portion where they rapidly increase in most cases. As a sequel, arises the idea of replacing the steady basis (as used in the preceding exemple) by an unsteady one. Figure 3 illustrates the estimation performance on a single component modulated by a function f of the form $f(t) = t^{\alpha} e^{(t/\tau)^{\beta}}$, normalized and starting at $t = 0.1$ s. The SNR is taken as high as 90 dB in order to assess the deterministic modeling error when projecting the original amplitude onto an exponential basis. Logically, the error obtained when using a 3-dimensional basis with damping factors $\{-0.1n^{-1},\ 0,\ 0.1n^{-1}\}$ (represented in gray) is much lower than that obtained when the basis is chosen 1-dimensional and steady (dashed), even if we can notice a slightly longer pre-echo effect.

4. APPLICATION TO A NOTE DECOMPOSITION

The application demonstrated here considers an E6 piano tone. The high register has been elected for simplicity's sake. To be