EXTRACTION OF SPECTRAL PEAK PARAMETERS USING A SHORT-TIME FOURIER TRANSFORM MODELING AND NO SIDELOBE WINDOWS.

Ph. Depalle, T. Hélie

IRCAM, Analysis/Synthesis Team, 1, Place Igor-Stravinsky, 75004 Paris, France.
Email : phd@ircam.fr, helie@email.enst.fr. Phone : 33-1-44-78-48-45, Fax : 33-1-44-78-15-40.

ABSTRACT

A new method which improves the estimation of frequency, amplitude and phase of the partials of a sound is presented. It allows the reduction of the analysis-window size from four periods to two periods. It therefore gives better accuracy in parameter determination, and has proved to remain efficient at low signal-to-noise ratios. The basic idea consists of using a parametric modeling of the short-time Fourier Transform. The method alternately estimates the complex amplitudes and the frequencies starting from the result of the classical analysis method. It uses least-square procedure and a first-order limited expansion of the model around previous estimations. This method lead us to design new windows which do not have any sidelobe in order to help the convergence. Finally an analysis algorithm which has been built according to the observed behavior of the method for various kinds of sound is presented.

1. INTRODUCTION

The additive synthesis model represents a sound $s(n)$ as a finite sum of K partials (sinusoids whose amplitude a_k, frequency f_k and initial phase ϕ_k, $k \in [1, K]$, vary in time), mixed with a time-varying spectrum envelope noise $b(n)$:

$$s(n) = \sum_{k=1}^{k=K} a_k(n)\cos(\Phi_k(n)) + b(n) \tag{1}$$

where the phase Φ_k is updated as follows :

$$\Phi_k(n+1) = \Phi_k(n) + 2\pi f_k(n) \quad \text{and} \quad \Phi_k(0) = \phi_k \tag{2}$$

In order to extract the temporal evolution of parameters a_k, f_k and ϕ_k, $k \in [1, K]$ from a recorded sound, the classical analysis procedure [8] consists of detecting and selecting spectral peaks from its Short-Time Fourier Transform (STFT) [9] given at time rI by:

$$S_r(f) = \sum_{n=-\infty}^{n=+\infty} w(n - rI)s(n)\exp(-2\pi i f n) \tag{3}$$

Spectral peaks are well defined when the size M of the window w becomes wider than four periods of the lowest frequency included in the analyzed signal. Then frequency and amplitude of each spectral peak are obtained by a polynomial interpolation around the maximum [10]. This procedure has two drawbacks: the constraint on the window size smooths rapid variations of the parameters and the interpolation procedure does not take into account the influence of neighbour peaks which slightly modify the frequency and even more the amplitude.

In order to alleviate these problems, we developed a method based on a parametric modeling of the STFT representation. This method allows to decrease the window size down to two periods and globally takes into account the mutual influence of the spectral peaks.

In this paper, we first describe the principle of the amplitude and frequency estimation method. As this method is very sensitive to the window's shape, we present two new families of windows without sidelobes in the spectral domain. Finally, we detail the synopsis of the analysis algorithm which has been built according to the observed behavior of the method for various kinds of sound.

2. THE METHOD

2.1. STFT parametric model

Lets consider the noiseless part of the additive representation of a sound: we obtain from Eq. (1) the pure sinusoidal model $\hat{s}(n) = s(n) - b(n)$. As we use a STFT to extract the parameters, we make the assumption that the amplitudes and frequencies are constant over the chosen window w. Then the initial phase ϕ_k is updated for each window in order to minimize the distortion due to frequency variations between the model $\hat{s}$ and the signal s. Thus we obtain the local model of the signal $\hat{s}_r(n_r)$, where $n_r = n - rI$ is the local time reference:

$$\hat{s}_r(n_r) = \sum_{k=1}^{k=K} a_k(rI)\cos(2\pi f_k(rI)n_r + \phi_k(rI)) \tag{4}$$

Henceforth we consider only variables defined at instant rI. Therefore, the subscript r and instant rI will be omitted. Then the STFT $\hat{S}(f)$ of the local model $\hat{s}(n)$ is given by the parametric expression:

$$\hat{S}(f) = \sum_{k=1}^{k=K} \frac{a_k}{2}\left(\mathrm{e}^{i\phi_k}W(f - f_k) + \mathrm{e}^{-i\phi_k}W(f + f_k)\right) \tag{5}$$

where $W(f)$ is the Fourier transform of the analysis window.

2.2. The estimation method

$\hat{S}(f)$ can be considered as a spectrum estimator of the STFT $S(f)$ of the observed signal $s(n)$. Now, the goal of our method is to identify the parameters for which the model $\hat{S}(f)$ best fits the observed spectrum $S(f)$ (Cf. Figure (1)) according to a least-square criterion. For that, the observed STFT measured at N equally spaced frequencies $F_j = (j - \frac{N}{2} - 1)(\frac{1}{N})$ for $j = 1, \ldots, N$; N being the power of 2 immediatly greater than M. Then, defining the notation $\underline{X}$ as an N-size column vector composed of $\underline{X}_j = X(F_j)$, the best estimate $\hat{S}$ of S is obtained by minimizing the cost :

$$\|\underline{S} - \underline{\hat{S}}\| \tag{6}$$

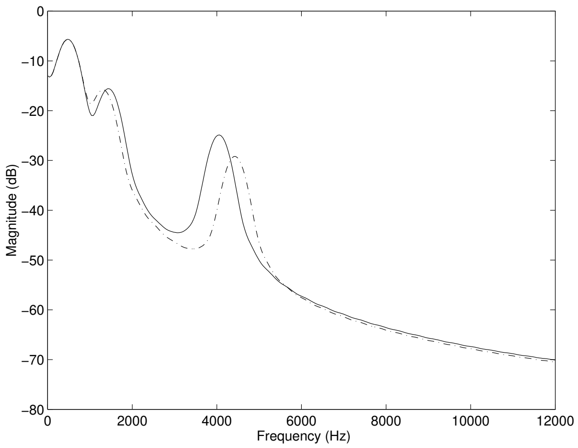

Figure 1: Observed Spectrum $S(f)$ (solid line) compared to its estimation $\hat{S}(f)$ (dashed dotted line).

The expression of $\hat{S}$ is linear in terms of a_k and $e^{i\phi_k}$ but non linear in f_k. And even if we linearize the dependence between $\hat{S}$ and the frequency parameters as in Eq. (13), the expression of $\hat{S}$ remains nonlinear as it contains products of unknown parameters. So we developed an iterative algorithm which alternately improves the estimates of amplitudes and frequencies starting from the results of the classical analysis method.

Amplitude and Phase estimation. For this step, we assume that the frequencies of the partials are known and we estimate their amplitude and their phase using the model defined in Eq. (5). The spectrum estimator takes the following linear structure:

$$\hat{S}(f) = \sum_{k=1}^{2K} p_k H_k(f) \tag{7}$$

in terms of the $2K$ unknown parameters p_k defined by:

$$\begin{cases} p_k & = & \frac{a_k}{2}\cos\phi_k & k \in [1, K] \\ p_{K+k} & = & \frac{a_k}{2}\sin\phi_k & k \in [1, K] \end{cases} \tag{8}$$

and the known $2K$ expressions related to the Fourier Transform of $w(n)$:

$$\begin{cases} H_k(f) & = & W(f - f_k) + W(f + f_k) \\ H_{K+k}(f) & = & i\left(W(f - f_k) - W(f + f_k)\right) \end{cases} \tag{9}$$

Defining $\mathcal{H}$ as a matrix of dimension $N \times 2K$ where $(\mathcal{H})_{j,k} = H_k(F_j)$ and $\underline{p}$ the vector of the unknown parameters, we obtain:

$$\underline{\hat{S}} = \mathcal{H}.\underline{p} \tag{10}$$

Then we deduct the least-square solution [7]:

$$\underline{p} = (\mathcal{H}^H.\mathcal{H})^{-1}.\mathcal{H}^H.\underline{S} \tag{11}$$

This procedure gives very good results and it has been shown [11] that, even with a strong presence of noise (-10 dB), amplitudes are well estimated and have very little fluctuations. Since there is no constant value in the model $\hat{S}(f)$, we center the signal in order to remove its mean value.

Frequency estimation. Let us assume that we know the amplitudes and the phases of the K sinusoids. Eq. (5) clearly shows that the dependence of the model on the frequencies is non-linear. In order to obtain a linear formulation, in addition we assume that we have a rough approximation $\mathcal{F}_k$ of each frequency f_k for $k = 1, \ldots, K$. The problem is now to estimate the distance Δ_k between f_k and $\mathcal{F}_k$ ($\Delta_k = f_k - \mathcal{F}_k$). For each frequency measurement point F_j, we linearize the frequency dependence by using a first-order limited expansion of the Fourier Transform of the analysis window W around $F_j - \mathcal{F}_k$ for $k = 1, \ldots, N$:

$$W(f \mp f_k) = W(f \mp \mathcal{F}_k) \mp W'(f \mp \mathcal{F}_k).\Delta_k + o(\Delta_k^2) \tag{12}$$

Thus rewriting $\underline{\hat{S}}$, we obtain the following expression [4]:

$$\underline{\hat{S}} = \underline{\tilde{S}} + \Omega.\underline{\Delta} \tag{13}$$

where $\tilde{S}$ is the STFT model evaluated with the rough approximation frequency vector $\underline{\mathcal{F}}$ and with matrix Ω defined by:

$$(\Omega)_{j,k} = \frac{a_k}{2}\left(-\mathrm{e}^{i\phi_k} W'(F_j - \mathcal{F}_k) + \mathrm{e}^{-i\phi_k} W'(F_j + \mathcal{F}_k)\right) \tag{14}$$

Finally, using the least-square solution, we obtain the estimation:

$$\underline{f} = \underline{\mathcal{F}} + (\Omega^H.\Omega)^{-1}.\Omega^H.\left(\underline{S} - \underline{\tilde{S}}\right) \tag{15}$$

3. DESIGN OF WINDOWS ADAPTED TO THE METHOD

3.1. Influence of the window shape

The STFT is expressed as a convolution product between spectral lines and the Fourier Transform of the window W. Furthermore, the expression of W appears at each step of the preceding algorithm and is actually a degree of freedom of the estimation method. It is then worthwhile to study the effects of the window's shape on the behavior of the algorithm.

The amplitude estimation is not very sensitive to the shape of the window except when two partials become very close in frequency. Classically, we have to choose windows with small bandwidth BW. Then it decreases the ill-conditioning of $\mathcal{H}^H\mathcal{H}$ by increasing the dissimilarity between the columns of the matrix $\mathcal{H}$. In order to minimize the smoothing effect of time variation of parameters and for a given bandwidth, we prefer windows with a small effective duration. This parameter is correlated [6] to the inverse of the Equivalent Noise Bandwidth EQN [5]. So we will search windows with a small ratio $\frac{BW}{EQN}$.

Finally we may notice that spectral leakage does not constitute a drawback to estimate amplitudes when frequencies are perfectly known since the information of each partial is spread over a great many measurement points. This gives also a better robustness when analyzing poor signal-to-noise ratio signals.

The frequency estimation is much more sensitive to the shape of the window as it is based on the first order expansion of W. By initializing the algorithm with a rough frequency approximation located far from the right value, it may converge to the position of the maximum of a sidelobe of W instead of converging to the center of the main lobe. And when the amplitude of the partial becomes very low, the algorithm may oscillate instead of converging. Then we can conclude that the presence of sidelobes in W is a serious problem to the robustness of the method.

3.2. New windows without sidelobes

Windows well adapted to our method may have a small ratio $R = \frac{BW}{EQN}$ and no sidelobes. To our knowledge we found in the literature only one window family without sidelobes: the Hanning-Poisson window [1] [5]:

$$w(n) = \frac{1}{2}\left[1 + \cos(\pi \frac{n}{M/2})\right] \mathrm{e}^{-\alpha \frac{|n|}{M/2}} n \in \left[-\frac{M}{2}, \frac{M}{2} - 1\right] \tag{16}$$

which has no sidelobe for $\alpha \geq 2$ and its smallest value of $R = 0.926$ for $\alpha = 2$.

To have a better control on the characteristics of W, we designed two families of windows called $A(a, b)$ and $B(a, b, c)$ [6]. In this paper, only the first one $A(a, b)$ is presented. To design this family of windows, the idea was to use a gaussian function which has no sidelobes but which is not time limited and to multiply it by a power of the triangular window (PTW) whose Fourier transform is always real and positive. When the variance of the gaussian is large enough, the smoothing in the spectral domain removes the sidelobes of PTW.

$$A(a, b)(n) = \left(1 - \frac{|n|}{M/2}\right)^a .\mathrm{e}^{-b.(\frac{|n|}{M/2})^2} n \in \left[-\frac{M}{2}, \frac{M}{2} - 1\right] \tag{17}$$

Then for each values of a, an optimal corresponding value of b which avoid the sidelobes is determined by dichotomy. Figure (2) shows the temporal and frequential shape of six windows of this family. For the spectral window W, notice the absence of sidelobe and the independant control of the narrowness of the main lobe and of the asymptotic level far from the main lobe. This can not be achieved by the Hanning-Poisson window.

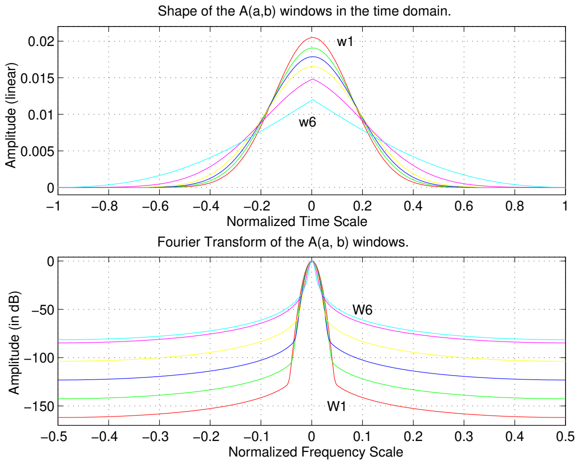

Figure 2: Six examples of $A(a, b)$ windows ($w_1 = A(10^{-4}, 21.6)$ and $w_6 = A(1.8, 0.92)$).

3.3. Example

We show an example which illustrates the ability of the method to converge to the right values even when it is initialized by very bad approximations. Let us consider a signal obtained by the superposition of three sinusoids whose parameters $\{f_k, a_k\}$ for $k = 1, \ldots, 3$ are given in the following table:

Order of Partial	Right Frequency	Right Amplitude	Frequency Initialization
1	$f_1 = 440$ Hz	$a_1 = 1.0$	$\mathcal{F}_1 = 50$ Hz
2	$f_2 = 1400$ Hz	$a_2 = 1.0$	$\mathcal{F}_2 = 2300$ Hz
3	$f_3 = 4000$ Hz	$a_3 = 1.0$	$\mathcal{F}_3 = 5000$ Hz

Notice that the frequencies $\mathcal{F}_k$ used to initalize the algorithm are far from the solution. The analysis parameters are defined as follows:

Window Type	Sampling Frequency	Window Size	FFT Channels
w_6	$F_s = 44100$ Hz	200	1024

The window size represents approximatively two periods of the first partial. Figure (3) shows the efficency of the method which converges in twelve iterations.

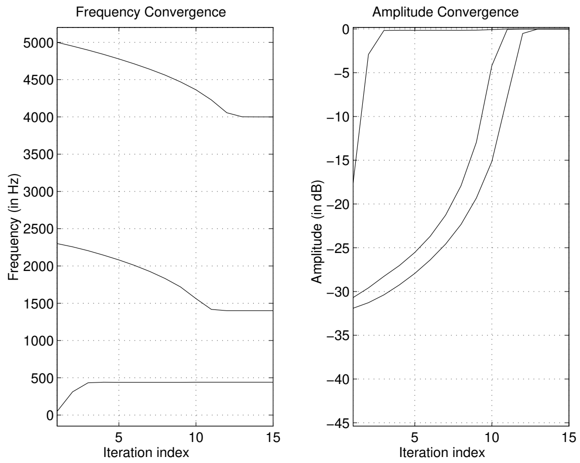

Figure 3: Convergence of the method on a mix of three sinusoids.

4. THE ALGORITHM

We used this method to analyze various kinds of sound. This led us to add several functionalities which help to converge or to speed up the computation. Figure (4) shows the resulting algorithm which is detailed now.

Frequency Initialization. To start the algorithm we perform a "classical analysis" using a window with a very low bandwidth (usually a rectangular window). Secondly we select the spectral peaks that have a power greater than a relative portion of the highest detected one and whose shape are close to that of W around the maxima. The size of the window is chosen to contain two periods of a 20 Hz sinusoid. According to the detected spectral peaks, the size is automatically modified.

Spectrum Splitting. We first detect the L lowest minima in the spectrum which are close to already detected maxima. Then we eventually catenate the bands which contain very few maximas with adjacent bands until they reach an assigned number of peaks.

Amplitude and Frequency Estimation. For each band, we perform the iterative method described in section (3).

Peak Fusion Manager. When a "spurious" peak remains selected at the frequency initialization step, it often becomes closer and closer

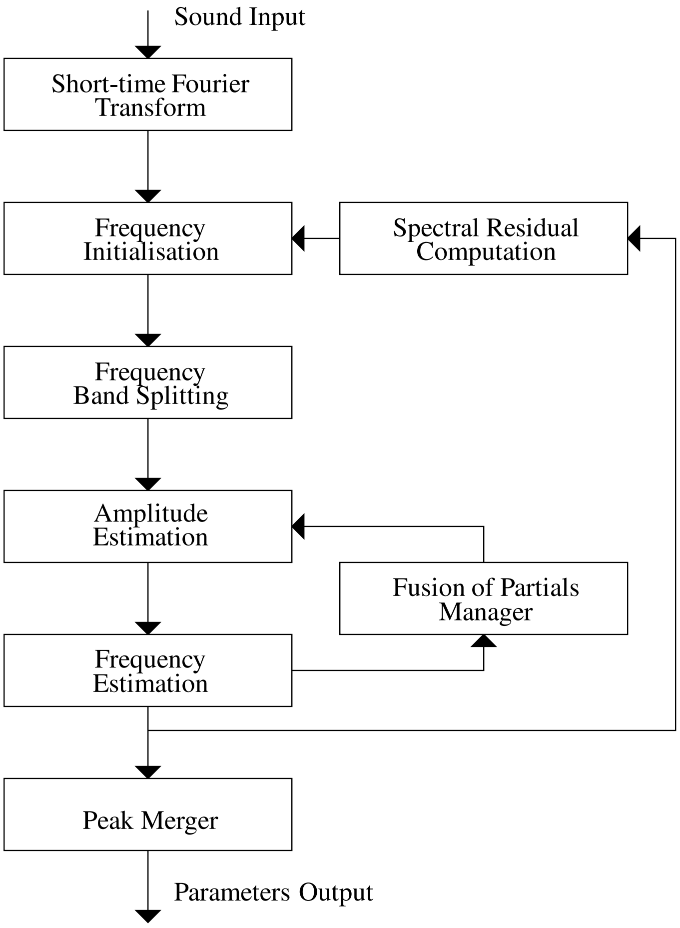

Figure 4: Synopsis of the Algorithm.

in frequency to another detected peak. This increases the conditioning of the matrice $\mathcal{H}^H\mathcal{H}$ and may give huge amplitudes and opposite phases to the two close peaks. To avoid this problem, we remove one of the two peaks when their frequencies become stable and their distance is lower than a threshold. If the signal actually contains the two peaks, the peak which has been eliminated will be selected again during the following spectral residual analysis.

Spectral Residual Computation. As we are using small window sizes, some peaks are not detected at the frequency initalisation step. It happens when their amplitudes are small, when they are too close to other peaks or when they have been rejected at the preceding step. Then we perform a new analysis on the difference between the observed spectrum S and the estimated spectrum $\hat{S}$. In practice the analysis of the spectrum and its spectral residual appears to be sufficient but one may iterate if needed.

Peak Merger. We have to merge the sets of peaks extracted from the observed spectrum $\hat{S}$ and the successive analyzed spectral residuals $\hat{S}_1, \ldots, \hat{S}_v$. We catenate them by pairs, going from the last obtained set_v to the first obtained set_0 (corresponding to $\hat{S}$). For each catenation, we remove the peaks of $\hat{S}_l$ which appear to be masked by peaks of $\hat{S}_{l-1}$.

Amplitude Estimation. Finally we estimate again the amplitude of the peaks selected at the preceding step.

This algorithm is then applied on each window centered at time rI to obtain successive sets of partials. To drive an additive synthesizer, these sets can be transformed in temporal trajectories of partials using hidden Markov models [2].

5. CONCLUSION

We have presented a new spectral analysis method which improves the estimation of time-varying frequency, amplitude and phase of the partials of a sound. It is based on a parametric modeling of the short-time Fourier transform of the sound. It reduces the size of the window by a factor of two and takes into account the mutual influence of the peaks for the amplitude estimation. This method has been proved to remain efficient at low signal-to-noise ratios. We have also designed new families of windows without sidelobe structure and manage the fusion of spectral peaks in order to improve the convergence of the algorithm. Finally we are able to detect low amplitude partials by an iterative analysis of the spectrum residual and we speed up the algorithm by splitting the spectrum in spectral bands.

References

1. N. K. Bary. *A Treatise on Trigonometric Series*. MacMillan, New York, 1964.
2. P. Depalle, G. Garcia, and X. Rodet. Tracking of partials using hidden Markov models. In *IEEE ICASSP-93, Minneapolis, Minesota*, April 1993.
3. P. Depalle and X. Rodet. A new additive synthesis method using inverse Fourier transform. In *Proc. Int. Computer Music Conf. (ICMC'92)*, pages 410–411, October 1992.
4. P. Depalle and L. Tromp. An improved additive analysis method using parametric modelling of the short-time Fourier transform. In *Proc. Int. Computer Music Conf. (ICMC'96)*, 1996.
5. F. J. Harris. On the use of windows for harmonic analysis with the discrete Fourier transform. In *IEEE Proc-78*, volume 66, pages 51–83, January 1978.
6. T. Hélie. Extraction des paramètres de partiels par modélisation de la transformée de Fourier à court-terme utilisant des fenêtres spectrales sans lobes. Master's thesis, Telecom-Paris, 1997.
7. Ch. L. Lawson and R. J. Hanson. *Solving least squares problems*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1974.
8. R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on sinusoidal representation. In *IEEE ASSP-34*, pages 744-754, August 1986.
9. L. R. Rabiner and R. W. Schafer. *Digital processing of speech signals*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1978.
10. X. Serra. A system for sound analysis-transformation-synthesis based on a deterministic plus stochastic decomposition. Phd, Stanford University, 1989.
11. L. Tromp. Amélioration de l'extraction de partiels dans les signaux sonores. Master's thesis, Université Paris-Sud Orsay, 1995.