EFFICIENT PARAMETRIC MODELING FOR AUDIO TRANSIENTS

Rémy Boyer and Karim Abed-Meraim

ENST, Department of Signal and Image Processing
46, rue Barrault, 75634 Paris Cedex 13
boyer,abed@tsi.enst.fr

ABSTRACT

In this work, we present an evolution of the DDS (Damped & Delayed Sinusoidal) model introduced within the framework of the general signal modeling. This model is named *Partial* Damped & Delayed Sinusoidal (PDDS) model and takes into account a single time delay parameter for a set of (un)damped sinusoids. This modification is more consistent with the transient audio modeling problem. Then, we develop model parameter high-resolution estimation algorithms. Simulations on a typical transient audio signals show the validity of this approach.

1. INTRODUCTION

The audio transient compact representation by parametric models is an up-to-date and difficult problem [1, 2]. Parametric EDS (Exponentially Damped Sinusoidal) model has been widely studied in the signal processing community [3]. However, its application to signal compression is quite recent [1, 4, 5]. This approach comes as a natural evolution of the sinusoidal model introduced by McAulay & Quateri [6]. In fact, sinusoidal models assume that model parameters have slow variation regarding the analysis time range. Yet, this is not always consistent with the fast-varying character of a transient signal. EDS model and its extensions [5] permits more appropriate fast time variations signal modeling since each sinusoidal component amplitude is allowed to vary exponentially over time. Based on this property, the EDS model presents a growing interest in the audio community since it allows to model compactly almost the totality of the audio signals. However, it is well known that this model becomes ineffective on sharp transient signals like some percussive sounds (castanets, gong, triangle, ...) [4, 7, 8]. Modeling characteristic artifacts are then created with two effects. First, the apparition of a pre-echo signal [2] *i.e* a distortion before the sound onset. Second, the signal dynamic is badly reproduced. These phenomena appear to be very prejudicial to the auditory perception of this sound category.
Recently, the parametric model, called DDS (Damped & Delayed Sinusoids) was presented in [7] as a generalization of the sinusoidal and EDS models. In this work, we make two realistic assumptions **(A.1)** A percussive audio signal can be seen as a set (sum) of damped sinusoids, all having a same time-delay. **(A.2)** Two successive audio transients are at "sufficient" relative distance one from an other to perform an efficient time-delay estimation/detection based on the signal envelop variation. In this context, we modify the general DDS model and introduce the Partial Damped & Delayed Sinusoidal (PDDS) model. This model can be seen as a generalization of the EDS model and a particular case of the DDS model.
After that, we propose model parameter High-Resolution (HR) estimation algorithms, named PDDS-D[1] and PDDS-MC[2] and we explain why it is necessary to use HR methods in the audio transient modeling problem context. Finally, we show the efficiency of this approach on typical percussive sounds.

2. PARTIAL MODEL : PDDS DEFINITION

In [7], we present the M-order parametric DDS (Damped & Delayed Sinusoidal) model. In this approach, every waveform 1-DDS possesses a delay parameter : $\{t_m\}_{1\leq m\leq M}$. Yet, in an audio modeling application, it is sufficient to consider a small number K of transient signals on a N-sample analysis such as $K \ll M$. We note k the index of the k-th transient signal and we fix $\sum_{k=0}^{K} M_k = M$ where M_k is the modeling partial order to represent the k-th transient signal with a support of $N_k = N - t_k$ samples. We denote $\{t_0, t_1, \ldots, t_{K+1}\}$ the delay parameter set with $t_0 = 0$, $t_{K+1} = N-1$, $t_k < t_{k+1}$, $0 \leq t_k \leq N-1$ and $B_k = t_{k+1} - t_k$. In relation with assumption **(A.1)**, we define the real M_k-PDDS model for $n = 0, \ldots, N-1$, by

$$\hat{x}_k(n) \triangleq \sum_{m=1}^{M_k} a_{m,k} e^{d_{m,k}(n-t_k)} \cdot \cos(\omega_{m,k}(n-t_k) + \phi_{m,k})\psi(n-t_k) \quad (1)$$

In the previous expression, $d_{m,k}$ is the (negative) damping factor, $\omega_{m,k}$ is the angular-frequency and $a_{m,k}$ and $\phi_{m,k}$ are respectively the m-th real amplitude and the m-th initial phase of the k-th M_k-PDDS model. The poles $z_{m,k}$ are defined by $z_{m,k} = e^{d_{m,k}+i\omega_{m,k}}$. Moreover, the Heaviside function $\psi(n)$ is defined by "1" for $0 \leq n \leq N-1$ and "0" otherwise. Note that there is a unique delay t_k for a set (sum) of M_k EDS waveforms (see figure 1).
Now, we can write the M-PDDS model expression as the sum of $(K+1)$ partial models, according to

$$\hat{x}(n) \triangleq \sum_{k=0}^{K} \hat{x}_k(n) \quad (2)$$

2.1. Models Equivalence

If we assume that the signal $\hat{x}_k(n)$ is time shifted of the quantity "$+ t_k$", we have, for $n = 0, \ldots, N_k - 1$

[1] D stands for Deflation.
[2] MC stands for Multi-Channel.

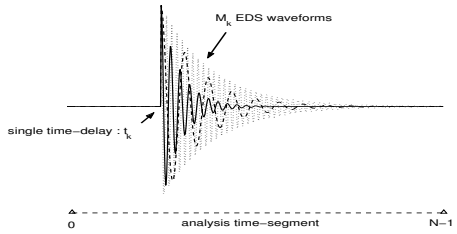

Figure 1: *M_k-PDDS model : one single time-delay for a set (sum) of M_k EDS waveforms.*

$$\hat{x}_k(n+t_k) = \sum_{m=1}^{M_k} a_{m,k} e^{d_{m,k} n} \cos(\omega_{m,k} n + \phi_{m,k}) \quad (3)$$

We recognize the expression of the real M_k-EDS model defined on a N_k-sample support. Moreover, we consider a second signal on the B_k-sample support $\{t_k, \ldots, t_{k+1} - 1\}$, defined by $\hat{x}_k(n+t_k)$ where $0 \leq n \leq B_k - 1$. The latter can be seen as a truncated version of the signal of expression (3) by discarding the $N_k - B_k$ last samples. This operation is made in a view to eliminate the perturbation of the $(k+1)$-th transient signal. We conclude that if we have the knowledge of the delay t_k, then the time translation of the quantity " $+ t_k$" of the M_k-PDDS model and the time support reduction (N_k to B_k) lead to consider an analysis by a M_k-EDS model on a B_k ($\leq N_k$) sample support. Once the model parameter estimation procedure is accomplished, we reconstruct the M_k-PDDS model by making the "inverse" operation, *i.e.* a time support extension (B_k to N_k) and a translation of the quantity " $- t_k$". We define, in a similar way, the t_k-sample shifted audio signal by $x_k(n) = x(n + t_k)$ for $0 \leq n \leq B_k - 1$.

2.2. Model parameters estimation

2.2.1. Efficient spectral analysis : High-Resolution (HR) method

Recalling that B_k is the analysis segment size of the k-th transient signal. This quantity can be quite small if t_k is large enough so it leads to a frequency resolution problem. Indeed, the Fourier resolution is of order $1/B_k$ for a B_k-sample segment. We can realize that the frequency resolution can be too large to make an efficient spectral analysis based on a Fourier-type method. Consequently, we use a HR method to jointly estimate the angular-frequencies and the damping-factors. These methods allow to overcome the Fourier resolution and perform well on very short time segment. More precisely, we use the Kung's algorithm [9]. This method is based on the fundamental shift-invariance property of the signal basis.

2.2.2. Delays estimation and detection

A transient signal can be seen as a very fast variation of the power of its time envelop. So, in relation with assumption **(A.2)**, it seems natural to compute the time envelop of the audio signal and to design an power transient detector based on the its variation. Consequently, we consider, here, a modified version of the detector, introduced in [10]. This modification consists of applying the detector on the signal smooth time envelop, instead of the audio signal. This improves slightly the detection/estimation performance. The smooth time envelop of the signal is computed by considering the median filtering of the modulus of the analytical signal $\nu_P(n)$. More precisely, we have

$$\nu_P(n) = |\nu(n)| * f_P(n) \quad (4)$$

where the analytical signal is defined by $\nu(n) = x(n) + i\Psi_x(n)$, $\Psi_x(n)$ being the Hilbert transform of the audio signal and $f_P(n)$ is a non-linear median filter of length $2P$. Note that using a non-linear median filter allows to obtain a smooth time envelop of the signal, *i.e.* without some awkward oscillatory phenomena. On the other hand, this filter with short duration, typically $P = 5$ or less, keeps unchanged the global variation of the signal time envelop. The used detection strategy is the one introduced in [10].

2.3. PDDS-D algorithm

2.3.1. Partial orders allocation

Basically, in an audio signal, we have two important features : the spectral content and the time waveform. For some kind of quasi-stationary signals, it is most important to well represent their spectral variations without considering too much the variation of the time waveform [6]. In the context of transient audio compact modeling, the signal time waveform is the main audio feature and has to be represented as best we can (*e.g.* without pre-echo). From this point of view, we choose to estimate the partial orders by the following empirical approach : a small partial order will be associated to low power signals since they do not need an accurate modeling. Conversely, higher power signals are associated to larger partial orders. Consequently, we introduce $\gamma \in \mathbb{R}$ according to $M_k = \lceil \gamma \, . \, \varepsilon_k \rceil$ where ε_k is the power of the B_k-sample audio signal $\boldsymbol{x}_k$, according to $\varepsilon_k = ||\boldsymbol{x}_k||_2^2 / B_k$. Afterwards, we fix

$$\gamma = \frac{M}{\varepsilon_0 + \varepsilon_1 + \ldots \varepsilon_K} \quad (5)$$

2.3.2. Poles and complex amplitudes estimation

We begin by estimating the delays $\{t_k\}$ and the partial orders $\{M_k\}$ according to the previous methodologies. The PDDS-D algorithm principle is as follows : for each $\boldsymbol{x}_k$, we estimate the signal poles $\{z_{m,k}\}_{1 \leq m \leq M_k}$, according to a HR method and the complex amplitude parameters $\{\alpha_{m,k} = a_{m,k} e^{i\phi_{m,k}}\}_{1 \leq m \leq M_k}$ by resolution of the following linear least squares criterion

$$\arg\min_{\boldsymbol{\alpha}_k} ||\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k||_2 = \arg\min_{\boldsymbol{\alpha}_k} ||\boldsymbol{x}_k - \boldsymbol{Z}_k^{(B_k)} \boldsymbol{\alpha}_k||_2 \quad (6)$$

where $\boldsymbol{Z}_k^{(B_k)}$ is the $B_k \times D_k$ Vandermonde matrix computed from the poles and their conjugates and $\boldsymbol{\alpha}_k = (\alpha_{1,k}, \alpha_{1,k}^*, \ldots, \alpha_{M_k,k}, \alpha_{M_k,k}^*)^T$. The solution of criterion (6) is $\boldsymbol{\alpha}_k = \boldsymbol{Z}_k^{(B_k)\dagger} \boldsymbol{x}_k$ where $\dagger$ denotes the pseudo-inverse. We, then, can synthesize the M_k-EDS model $\hat{x}_k(n + t_k)$. After that, we build the M_k-PDDS N-sample signal $\hat{x}_k(n)$ by a time support extension (B_k to N_k) and a " $- t_k$" shifting of the model M_k-EDS.

2.3.3. Deflation procedure

In a deflation procedure context, the algorithm begins by initializing the first residual signal $r_0(n) = x(n)$. At the k-th iteration and for the k-th residual signal $r_k(n)$, we estimate the M_k-EDS model : $\hat{r}_k(n + t_k)$ and we reconstruct the M_k-PDDS signal,

$\hat{r}_k(n)$. Then, we add it to the synthesis signal $\hat{x}_{k-1}(n)$. This operation is named the synthesis stage. And finally, we remove its contribution to the last residual signal $r_k(n)$ to compute the next residual signal $r_{k+1}(n)$.

2.4. PDDS-MC algorithm : "Multi-Channel" approach

2.4.1. First Hankel matrix factorization : PDDS-MC1 algorithm

It is possible to consider the analyzed segment as a set of "multi-channel" signals. In this approach, we estimate jointly the damping-factor and the angular-frequency parameters for the $(K+1)$ signals $\{\hat{x}_k(n+t_k), n=0,\dots B_k-1\}_{0\leq k\leq K}$. We define the non-square $L_\nu \times L_k$ Hankel matrix $\mathcal{H}(\hat{\boldsymbol{x}}_k)$ such as $L_\nu + L_k = B_k$ and $2M \leq L_\nu$. We introduce the block-Hankel matrix according to

$$\boldsymbol{H}(\hat{\boldsymbol{x}}) \stackrel{\triangle}{=} \left[\mathcal{H}(\hat{\boldsymbol{x}}_0) \quad \mathcal{H}(\hat{\boldsymbol{x}}_1) \quad \dots \quad \mathcal{H}(\hat{\boldsymbol{x}}_K)\right] \tag{7}$$

Its rank is $2M$ under condition that all the poles are different and without modeling noise. Every matrix $\mathcal{H}(\hat{\boldsymbol{x}}_k)$, represents the Hankel data matrix of the k-th channel of B_k samples size and verifies a factorization in a Vandermonde basis [9], such as $\mathcal{H}(\hat{\boldsymbol{x}}_k) = \boldsymbol{Z}_k^{(L_\nu)} \boldsymbol{\Gamma}_k \boldsymbol{Z}_k^{(L_k)T}$ with $\boldsymbol{\Gamma}_k = \text{diag}(\boldsymbol{\alpha}_k)$. Consequently, $\boldsymbol{H}(\hat{\boldsymbol{x}})$ admits the following factorization :

$$\boldsymbol{H}(\hat{\boldsymbol{x}}) = \boldsymbol{\Theta} \,.\, \boldsymbol{\Lambda}_1 \tag{8}$$

where $\boldsymbol{\Theta} = [\boldsymbol{Z}_0^{(L_\nu)} \dots \boldsymbol{Z}_K^{(L_\nu)}]$ and $\boldsymbol{\Lambda}_1 = \text{Bdiag}([\boldsymbol{\Gamma}_0 \boldsymbol{Z}_0^{(L_0)T} \dots \boldsymbol{\Gamma}_K \boldsymbol{Z}_K^{(L_K)T}])$ with Bdiag(.) denotes a block diagonal matrix. We notice that factorization (8) highlights the row-shift invariance property of matrix $\boldsymbol{\Theta}$ which is a block-Vandermonde matrix. It is thus possible to use a HR method on $\boldsymbol{H}(\hat{\boldsymbol{x}})$ and to jointly determine the signal poles.

2.4.2. Second Hankel matrix factorization : PDDS-MC2 algorithm

Another approach is to consider the $(B_\nu - B_k)$ sample zero-padded signals $\hat{\boldsymbol{x}}_k^{(zp)}$ with $B_\nu = \max_k B_k$, according to $\hat{\boldsymbol{x}}_k^{(zp)} = [\hat{\boldsymbol{x}}_k^T \; \boldsymbol{0}_{B_\nu - B_k}^T]^T$. Based on the properties of the Hankel operator, we have

$$\mathcal{H}\left(\sum_{k=0}^{K} \hat{\boldsymbol{x}}_k^{(zp)}\right) = \sum_{k=0}^{K} \mathcal{H}(\hat{\boldsymbol{x}}_k^{(zp)}) \approx \boldsymbol{\Theta} \,.\, \boldsymbol{\Lambda}_2 \tag{9}$$

where $\boldsymbol{\Lambda}_2 = [\boldsymbol{Z}_0^{(L_\nu)} \boldsymbol{\Gamma}_0 \dots \boldsymbol{Z}_K^{(L_\nu)} \boldsymbol{\Gamma}_K]^T$ with $B_\nu = 2L_\nu$ (square Hankel matrix). Due to the zero-padding, factorization (9) is only an approximation. However this approximation does not affect much the performance of the method. Note, we have to satisfy the constraint $4M \leq B_\nu$.

2.4.3. Poles processing

The $2M$ poles are estimated in the following manner

$$\{z_{m,k}\} = \lambda_{2M}\left\{\boldsymbol{U}_\downarrow^\dagger \boldsymbol{U}_\uparrow\right\}, \;\; \forall m, \; \forall k \tag{10}$$

where $\boldsymbol{U}$ is the matrix containing the $2M$ left singular vectors of $\boldsymbol{H}(\hat{\boldsymbol{x}})$ or $\mathcal{H}(\hat{\boldsymbol{x}})$, $\lambda_{2M}\{.\}$ is the set of $2M$ eigenvalues and $\downarrow$ (respectively $\uparrow$) stands for deleting the bottom (respectively top) row. In presence of audio data (noisy data), we, simply, substitute $x_k(n)$ for $\hat{x}_k(n)$ without too much decreased the performance of the HR algorithm.

2.4.4. Pairing operation

For the two PDDS-MC methods, there is a pairing problem between the time-delays $\{t_k\}_{0\leq k\leq K}$ and the couples $\{\omega_\ell, d_\ell\}_{1\leq \ell\leq M}$. In other words, we have to associate the right time-delays to the right angular-frequencies and damping-factors. A simple way, to resolve this problem is to compute a "collection" of waveforms $g_\ell(n) = e^{d_\ell n}\cos(\omega_\ell n)$ from the set of estimated couples $\{\omega_\ell, d_\ell\}$ and to maximize over k the normalized correlation coefficient $\rho_{\ell,k}$ between each possible B_k-sample waveforms $\boldsymbol{g}_\ell$ and the audio signal $\boldsymbol{x}_k$. Then, for a given index ℓ, we have

$$\arg\max_k \rho_{\ell,k} \quad \text{where} \quad \rho_{\ell,k} \stackrel{\triangle}{=} \frac{\left|\boldsymbol{g}_\ell^T \,.\, \boldsymbol{x}_k\right|}{||\boldsymbol{g}_\ell||_2 \,.\, ||\boldsymbol{x}_k||_2} \tag{11}$$

Note that from expression (11), we can, easily, deduce the modeling partial orders $\{M_k\}$.

2.4.5. Complex amplitudes estimation

The complex amplitudes are determined by solving the criterion $\arg\min_{\tilde{\boldsymbol{\alpha}}} ||\boldsymbol{x} - \hat{\boldsymbol{x}}||_2^2$ where $\boldsymbol{x}$ is the N-sample audio signal and $\hat{\boldsymbol{x}}$ is the N-sample modeled signal.

3. MODELING WITH THE PDDS MODEL

3.1. Time modeling

We test and compare the PDDS-D and PDDS-MC algorithms with the EDS approach on a typical audio transient signal : castanet onsets. Note that we have $4M$ parameters for the EDS model and $4M+K$ for the PDDS model. Moreover, according to our initial assumption, we have $K \ll M$. Consequently, the total number of model parameters is almost the same for the two models. We fix the modeling orders to 20. On the top of figure 2, we have represented the original signal. On the middle of figure 2, we note the pre-echo phenomenon and the weak dynamic onset for the EDS modeling. On the bottom of figure 2 and for the three methods, we can point out the total absence of pre-echo and the great reproduction of the onset dynamic. Clearly, the PDDS model outperforms the EDS approach.

3.2. Perturbation of the estimated time-delay

Hereafter, we study the robustness of the PDDS-D and PDDS-MC algorithms to a perturbation Δ_t of the time-delay according to $t_1 + \Delta_t$ with $\Delta_t = \{-10, \dots, 10\}$ on the castanet onset signal. The estimated time-delay t_1 for the castanet signal equals 223 samples. On figure 3-a, we can see that the PDDS-D algorithm is the most robust algorithm, especially for time-delay under-estimation. The PDDS-MC1 is the least robust to the time-delay variation in the context of this simulation. The PDDS-MC2 shows intermediate robustness.

4. ALGORITHMIC COMPLEXITY AND CHOICE OF THE ALGORITHM

The complexity of the EDS algorithm can be evaluated to $O(NM^2)$ if we use an iterative processing of the SVD [11]. The

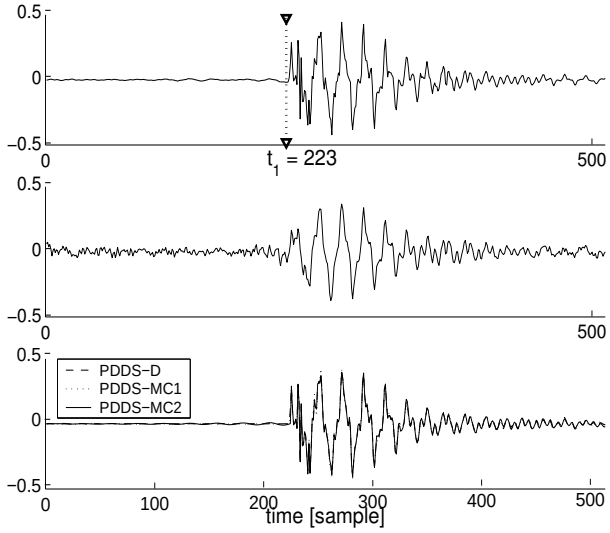

Figure 2: *(top) castanets onset (normalized amplitude), (middle) 20-EDS modeling, (bottom) 20-PDDS modeling.*

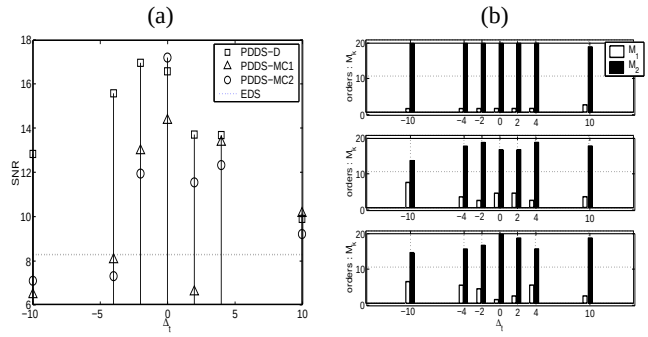

Figure 3: *(a) SNR with respect to the time-delay variation, (b) partial orders (top) PDDS-D, (middle) PDDS-MC1, (bottom) PDDS-MC2.*

complexity of the PDDS-MC1 is similar to the EDS one. The computational cost of the PDDS-D algorithm can be evaluated to $O(\sum_k B_k M_k^2)$ and $O(B_\nu M^2)$ for the PDDS-MC2.
From the simulation section, we conclude that the PDDS-D and PDDS-MC algorithms are well adapted to the transient audio modeling problem. Note that the allocation procedure for the PDDS-D algorithm is based on some empirical considerations on the "nature" of transient audio signal. Conversely, in the context of the PDDS-MC algorithms, the partial orders estimation is automatic since it is essentially a simple "re-allocation".
To conclude, we can say : for the true time-delay estimated, the PDDS-MC2 is the most efficient method since it has the lowest computational cost for slightly higher performance (see figure 3-a for $\Delta_t = 0$). In case of errors in the time delay estimation, we choose the PDDS-D method since this method presents the best trade-off between complexity and robustness.

5. CONCLUSION

In this paper, we have introduced an efficient non-stationary model for the transient compact representation problem. This model is an evolution of the DDS model introduced in the general context of signal modeling. This approach uses *a piori* information on percussive audio signal, *i.e.*, an audio transient signal can be seen as a set of damped sinusoids with a single time-delay. This natural consideration leads to propose the PDDS model and three high-resolution estimation methods. Finally, we show that the PDDS approach outperforms the EDS approach with an identical total number of modeling parameter on a typical transient audio signal. This conclusion is confirmed by intensive and informal listening tests.

6. REFERENCES

[1] B. Edler and H. Purnhagen, "Parametric audio coding", *Proc of the 5th International Conference on Signal Processing (ICSP 2000)*, Beijing, August 2000.

[2] T. Painter and A. Spanias, "Perceptual Coding of Digital Audio", *Proc of the IEEE*, Vol. 88, No 4, April 2000.

[3] K. Abed-Meraim, A. Belouchrani, Y. Hua and A. Mansour, "Parameter Estimation of Exponentially Damped Sinusoids using Second Order Statistics", *Proc. of Europ. Signal Processing Conf. (EUSIPCO 98)*, September 1998.

[4] R. Boyer, S. Essid and N. Moreau, "Dynamic temporal segmentation in parametric non-stationary modeling for percussive musical signals", *Proc. of IEEE Int. Conf. on Multimedia and Expo (ICME 02)*, August 2002.

[5] R. Boyer and J. Rosier, "Iterative method for Harmonic and Exponentially Damped Sinusoidal model", *Proc. of 5th Int. conf. on Digital Audio Effects (DAFx02)*, September 2002.

[6] R.J. McAulay and T.F. Quatieri, "Speech analysis & synthesis based on a sinusoidal reprsentation", *IEEE Trans. on ASSP*, Vol. 34, No. 4, August 1986.

[7] R. Boyer and K. Abed-Meraim, "Audio transients modeling by Damped & Delayed Sinusoids (DDS)", *Proc. of IEEE Int. Conf. on Acoustic, Speech and Signal Processing (ICASSP 02)*, May 2002.

[8] J. Nieuwenhuijse, R. Heusdens and E.F. Deprettere, "Robust Exponential Modeling of Audio Signal", *Proc. of Int. Conf. on Acoustic, Speech and Signal Processing (ICASSP 98)*. Vol. 6 , 1998.

[9] S.Y. Kung, K.S. Arun and D.V. Baskar Rao, "State-space and singular-value decomposition-based approximation methods for harmonic retrieval problem", *J. Opt. Soc. Am.*, 73(12):1799-1811, December 1983.

[10] J. Kliewer and A. Mertins, "Audio Subband Coding With Improved Representation Of Transient Signal Segments", *Proc. of Europ. Signal Processing Conf. (EUSIPCO 98)*, September 1998.

[11] R. Badeau, R. Boyer and B. David, "EDS parametric modeling and tracking of audio signals", *Proc. of 5th Int. conf. on Digital Audio Effects (DAFx02)*, September 2002.