DYNAMIC TEMPORAL SEGMENTATION IN PARAMETRIC NON-STATIONARY MODELING FOR PERCUSSIVE MUSICAL SIGNALS

Rémy Boyer, Slim Essid, and Nicolas Moreau

ENST, Dept. of Signal and Image Processing
46, rue Barrault - 75634 Paris cedex 13
{boyer,essid,moreau}@tsi.enst.fr

ABSTRACT

An audio signal parametric modeling scheme is proposed that permits higher performance for representing strong sound transients. The Exponentially Damped Sinusoids (EDS) model is considered in association with a high resolution parameter estimation approach. Such a technique is well adapted to almost every audio signal but is unfortunately not efficient when dealing with signals presenting strong temporal variations, such as percussive music signals, and causes pre-echo artifacts and weak onset dynamic reproduction which are prejudicial to listening. A system, based on the EDS model, has been developed with a transient detector and dynamic time segmentation and modeling that allows to overcome such artifacts.

1. INTRODUCTION

Audio coding has been very active over the last fifteen years. Number of recommendations, within the UIT-T and the ETSI, and normalization within the ISO/MPEG have been made. MPEG-4 audio has recently issued a call for proposals [1] to set up a high-quality audio coder for music and speech at a target bit-rate of 24 kbits/s for the 20 Hz-15 kHz band aiming at better quality Internet streaming over the PSTN (Public Switched Telephone Network).

Audio signals presenting a stressed non-stationary character are, by nature, difficult to model. Several representations have been adopted in previous work utilizing fast time varying amplitude parameters such as Gabor models [3], Damped & Delayed Sinusoids [4] and Exponentially Damped Sinusoids (EDS) [5]. Such non-stationary models can be considered as a generalization of the stationary sinusoidal model [12].

The utilization of EDS by the audio community is quite recent and promessing [3, 6, 7, 8], since it allows sparse representations of the largest number of audio signals. Nevertheless, most effort has been focused on EDS-based speech modeling.

Towards a better representation of strong transient character signals, we have looked up for an improvement of EDS modeling for the non-stationary context of percussive signals, typically, beats, drums, castanets, gongs, which are of great perceptual importance in music. This work is a fundamental unity of a larger framework of a parametric audio coder (speech and music) at a target bit-rate smaller than 30 kbit/s for a 32 kHz sampling frequency.

Our contribution takes inspiration from the ISO-MPEG-1 layer III [2] work since parametric modeling is performed over variable length windows, switching from long to short ones under special conditions described hereafter. Moreover, a dynamic model order fitting is achieved. The hole system is driven by a powerful onset detector based on a perceptual filter bank [10].

2. THE PARAMETRIC NON-STATIONARY MODEL AND THE HIGH RESOLUTION (HR) METHOD

2.1. The Exponentially Damped Sinusoidal (EDS) model

For the audio signal $s(n)$ analysis, with $n = 0, \dots, N-1$, we define the Exponentially Damped Sinusoidal (EDS) model as

$$s_M(n) = \sum_{m=1}^{M} a_m e^{d_m n} \cos(\omega_m n + \phi_m) \qquad (1)$$

or

$$\boldsymbol{s}_M = \frac{1}{2}\left\{\sum_{m=1}^{M} (\alpha_m z_m^n + \alpha_m^* z_m^{n*})\right\}_{0 \le n \le N-1} \in \mathbb{R}^{N\times 1} \qquad (2)$$

where M is the modeling order, $\alpha_m = a_m e^{i\phi_m}$ is the complex amplitude and $z_m = e^{d_m + i\omega_m}$ is the complex pole. We, also, note a_m the m-th real amplitude, d_m the m-th real damping factor, ω_m the m-th angular frequency and ϕ_m the m-th initial phase belonging to $[0, 2\pi[$ and let $D = 2M$. $\boldsymbol{s}_M$ is determined on the basis of a non-linear Least-Squares criterion on the audio signal $\boldsymbol{s}$ with respect to the model parameters.

2.2. The High-Resolution (HR) method

The poles $\{z_m, z_m^*\}_{1 \le m \le M}$ of the signal (2) are calculated by exploiting the row(column)-shift invariance of the dominant space basis vectors. Let $\mathcal{S}_D$ the dominant space basis and let $\mathcal{H}_L(\boldsymbol{s})$ the $K \times L$ real Hankel matrix as defined in [5]. Note that $K + L = N - 1$ and this matrix is full rank ($= \min(K, L)$). An optimal choice is $L = N/2$ [5]. Then, $\mathcal{S}_D = \mathcal{R}(\boldsymbol{U}^{(D)})$ according to

$$\{z_1 z_1^* \dots z_M z_M^*\} \approx \lambda_D \left\{(\boldsymbol{U}_{\downarrow}^{(D)})^{\dagger} \boldsymbol{U}_{\uparrow}^{(D)}\right\} \qquad (3)$$

where $(.)^{\dagger}$ is the pseudo-inversion operation, $\mathcal{R}(.)$ denotes the range space and $\boldsymbol{U}_{\downarrow}^{(D)}$ (respectively $\boldsymbol{U}_{\uparrow}^{(D)}$) is the matrix $\boldsymbol{U}^{(D)}$ where the last (respectively the first) row has been deleted. $\lambda_D\{.\}$ is the set of the D eigen-values. The matrix $\boldsymbol{U}^{(D)}$ is explicitly calculated through the SVD factorization applied to $\mathcal{H}_L(\boldsymbol{s})$ according to $\mathcal{H}_L(\boldsymbol{s}) = \sum_{\ell=1}^{\min(K,L)} \sigma_\ell \boldsymbol{u}_\ell \boldsymbol{v}_\ell^T$ with $\sigma_\ell \ge \sigma_{\ell+1}$ where $\{\sigma_\ell\}$ is the singular values set, $\{\boldsymbol{v}_\ell\}$ (respectively $\{\boldsymbol{u}_\ell\}$) is the right (left) singular vectors set. By truncating the matrix $\boldsymbol{U}$ according to the

0-7803-7304-9/02/$17.00 C2002 IEEE

expression $\boldsymbol{U}^{(D)} = \boldsymbol{U}\boldsymbol{T}_D$ where $\boldsymbol{T}_D$ is a D first column selection matrix, we obtain the representation of the dominant space basis. Finally, for $m = 1, \ldots, M$, the m-th angular frequency and m-th the damping factor are deduced using $\omega_m = \Im m\{\log z_{2m-1}\}$ and $d_m = \Re e\{\log z_{2m-1}\}$. Note that a subspace separation procedure can be used to improve the angular frequency and damping factor estimation [9].

The complex amplitudes parameters $\{\alpha_m\}$ are determined by resolving the linear LS criterion $\min_{\boldsymbol{\alpha}} \|\boldsymbol{s} - \boldsymbol{C}(\boldsymbol{z})\boldsymbol{\alpha}\|_2^2$ where $\boldsymbol{\alpha} = \frac{1}{2}(\alpha_1\, \alpha_1^* \ldots \alpha_M \alpha_M^*)^T$, $\boldsymbol{C}(\boldsymbol{z}) = [\boldsymbol{c}_1\, \boldsymbol{c}_1^* \ldots \boldsymbol{c}_M \boldsymbol{c}_M^*]$ and $\boldsymbol{c}_m = \left(1\, z_m \ldots z_m^{N-1}\right)^T$. The solution to this criterion is $\boldsymbol{\alpha} = \boldsymbol{C}(\boldsymbol{z})^\dagger \boldsymbol{s} = \left(\boldsymbol{C}(\boldsymbol{z})^H \boldsymbol{C}(\boldsymbol{z})\right)^{-1} \boldsymbol{C}(\boldsymbol{z})^H \boldsymbol{s}$. Hence, for $m = 1, \ldots, M$, the m-th real amplitude and m-th initial phase are $a_m = 2|\alpha_{2m-1}|$ and $\phi_m = \Im m\{\log \alpha_{2m-1}\}$. Hereafter, this algorithm associated with the EDS model will be referred to as HR-EDS.

3. EDS MODEL LIMITATIONS

With EDS, a loss of modeling quality is observed whenever a transient does not start at the beginning of the analysis segment. Compared to the case where a transient starts at the beginning of a segment, the number of sinusoids needed to model it with a certain quality increases considerably [6]. Signal to Noise Ratio (SNR) decrease as a function of the distance τ between the "beginning" of the onset and the beginning of the 512 samples length segment (i.e 16 ms) with a 35-sinusoid model is shown on figure 2 on a castanets sample. On figure 3, we can point out the EDS model defects (modeling artifacts) materialized in two aspects : pre-echo and bad attack dynamic reproduction. The human auditory system is particularly sensitive to such modeling artifacts which are to be minimized.

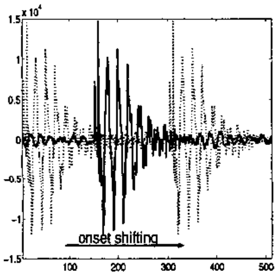

Figure 1: Onset shifting for $\tau = \{0, 150, 300\}$

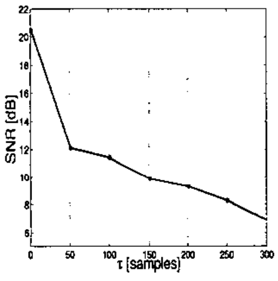

Figure 2: SNR for $\tau = \{0, \ldots, 300\}$

Basically, the pre-echo phenomenon can be described by the apparition of a significant amount of energy before the sound attack, *i.e* in the weak energy region of the signal, as well as a bad reproduction of the onset dynamic. As a result, the percussive character is no longer perceived during listening. One way to remedy this is to ensure onsets are modeled through a process where they are always positioned at the very beginning of analysis windows. Towards this purpose, an onset detector has been developed based on [10], and a dynamic windowing and modeling of the analyzed signal has been designed.

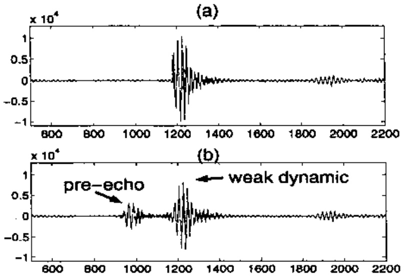

Figure 3: Pre-echo phenomenon and weak onset dynamic on Castanets onset modeling; (a) original signal ; (b) modeled signal with $M = 35$

4. TRANSIENT EXTRACTION AND DYNAMIC TIME SEGMENTATION

Modeling strong transient signals is a difficult matter since the abrupt temporal variations of the waveform are not easy to represent even with non-stationary models as EDS. In fact, whenever the beginning of the onset is "far" from the initial analysis instant, pre-echo is generated and a weak signal dynamic is reproduced. To overcome these defects, a sound onset detector is used that allows to locate attacks instants, thus, permitting the use of a dynamic time windowing. Whenever an onset is detected "far" from the initial analysis time, a shorter window is used that limits pre-echo. A better reconstruction of the attack is accomplished by means of a dynamic repartition of the model orders around this transient. Figure 4 presents a block diagram of the modeling system.

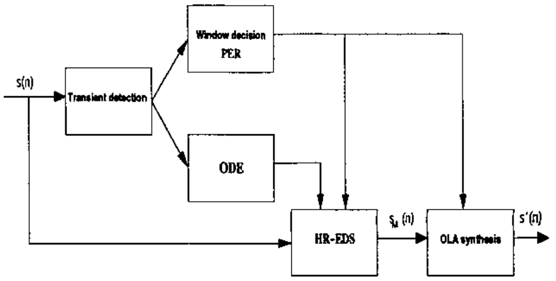

Figure 4: Modeling system block diagram, PER : Pre-Echo Reduction, ODE : Onset Dynamic Enhancement

4.1. The transient detector

The transient detector that is used is based on [10] and is here briefly described (see figure 6).

1. $s(n)$ signal filtering with a critical band filter bank [3] $\{h_k(n)\}$, $k = 1, \ldots, 24$ in the frequency range 0-16 kHz for a 32 kHz sampling frequency (see figure 5);

2. sub-band amplitude envelopes $A_k(n)$ calculations by means of a low-pass filtering with a 100 ms half-Hanning window;

3. computations of the Relative Difference Functions (RDF) of $A_k(n)$, onset candidates correspond to the maximum rising slopes i.e the dominant RDF values;
4. onset candidates loudness computations on the basis of the loudness model in [11];
5. selection of the loudest candidates in a 50 ms neighborhood.

This system, initially proposed in the rhythm analysis context, has been tested on several audio sample signals and provided satisfactory results.

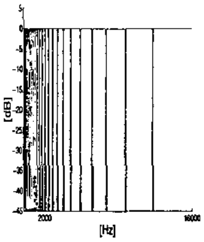

Figure 5: Filter bank freq. response

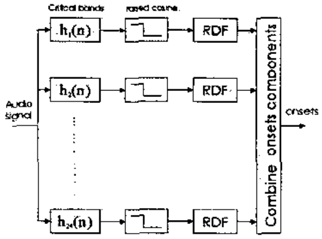

Figure 6: Transient detector block diagram

4.2. Dynamic time-segmentation and modeling

4.2.1. Pre-Echo Reduction (PER)

The signal analysis/synthesis is Overlap-Add based. After the onset detection, two window types can be used : long windows or short windows. HR-EDS modeling over long windows is referred to as HR-EDS(L), while modeling over short windows is noted HR-EDS(S). Long windows are used to model quasi-stationary and soft transients parts of the signal. Note that small overlap can be kept between long analysis windows (12.5%) without any loss of performance since the considered signal is modeled by HR-EDS in a satisfactory manner yielding non significant modeling discontinuities between successive long frames. Short windows are used for modeling the onsets. However, an overlap of 50% is maintained between these windows. This ensures the onset is fitted at the beginning of one short window. Therefore, the overlap between long windows can be the same as the one between short windows which prevents from using transition windows and allows simple implementations and less overhead. When a transient is detected, a decision must be taken whether to switch to a short window or not. In fact, if the onset is estimated at a position smaller than the overlap length (see figure 7), it is not necessary to use short windows and HR-EDS(L) is still appropriate. However, if the onset is detected further, the long window is subdivided into short overlapping windows over which the signal is successively modeled (see figure 8).

4.2.2. Onset Dynamic Enhancement (ODE)

Further improvement can be achieved by using a dynamic modeling order around the onset as it will be discussed in section 6. In effect, a transient can be decomposed into three segments in terms of energy : a first low energy part, followed by a strong dynamic part which is actually the attack and finally a progressive energy decreasing part. Hence, it seems natural to have a varying model order with respect to the considered transient segment. The allocation strategy consists in dedicating half of the parameters to the attack segment (2 successive windows) and the rest to the other segments.

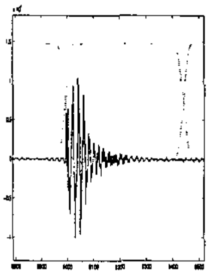

Figure 7: Long window

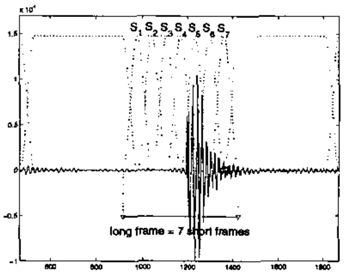

Figure 8: Dynamic windowing of the original signal

5. HR APPROACH JUSTIFICATION

5.1. HR Vs peak-picking in Spectral Modeling Synthesis

In [12] sinusoidal model parameters are determined through a peak picking operation based on the maximization of the Short-Time Fourier Transform (STFT) module. In this work, a high-resolution estimation approach, as described in section 2 has been preferred. In fact, the HR method is able to provide a good estimation of the parameters over a low number of samples. The use of short-time analysis windows is then possible without any important loss of estimation performance in the frequency domain which allows a good time resolution. By contrast, in a classical STFT analysis, the window size alters the frequency resolution and a short-time window analysis is not sufficiently accurate. For example, on 128-samples frames, the resolution is only $f_e/N = 250$ Hz.

5.2. HR Vs Matching-Pursuit (MP)

In [3], Matching Pursuit on a redundant EDS family is described. We have preferred the HR approach to such an iterative algorithm since the EDS family discretization for finite dictionary building, intimately conditions modeling performance. A too sharp discretization leads to a high computational cost. On the contrary, a rough discretization leads to less efficient modeling. In the HR approach, model parameters are not bounded by construction.

5.3. Computational complexity and fast algorithm

In [13], a fast algorithm is presented for determining the Signal space that is based on an iterative approach called OI-QR and the Hankel structure of the matrix $\mathcal{H}_L(s)$. It is shown that it is possible to achieve very important complexity reduction for an identical time waveform modeling. Moreover, the processing of 32-ms (1024 samples) frames demands, approximately, 21 s for the HR-EDS algorithm based on the SVD versus 0.2 s for the HR algorithm based on the OI-QR processed on a Pentium III computer under MatLab 5.3. So, the mean execution time to real time ratio, is around 669 for the HR-EDS algorithm based on SVD and 6.5 for the HR algorithm based on OI-QR.

6. SIMULATION ON TRANSIENT SIGNALS

In the following, dynamic window based modeling (HR-EDS(L/S)) is compared to fixed length window modeling (HR-EDS(L)). The windows are modified Hanning's satisfying the perfect reconstruction condition as shown on figure 8. The length of the long windows, respectively, short windows is, respectively, 512 and 128 samples, *i.e* 16 and 4 ms for a 32 kHz sampling frequency. This implies a long window is replaced by 7 short overlapping windows when switching to transient modeling. Note that, shorter windows would lead to non sparse representations. Model orders have been set to $M_L = 35$ for HR-EDS(L) and $M_S = 5 = M_L/7$ for HR-EDS(S) thus ensuring a constant number of parameters over different length frames. Once a transient has been detected and short windows have been decided, we are able to locate it in two successive windows due to their construction. Greater model order is then used over these two windows. For the modeling example shown on figure 9-(d), the orders are set as follows : $M_{S_4} + M_{S_5} = [M_L/2]$. The original signal, 17 ms of a Castanets signal, is presented on figure 8 with the two windowing possibilities. On figure 9-(b), fixed window-length based HR-EDS(L) modeling is presented. Modeling with transient detection and window switching is shown on figure 9-(c). Finally, HR-EDS(L/S) modeling with onset dynamic enhancement is presented on figure 9-(d). One can easily figure out the presence of strong pre-echo on figure 9-(b) which is eliminated thanks to the PER approach as shown on figure 9-(c). Moreover, it can be depicted from figure 9-(d) that a better onset dynamic reproduction is achieved by means of ODE. This approach has been validated on a set of percussive music signals as it has turned out to be successful through informal listening tests.

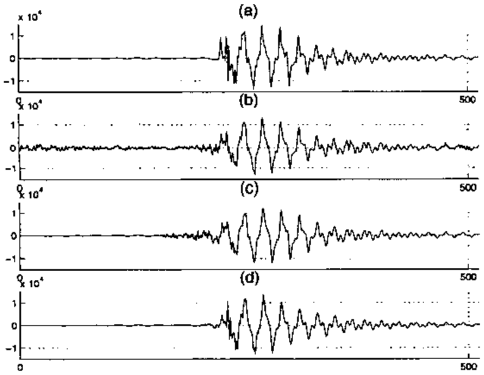

Figure 9: (a) original signal (b) HR-EDS(L) (c) HR-EDS(LS) (d) HR-EDS(LS) with onset dynamic enhancement

7. CONCLUSION

In this communication, we have presented a parametric modeling technique based on HR-EDS and well adapted to strong transient representations, which is fundamental for musical signals. The key to it is the use of a psycho-acoustically relevant onset detector associated with a dynamic time-segmentation and modeling approach. Fitting appropriate variable length windows, pre-echo has been drastically reduced (PER) while allocating model orders in a dynamic manner has permitted an enhancement of the onset dynamic (ODE). The technique has been tested on a wide variety of audio signals and listening tests have allowed to work out the superiority of the system by comparison to HR-EDS(L) at a similar bitrate of 30 kbit/s after specific quantization that will be described in future communications.

8. REFERENCES

[1] ISO-MPEG, *Call for proposals for new tools for audio coding*, ISO/IEC JTC1/SC29/WG11 MPEG2001/N3793, January 2001

[2] K. Banderburg and G. Stoll, "ISO-MPEG-1 Audio : a generic standard for coding of high-quality digital audio", *JASA*, Vol. 42, October 1994

[3] M. Goodwin, *Adaptive Signal Models: Theory, Algorithms, and Audio Applications*, PhD thesis, University of California, Berkeley, 1997.

[4] 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*, May 2002

[5] Y. Hua and T.K. Sarkar, "Matrix Pencil Method for Estimating Parameters of Exponentially Damped/Undamped Sinusoids in Noise", *IEEE Trans. on ASSP*, Vol 38 Issue: 5, May 1990

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

[7] J. Jensen, S.H. Jensen and E. Hansen, "Exponential sinusoidal modeling of transitional speech segments.", *Proc. of IEEE Int. Conf. on Acoustic, Speech and Signal Processing*, 1999

[8] R. Boyer, S. Essid and N. Moreau, "Non-stationary signal parametric modeling techniques with an application to low bitrate audio coding" *Proc. of IEEE Int. Conf. Signal Processing*, August 2002

[9] S. Van Huffel, "Enhanced resolution based on minimum variance estimation and exponential data modeling", *Signal Processing*, vol.33, no.3, 1993

[10] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge", *Proc. of IEEE Int. Conf. on Acoustic, Speech and Signal Processing*, 1999.

[11] B. Moore, B. Glasberg and T. Baer, "A Model for the Prediction of thresholds, Loudness and Partial Loudness", *J. Audio Eng. Soc.*, Vol. 45, No. 4, pp. 224-240. April 1997

[12] X. Serra and J. Smith III, "Spectral Modeling Synthesis : A Sound System Based on a Deterministic plus Stochastic Decomposition", *Computer Music Journal*, Vol. 14, No. 4, Winter 1990.

[13] R. Boyer, "Fast algorithm and non-stationary model for high-resolution audio signal modeling", ENST internal report : `http:/www.tsi.enst.fr/ boyer/`, 2002