Non-stationary signal parametric modeling techniques with an application to low bitrate audio coding

Rémy Boyer, Slim Essid and Nicolas Moreau
ENST, Department of Signal and Image Processing
46, rue Barrault, 75634, Paris Cedex 13
{ boyer/essid/moreau }@tsi.enst.fr

ABSTRACT

Low bitrate audio coding often relies on Fourier representation despite its limitations for transient signal modeling. This study proposes alternative decompositions and expansion strategies that lead to more accurate modeling. Two classes of methods are considered, subspace decomposition methods, and atomic decomposition methods and their performances are compiled to propose an audio modeling scheme amenable to low bitrate coding.

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. MPEG4 audio has recently issued a call for proposals to set up a high-quality audio coder for music and speech at a target bitrate of 24 kbits/s for the 20 Hz-15 kHz band aiming at better quality internet streaming. Our contribution builds upon the sinusoidal modeling philosophy initiated in the 1980s for speech coding in the telephony band [1]. Such modeling was then extended to music signal in proposing a musical sound analysis/synthesis based on a "Sinusoidal+Noise" approach [2]. More recently, this approach has been adopted for the low-bitrate coder MPEG4-HILN. Modifications of the "Sinusoid+Noise" model have been proposed to provide a better representation of transients (sound onsets or fadings) yielding the "Sinusoidal+Transient+Noise" model [9].
In this paper, we present an analysis of modeling techniques amenable to an appropriate representation of both stationary and transient signals. In this optics, two approaches have been investigated. The first one is based on an Exponentially Damped Sinusoid (EDS) parametric representation of the signal whose parameters are determined by means of a High-Resolution (HR) approach. This is achieved through subspace methods [3] that are preceded by a subspace separation procedure [4]. The second approach exploits an iterative algorithm proceeding to successive projections of the signal to be modeled on Gabor atoms of a pre-defined dictionary [5] (Matching Pursuit : MP and Orthogonal Matching Pursuit : OMP). It will be deduced that a satisfying approach would consider to split the analysis into three components : "strong transient", "soft transient + sinusoidal" and "noise".

The outline of this paper is the following. In Section 2 we present an overview of HR and MP methods as well as the waveforms used for signal representation. Simulation results and discussion are presented in Section 3. We finally make considerations concerning the computational cost and suggest a modeling block diagram allowing low bitrate coding.

2 Parametric Modeling and Atomic Decompositions

2.1 Real expansions on complex waveform collections

It is assumed that the audio signal can be modeled as a sum of two components : a deterministic component plus a "stochastic" component [2]. Let, then, $\mathcal{O}$ the real observation space such that $\dim(\mathcal{O}) = L$ and let $\boldsymbol{s} = \{s(n)\}_{n=0}^{N-1}$ an N-sample audio signal such that $\boldsymbol{s} \in \mathcal{O}$. $\mathcal{O}$ can be decomposed into two subspaces : $\mathcal{S}$ the $2M$-dimension signal subspace and $\mathcal{B}$ the noise subspace such that $\mathcal{O} = \mathcal{S} \oplus \mathcal{B}$. Thus $\boldsymbol{s} = \boldsymbol{s}_M + \boldsymbol{r}$ with $\boldsymbol{s}_M \in \mathcal{S}$ and $\boldsymbol{r} \in \mathcal{B}$ and one can write

$$\boldsymbol{s}_M = \frac{1}{2}\sum_{m=1}^{M}(\alpha_m \boldsymbol{g}_m + \alpha_m^* \boldsymbol{g}_m^*) \in \mathbb{R}^{N\times 1} \quad (1)$$

where

$$\boldsymbol{g}_m = \left\{g_m(n)e^{i\omega_m n}\right\}_{n\leq 0\leq N-1} \quad (2)$$

and $\alpha_m = a_m e^{i\phi_m}$, ω_m is the angular frequency, a_m is the real amplitude, ϕ_m is a phase term in the range $[0, 2\pi[$ and $(\boldsymbol{g}_1, \boldsymbol{g}_1^*, \ldots, \boldsymbol{g}_M, \boldsymbol{g}_M^*)$ is a basis of $\mathcal{S}$ eventually orthonormal.

2.2 Representation Model

Sinusoidal modeling has proven its efficiency in modeling harmonic or quasi-harmonic signals that present slow time variations [2]. However, such modeling provides poor performance when representing transient signals, typically signal onsets or fadings, which are, by nature, localized both in time and frequency. It thus seemed interesting to produce a signal representation based on waveforms allowing a better modeling of fast time varying signals.

2.2.1 EDS and Fourier Model

We define the Exponentially Damped Sinusoid (EDS) model of order M, M-EDS, by letting $g_m(n) = e^{d_m n}$ in

0-7803-7488-6/02/$17.00 © 2002 IEEE.

expression (2), where d_m is the m-th real damping factor. Note that if $\forall m,\ d_m = 0$ the model is a Fourier one.

2.2.2 Gabor Model

A Gabor atom-collection can be generated by dilating, translating and modulating a same Gaussian window $g(n) = e^{-\pi n^2}$. For each scale factor $c_m > 0$ and translation parameter u_m, the Gabor model can be defined by letting $g_m(n) = g\left(\frac{n-u_m}{c_m}\right)$ in expression (2). With the notation $\gamma_m = (c_m, u_m, \omega_m)$, the normalized Gabor atom is $g_{\gamma_m}(n) = K_c g\left(\frac{n-u_m}{c_m}\right) e^{i\omega_m n}$ where K_c ensures $\|g_{\gamma_m}\|_2 = 1$.

2.3 Expansion Strategy

2.3.1 High-Resolution Method

The M-EDS model can be represented by the matrix formalism

$$s_M(n) = 1_{2M} Z^n \alpha \tag{3}$$

where $1_{2M} = (1 \dots 1) \in \mathbb{R}^{1\times 2M}$, $Z = \mathrm{diag}\{z_1, z_1^*, \dots, z_M, z_M^*\}$ and $\alpha = \frac{1}{2}(\alpha_1\ \alpha_1^* \dots \alpha_M \alpha_M^*)^T$ with $z_m = e^{d_m + i\omega_m}$ the m-th pole. We then build the $(N-L)\times(L+1)$ Hankel matrix $\mathcal{H}_L(s_M)$ and proceed to its factorization into two matrices $O_{(N-L)\times 2M}$ and $C_{2M\times(L+1)}$, respectively the observability matrix and the controlability matrix [3]. In addition, defining the row-shifting matrix operators, $(.)_\uparrow$ which deletes the first row, and $(.)_\downarrow$ which deletes the last row, one can easily figure out the row-shifting invariance property : $O_\downarrow Z = O_\uparrow$. This leads to compute $O_\downarrow^\dagger O_\uparrow$, where $(.)^\dagger$ denotes the pseudo-inverse, and find out its eigen values decomposition. In fact, the eigen-values set of $O_\downarrow^\dagger O_\uparrow$ is $\{z_1, z_1^*, \dots, z_M, z_M^*\}$. Let us now explicit the matrix O. We choose $O = UT_{2M} = [u_1 \dots u_{2M}]$ with T_{2M} a U $2M$ first-columns selection matrix, U being the left singular values matrix.

For real signals, truncating the Singular Values Decomposition (SVD) [11] is not so evident. This is why we use the Composite Property Mapping (CPM) pre-processing algorithm introduced in [4] on $H = \mathcal{H}_L(s)$ which allows us to separate $\mathcal{S}$ from $\mathcal{B}$. Defining a truncation operator $\mathcal{T}_{2M}(H) = \sum_{m=1}^{2M} \lambda_m u_m v_m^H$ as well as an averaging on anti-diagonals operator $\mathcal{M}(H)$, one can figure out an important property that is $\mathcal{M}\circ\mathcal{T}_{2M}(H) \approx s_M$. So, the CPM algorithm can be summed-up in

$$[\mathcal{H}_L \circ \mathcal{M} \circ \mathcal{T}_{2M}]^k(H) \xrightarrow{k\nearrow} \mathcal{H}_L(s_M) \tag{4}$$

which has a fast convergence rate. In other words, this iterative algorithm leads to solve

$$\arg\min_{s_M} \|H - \mathcal{H}_L(s_M)\|_F^2 \ \text{ with } \begin{cases} \mathcal{H}_L(s_M) \text{ having a Hankel structure} \\ \mathrm{rank}(\mathcal{H}_L(s_M)) = 2M \end{cases} \tag{5}$$

where $\|.\|_F$ is the Frobenius norm.

The next step consists in computing the real amplitudes and phases by solving the least-squares problem $\min_\alpha \|s - V\alpha\|_2^2$ where V is the $N \times 2M$ Vandermonde matrix whose general term is $\{e^{(d_m \pm i\omega_m)n}\}$. The solution to this problem is $\alpha = V^\dagger s$.

2.3.2 HR method Discussion

High-Resolution methods based on subspace decomposition are advantageous in the sense that they are not bound to the Fourier resolution limitation. This turns out to be particularly consistent whenever two spectral components of a complex sound are separated by a distance smaller than the frequency resolution. Also, when there is frequency sliding, this method can be efficient.

The second advantage is its ability to provide a good estimation of the parameters over a low number of samples. The use of very short-time analysis windows is then possible without any important loss of estimation performance which allows a good time resolution. By contrast, in a classical Fourier analysis, the window size alters the frequency resolution and a short-time window analysis is not sufficiently accurate.

2.3.3 Matching Pursuit and Orthogonal Matching Pursuit (MP & OMP)

Matching Pursuit is a greedy iterative algorithm introduced by Mallat & Zhang [5] that produces a signal decomposition over a collection of atoms chosen in a redundant $N\times D$ dictionary ($D \gg N$) such that $G = \{g_\gamma\}_{\gamma\in\Gamma}$ where Γ is the atoms index set. The principle had already been used in CELP speech coding, for excitation vectors selection [7]. The signal s is successively approximated in pursuit stages by orthogonal projections on "di-atomic molecules" built from G. In fact, real decompositions of the signal are obtained with an underlying complex atoms dictionary so as to avoid discretizing a supplementary phase parameter which would lead to an increased dictionary-size [6]. This is made possible thanks to conjugate subspace pursuit. At each pursuit iteration, should be selected a subspace of dimension 2 spanned by an atom and its conjugate in such a way that, at the m-th iteration, the algorithm looks for a "molecule", materialized by a $N\times 2$ matrix G_{γ_m} whose two columns consist of a dictionary vector and its conjugate which minimizes the norm of the residual $r_{m+1} = r_m - G_{\gamma_m}\alpha_m$ where α_m is a vector of two correlation coefficients. G_{γ_m} is then such that

$$G_{\gamma_m} = \arg\max_{\gamma\in\Gamma} |< G_\gamma, r_m >| . \tag{6}$$

The orthogonality constraint $< r_m - G_{\gamma_m}\alpha_m, G_{\gamma_m} >= 0$ implies a solution on the weights $\alpha_m = G_{\gamma_m}^\dagger r_m$ which yields[1]

$$r_{m+1} = r_m - 2\Re e(\alpha_m(1) g_{\gamma_m}). \tag{7}$$

Such a process leads to signal decompositions of the form $s_M = 2\sum_{m=1}^M \Re e(\alpha_m(1) g_m)$. A variation can be considered[2] that ensures that the residual r_m is orthogonal to every dictionary vector already selected in the $m-1$ previous iterations. This guarantees that MP stops after a finite number of iterations in a perfect signal reconstruction scheme [5]. The original algorithm is modified as follows :

[1] calculus details can be found in [6]

[2] here again the principal had been used in speech coding [7]

at iteration m we orthogonalize the vector g_{γ_m} with respect to the $m-1$ vectors $g_{\gamma_1}, \ldots, g_{\gamma_{m-1}}$ already selected by means of a Gram-Schmidt procedure thus having

$$g^{\perp}_{\gamma_m} = \frac{g_{\gamma_m} - P_{\mathcal{V}_{m-1}} g_{\gamma_m}}{\left\| g_{\gamma_m} - P_{\mathcal{V}_{m-1}} g_{\gamma_m} \right\|_2} \quad (8)$$

where $P_{\mathcal{V}_{m-1}}$ is the orthogonal projector on the subspace $\mathcal{V}_{m-1} = \text{span}\{g_{\gamma_1}, \ldots, g_{\gamma_{m-1}}\}$, and we project back r_m on $g^{\perp}_{\gamma_m}$.

2.3.4 MP & OMP Discussion

MP is an algorithm that can be regarded as a general problem of decomposition of an M-dimensional optimization problem into M one-dimensional optimization problems. This approach is, thus, by construction, sub-optimal. Yet, it provides the capability of making no *a priori* assumptions on the signal to be modeled and allows the use of a wide variety of waveforms. Waveforms such as wavelets, chirps, chirplets, EDS, etc. can be added in the dictionary. However, the waveforms parameter-discretization step conditions its performance. Besides, MP has the ability to compensate at advanced pursuit stages for the modeling artifacts created at first stages. This enables MP to be more stable than parametric modeling.
Note that OMP increases the residual norm convergence speed [5]. However, it is only meaningful for redundant dictionaries, i.e consisting of a linearly dependent vector-family. This implies a loss of orthogonality and thus requires a re-orthogonalization procedure to obtain an orthogonal basis of the signal subspace.

3 Simulations on real signals

Modeling techniques described above have been implemented and intensive listening tests have been carried out on a wide variety of audio signals allowing to compare HR and MP modeling performance with Fourier, EDS and Gabor waveforms. Even though the Signal to Noise Ratio (SNR) cannot be trusted in audio coding context, it has provided results confirming the listening tests results and is merely used here as an illustration to our conclusions. SNR measures on three typical signals are thus presented : a speech signal (harmonics + soft transients) shown on figure 1-a, a castanets signal (strong transients) on figure 1-b and a bells signal (harmonics + transients) on figure 1-c. Note that these methods are compared according to a bitrate criterion in a way that ensures $\rho^{(HR)} \approx \rho^{(MP)}$ where ρ is the estimated bitrate and is computed using the formula $\rho = \frac{f_e}{\beta N} \mathcal{N}_{bit} M$ [bit/s] with f_e, the sampling frequency, β an overlap parameter in the range $[\frac{1}{2}, 1]$ corresponding to $\frac{N}{2}$ down to 0 overlapping samples between successive analysis windows and $\mathcal{N}_{bit}$ the total number of bits needed to code the parameters of a waveform in the model. In these simulations, we make no distinction between MP and OMP since both algorithms provide comparable performance in terms of SNR on the tested signals and assuming M should be small. Therefore, we have chosen MP since it has been used for sinusoidal modeling in previous work [9] thus providing reference results. HR-Fourier has turned out to have poor performance which is why it has been removed from the simulations. Note that a "HR-Gabor" model is not possible since Gabor waveforms do not obey to an auto-regressive model of order M by opposition to EDS. Now this is necessary for the algorithm in [3] to work out. Additionally, MP/OMP-EDS is not tested for a given parameter-discretization would lead to a too huge dictionary and consequently to a high cost dictionary index coding in terms of bitrate. Thus, three models simulations are presented : MP-Fourier, HR-EDS and MP-Gabor on three typical signals.

Comments on figure 1-a : speech signal. On figure 1-d , it can be noticed that HR-EDS is the most successful method with an average SNR of 30 dB. MP-Fourier and MP-Gabor provide similar performance. This can be explained by the fact that the former take profit from a High-Resolution joint estimation scheme while the latter is an iterative dictionary approach which is conditioned by the parameters discretization.

Comments on figure 1-b : castanets signal. On this strongly non-stationary signal, MP-Gabor provides superior performance in terms of SNR relatively to the other methods (see figure 1-e). This is due to the fact that it utilizes a time delay parameter that enables it to better match the onset start. The other models are incapable of efficiently modeling short-time signals far from the beginning of the analysis window.

Comments on figure 1-c : bells signal. This signal is interesting as it mixes a strong transient (the bell onset) with a highly stationary signal. On figure 1-f, it can be observed that MP-Gabor provides a better SNR on the onset. On the contrary, HR-EDS has an SNR greater than MP-Gabor SNR and MP-Fourier SNR on the signal portion before and after the onset.

Conclusion. These results suggest a modeling scheme where HR-EDS represents the sinusoidal and soft transient part and MP-Gabor is used to model strong signal transients. This process has been implemented and intensive listening tests have confirmed the expected results.

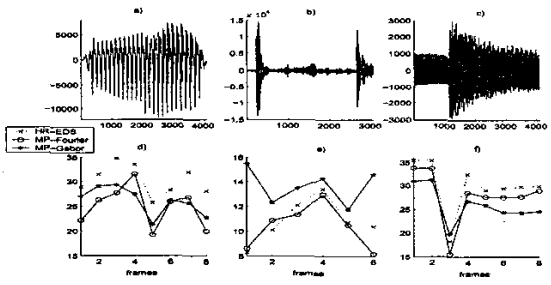

Figure 1: Time-domain signals (a,b,c) and SNR [dB] per frame (d,e,f) with $N = 512$

4 Computational cost considerations

The methods developed here present a high computational cost which is constraining for audio coding purposes. Therefore, complexity reduction techniques have been considered that allow fast modeling. These can summed-up as follows :

- For the MP algorithm, we use a fast search procedure running in local dictionaries with a tree structure as in [5]. We also, make an original use of the FFT algorithm to compute signal residuals correlations with dictionary atoms. These two approaches lead to a strongly reduced calculation time.
- For the HR method, a SVD is usually used to determine the signal subspace basis. This factorization is expensive and can be iteratively approximated [11] and limited to the $2M$ dominant singular vectors. This approach finds its justification in that we are only interested in the signal subspace basis. In a modeling application of small order M, this approach considerably reduces the algorithmic complexity.

On a Pentium III computer under MatLab 5.3, these techniques lead to a mean execution time to real time ratio of order 6. Further details are to appear in a future paper.

5 Coding Considerations and Conclusion

After several testing procedures, we propose the coding block diagram shown on figure 2. The audio signal is seen as a sum of four components coded in three successive stages : "soft transient + sinusoidal", "strong transient" and "noise". Soft transient and sinusoidal component are analysed with the HR-EDS model. The model parameters are coded as follows. Damping factors as well as phases are vector quantized and coded over 5 bits. A non-uniform scalar quantization is used for amplitudes leading to a 5 bit per amplitude coding. Frequencies are continued into trajectories over successive frames. Trajectory initiating frequencies are coded over 9 bits and continued frequencies are relatively coded over 4 bits. The first residual is then computed wich consists of noise plus strong transients. The strong transient component is modeled in a DCT (Discrete Cosine Transform) domain so that MP-Gabor models the transients in priority[3]. MP-Gabor parameters quantization is integrated into the iterative process as shown on figure 2. Correlation coefficients magnitudes and phases are coded similarly to HR-EDS amplitudes and phases. Indexes are coded over 10 bits. The last residual is modeled using a 20-order warped LPC filter [8] which turned out to provide very satisfactory results since the residual is not assimilabe to "pure" noise. Each filter coefficient is coded over 5 bits. The analysis/synthesis is frame-based. The frame-size is 1024 samples and the overlap is smaller than half the analysis frame-size. The synthesis is OverLap-Add-based. With a 32 kHz sampling frequency a 25 to 30 kbit/s bitrate has been reached. Listening results are encouraging since we have been able to notice, through informal testing, that the majority of the coded signals have acceptable quality.

[3] justification can be found in [9] and [10]

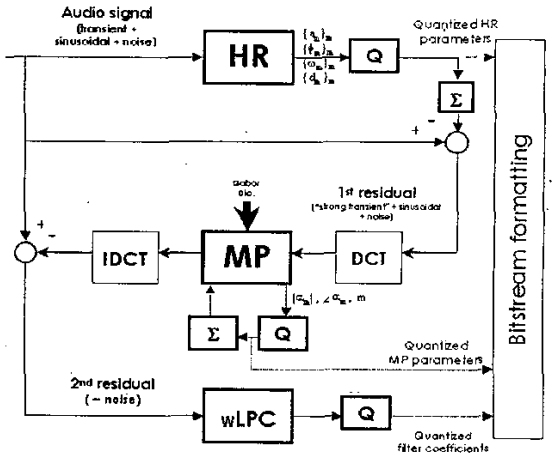

Figure 2: Coding block diagram

References

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

[2] X. Serra, 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.

[3] J. Nieuwenhuijse, R. Heusens, Ed. F. Deprettere, "Robust exponential modeling of audio signals", *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Vol. 6, 1998

[4] J.A. Cadzow, "Signal Enhancement - A Composite Property Mapping Algorithm", *IEEE Trans. on ASSP*, Vol. 36, No. 1, January 1988.

[5] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, second edition, 1999.

[6] M. Goodwin, M. Vertterli, "Matching Pursuit and Atomic Signal Models Based on Recursive Filter Banks", *IEEE Trans. on SP*, Vol. 47, No. 7, July 1999.

[7] N. Moreau, P. Dymarski, "Successive orthogonalizations in the multistage CELP coder", *Proc. Int. Conf. Acoust., Speech, Signal Processing*, 61-64,1992

[8] A. Harma, U.K. Laine, "A comparison of warped and conventional linear predictive coding", *IEEE Trans. on ASSP*, Vol. 9 Issue: 5, July 2001

[9] T. Verma, T. Meng, *A Perceptually Based Audio Signal Model with Application to Scalable Audio Compression*, PhD thesis, Stanford University, 1999.

[10] R. Boyer and S. Essid, "Transient modeling with a Frequency-Transform Subspace Algorithm and "Transient + Sinusoidal" scheme" *Proc. of IEEE Int. Conf. on Digital Signal Processing*, July 2002, *Accepted*

[11] G.H. Golub, C.F. Van Loan, *Matrix Computation*, North Oxford Academic, Oxford, second edition, 1983.