TRANSIENT MODELING WITH A FREQUENCY-TRANSFORM SUBSPACE ALGORITHM AND "TRANSIENT + SINUSOIDAL" SCHEME

Rémy Boyer, Slim Essid

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

Abstract: In this paper, we present an efficient modeling method for strong transient character audio signals. It is shown that the parametric non-stationary Exponentially Damped Sinusoids (EDS) model permits good performance for time domain modeling of quasi-stationary signals or "weak" transients. However, a decay in modeling performance is observed when dealing with highly non-stationary signals as in a variety of musical sound (various percussions, castanets, triangle, ...). The idea is then to process the signal in a well chosen frequency-transform domain in which the transient temporal characteristics are better modeled by EDS. As a result, better representations of the transient signal class are obtained with no pre-echo artifacts (energy before the attack) and a very good signal onset dynamic reproduction. Finally, an original "Transient+Sinusoidal" modeling scheme is proposed.

1. INTRODUCTION

Audio signals (speech and music) present a wide diversity. A rough classification would consider three main categories. The first would consist of stationary and quasi-stationary signals such as horn sounds or speech voiced sounds. The second would be the class of "weak" transients such as speech plosives for example and finally, the "strong" transient signals category. The first two classes of sound are well represented with parametric stationary models *i.e* sinusoids [1] or non-stationary models *i.e* Exponentially Damped Sinusoids (EDS) [2], [3], [4] which has been mostly used for EDS-based speech modeling. Now, it has been shown in [2], [5] on a castanets sound sample that a decay in modeling performance is observed with such "strong" transient signals. Modeling characteristic artifacts are then created with two effects. First, the apparition of a pre-echo signal *i.e* a distortion before the sound attack [6]. Second, the signal dynamic is badly reproduced. These phenomena appear to be very prejudicial to the auditory perception of the third category sound.

Based on the time-frequency duality principle, we propose to show that EDS modeling of strong transient signals in a well chosen frequency-transform domain enables pre-echo cancellation as well as reproducing the proper signal dynamic. We therefore, introduce the Frequency-Transform Subspace Algorithm based on the EDS model (FTSA-EDS).

The outline of the paper is the following. In section 2, we present the EDS model and a brief sum-up of the Subspace Algorithm (SA-EDS) for the model parameters determination. In section 3, we introduce time-frequency duality as well as the chosen transform. In section 4, we exhibit the FTSA-EDS algorithm and show its performance on a typical example of real strong transient signals. Finally, in section 5, an original "Transient + Sinusoidal" modeling scheme is presented.

2. THE EDS MODEL AND THE SUBSPACE ALGORITHM (SA)

2.1. The Non-stationary model : EDS signals

We define the Exponentially Damped Sinusoidal (EDS) model by

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

and

$$\hat{\boldsymbol{s}}(\boldsymbol{\alpha}, \boldsymbol{z}) = \frac{1}{2} \left\{ \sum_{m=1}^{M} (\alpha_m z_m^n + \alpha_m^* z_m^{n*}) \right\}_{0 \leq n \leq N-1} \in \mathbb{R}^{N \times 1} \tag{2}$$

where $D = 2M$ and 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 denote the m-th real amplitude by a_m, the m-th real damping factor by d_m, the m-th angular frequency by ω_m and the m-th initial phase belonging to $[0, 2\pi[$ by ϕ_m. We define the vectorial notations as follows : $\boldsymbol{\alpha} = (\alpha_1 \alpha_1^* \ldots \alpha_M \alpha_M^*)^T$ and $\boldsymbol{z} = (z_1 z_1^* \ldots z_M z_M^*)^T$.

2.2. The SA-EDS algorithm: model parameters processing

The D model parameters $\{\alpha_m, z_m\}$ are determined through the minimization of the following quadratic criterion

$$\arg\min_{\boldsymbol{\alpha}, \boldsymbol{z}} \|\boldsymbol{s} - \hat{\boldsymbol{s}}(\boldsymbol{\alpha}, \boldsymbol{z})\|_2^2 \tag{3}$$

where $\boldsymbol{s} \in \mathbb{R}^{N \times 1}$ is the signal to be modeled. A joint optimization with respect to $\{\boldsymbol{\alpha}, \boldsymbol{z}\}$ is not possible in practice. Thus, signal poles $\boldsymbol{z}$ are computed thanks to the line-shift invariance property of the signal basis vectors [7]. We then solve the quadratic criterion (3) with respect to $\boldsymbol{\alpha}$.

2.2.1. *Poles processing*

The poles are computed as follows.

1. Build $\boldsymbol{H} = \mathcal{H}_L(\boldsymbol{s})$ where $\mathcal{H}_L(\boldsymbol{s})$ defines the Hankel operator within $\mathbb{R}^{N \times 1} \to \mathbb{R}^{L \times L}$ (such that $2L = N - 1$) [7] to be $\mathcal{H}_L(\boldsymbol{s}) = [\boldsymbol{s}_1 \mid \ldots \mid \boldsymbol{s}_L]$ with $\boldsymbol{s}_\ell = (s(\ell) \ldots s(\ell + L - 1))^T$. Note that with no prior assumption on $\boldsymbol{s}$, this matrix is full rank $(= L)$.
2. Determine the Hankel matrix $\mathcal{H}_L(\hat{\boldsymbol{s}})$ of rank D[1] $(\leq L)$ that minimizes $\|\boldsymbol{H} - \mathcal{H}_L(\hat{\boldsymbol{s}})\|_F^2$, *i.e*, according to [8],

   $$[\mathcal{H}_L \circ \mathcal{M}_N \circ \mathcal{T}_D]^\eta(\boldsymbol{H}) \xrightarrow{\eta \nearrow} \mathcal{H}_L(\hat{\boldsymbol{s}}) \tag{4}$$

   where $\mathcal{M}_N(.)$, defined within $\mathbb{R}^{L \times L} \to \mathbb{R}^{N \times 1}$, is the averaging on anti-diagonals operator and $\mathcal{T}_D(.)$ defined within $\mathbb{R}^{L \times L} \to \mathbb{R}^{L \times L}$, is the rank reduction operator or "zero forcing" of the $L - D$ smallest singular values. Let η be the number of iterations and assume that $\eta = 0$

[1] In an audio compression application, the parameter D is chosen to reach a target bitrate.

0-7803-7503-3/02/$17.00 ©2002 IEEE

implies only the rank reduction operator is used. Then, $\mathcal{H}_L(\hat{\boldsymbol{s}}) \approx \mathcal{T}_D(\boldsymbol{H})$. Note that, the Hankel character is then not conserved. In practice, this approximation provides satisfactory results.

3. Extract the matrix of the left D dominant vectors $\boldsymbol{U} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_D]$ from $\mathcal{H}_L(\hat{\boldsymbol{s}})$.
4. Determine signal poles through

$$\boldsymbol{z} = \lambda_D \left\{ \boldsymbol{U}_{\downarrow}^{\dagger} \boldsymbol{U}_{\uparrow} \right\} \tag{5}$$

where $(.)^{\dagger}$ is the pseudo-inversion symbol, $\boldsymbol{U}_{\downarrow}$ (respectively $\boldsymbol{U}_{\uparrow}$) is the matrix $\boldsymbol{U}$ from which the last (respectively, first) line has been removed and $\lambda_D\{.\}$ is the set of the D eigen values. Finally, extract the angular frequencies $\{\omega_m\}$ and the damping factors $\{d_m\}$ from the poles $\{z_{2m-1}\}$, for $m = 1, \dots, M$.

2.2.2. *Complex amplitudes processing*

The criterion (3) has to be solved for $m = 1, \dots, M$ according to

$$\boldsymbol{\alpha}_m = [\boldsymbol{c}_m(\boldsymbol{z}) \ \boldsymbol{c}_m^*(\boldsymbol{z})]^{\dagger} \boldsymbol{s} \tag{6}$$

where $\boldsymbol{c}_m(\boldsymbol{z}) = (1 \ z_m \dots z_m^{N-1})^T$ and $\boldsymbol{\alpha}_m = \frac{1}{2}(\alpha_m \ \ \alpha_m^*)^T$. We have adopted this method for determining the complex amplitudes rather than a joint resolution of the criterion (3) with respect to $\{\boldsymbol{\alpha}_m\}_{1 \leq m \leq M}$ since the latter is more greedy in terms of computational cost as it demands to pseudo-invert a $N \times D$ matrix, N and M being large. In fact, it is shown in [9] that estimating α_m according to expression (6) is asymptotically consistent (for N great enough). Finally, the real amplitudes $\{a_m\}$ and the initial phases $\{\phi_m\}$ can be extracted from $\{\alpha_{2m-1}\}$, for $m = 1, \dots, M$.

3. FREQUENCY-TRANSFORM

3.1. Duality between transients and sinusoids

Frequency-transforms such as Fourier's, Hartley's, cosine transform (DCT-I-II-III-IV) and sine transform obey to the time-frequency duality principle [10]. A Dirac-like shape in the time domain turns into an oscillating shape in the frequency domain and *vice versa*. To illustrate this phenomenon, we consider the Fourier Transform $X(\lambda)$ of a synthetic strong transient, the DDS (Damped & Delayed Sinusoid) signal [5] defined as

$$x(n) = ae^{i\phi} e^{(n-t)(i\omega+d)} \psi(n-t) \tag{7}$$

(see figure 1-b), according to

$$X(\lambda) = ae^{i\phi} S(\lambda) e^{-i\lambda t}, \ \lambda \in [0, \pi] \tag{8}$$

where

$$S(\lambda) = \frac{1 - e^{(N-t)(i(\omega-\lambda)+d)}}{1 - e^{i(\omega-\lambda)+d}}. \tag{9}$$

We denote by a, ϕ, ω, d, t, respectively, the amplitude, the phase, the angular frequency, the damping factor and the delay parameter and $\psi(n)$ is the Heaviside function. The $X(\lambda)$ expression indicates that a delay t results in an oscillating term "'$e^{-i\lambda t}$'" in the Fourier domain. This classical result is illustrated on figure 1.

This principle in conjunction with the EDS model can be used profitably. In effect, a sharp transient presents a singularity that is similar to the signal in figures 1-(b)-(d). In this case, temporal modeling of the signal by SA-EDS provides poor performance [2], [5]. Yet, as the corresponding frequency-transformed signal is essentially oscillating it is advantageous to proceed to modeling in the frequency domain where SA-EDS is more efficient. The fact is the signal on figure 1-d is much better represented by SA-EDS comparatively with the signal on 1-b. Note that this approach is more consistent whenever the term $-|dN|$ in the expression of $X(\lambda)$ is large enough (in absolute value). In this case, $S(\lambda) \approx 1$. The ideal situation takes place when either the analysis length N is large or the transient signal fades out abruptly *i.e* $|d|$ is large. In other words, the transient temporal length should be smaller than the analysis length, which can be controlled through a proper choice of N. A rough estimation of the delay t is then $\hat{t} = N\hat{\omega}/\pi$, where $\hat{\omega}$ is the estimated value of the angular frequency ω of the transformed signal.

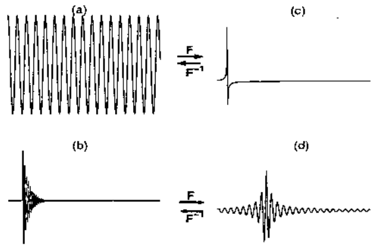

Fig. 1. time domain, (a) sinusoid (DDS with $t = d = 0$) (b) DDS with $t = 50$;(c) positive frequency domain (d) $\Re e\{X(\lambda)\}/2$

3.2. DCT-IV transform expression

Let $\boldsymbol{x} \in \mathbb{R}^{N \times 1}$ the real audio signal observed on a N length frame. The transformed audio signal $\boldsymbol{X} \in \mathbb{R}^{N \times 1}$ is defined as

$$\boldsymbol{X} = \mathcal{F}_N(\boldsymbol{x}) = \boldsymbol{F}\boldsymbol{W}\boldsymbol{x} \tag{10}$$

where $\boldsymbol{W} = \text{diag}\{h(0), \dots, h(N-1)\}$ and $\forall n, \ h(n) \neq 0$ is a temporal weighting window such that $\boldsymbol{W}^{-1} = \text{diag}\{1/h(0), \dots, 1/h(N-1)\}$ *i.e* $\boldsymbol{W}^{-1}\boldsymbol{W} = \boldsymbol{I}_N$. $\mathcal{F}_N(.) : \mathbb{R}^{N \times 1} \rightarrow \mathbb{R}^{N \times 1}$ is the windowed DCT-IV transform. Note that the $N \times N$ matrix

$$\boldsymbol{F} = (\vartheta_{nk})_{n,k} \tag{11}$$

such that [10]

$$\vartheta_{nk} = \sqrt{2/N} \cos\left((n+0.5)(k+0.5)\pi/N\right) \tag{12}$$

is real, symmetric and unitary ($\boldsymbol{F}^{-1} = \boldsymbol{F} = \boldsymbol{F}^T$). Then $\boldsymbol{x} = \mathcal{F}_N^{-1} \circ \mathcal{F}_N(\boldsymbol{x})$. Choosing the DCT-IV rather than other frequency-transforms is justified by the fact that it is well adapted to real signals and does not require any inversion processing.

4. TRANSIENT MODELING METHOD

4.1. The FTSA-EDS algorithm

FTSA-EDS refers to Frequency-Transform Subspace Algorithm with an EDS model and is described in this subsection. First, the frequency transform of the transient signal $\boldsymbol{x}$ is computed after it has been windowed. Then the signal is modeled in the frequency domain according to the algorithm presented in section (2.2). $\hat{\boldsymbol{X}}$ denotes the estimated version of the transformed signal $\boldsymbol{X}$. Inverse transform and windowing is finally applied to deduce the time domain estimated signal $\hat{\boldsymbol{x}}$ as shown on figure 2. The process is summed-up in the table (1).

4.2. Simulation on real transient signals

The considered signal is 16 ms of a castanets sample (typical percussive signal) which is shown on figure 3-a. This sample

(1)	$\mathcal{H}_L(X)$	$:=$	$\mathcal{H}_L \circ \mathcal{F}_N(x)$
(2)	$\mathcal{H}_L(X)$	$\xrightarrow{\boxed{\text{SA-EDS}}}$	$\hat{X}$
(3)	$\hat{x}$	$:=$	$\mathcal{F}_N^{-1}(\hat{X})$

Table 1. FTSA-EDS algorithm

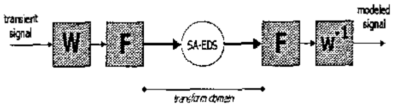

Fig. 2. Bloc diagram of the FTSA-EDS algorithm

is often used to illustrate a strong transient character audio phenomenon [2], [5] and [6]. The sampling frequency is 32 kHz. The modeling order is $M = 40$. The temporal weighting is achieved with a Hamming window $h(n)$.

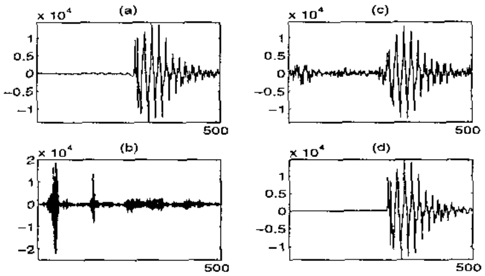

Fig. 3. castanets onset, (a) original signal, (b) transformed signal (c) SA-EDS ($M = 40$), (d) FTSA-EDS ($M = 40$)

Figure 3-b presents the castanets signal in the DCT-IV domain. Figures 3-c and 3-d give modeling results with SA-EDS and FTSA-EDS. Two aspects should be stressed. The first is the total absence of pre-echo with the FTSA-EDS modeling (*c.f.* figures 3-d) while such a distortion is observed with SA-EDS modeling (c.f. figures 3-c). The second is the great dynamic reproduction of the attack with FTSA-EDS comparatively to the limited SA-EDS modeling case.

Figure 4 allows to work out both algorithms behavior. Modeling order is progressively increased, $M = \{2; 10; 20; 35\}$ going from the top to the bottom.

It can be noticed that the SA-EDS models signal all over the analysis length which brings pre-echo even for a low modeling order (*c.f.* figure 4-e). On the contrary, FTSA-EDS only models signal consistent part of the analysis window (*c.f.* figure 4-a,b,c et d) and no pre-echo is thus created.

5. PROPOSED ALGORITHM LIMITS AND "TRANSIENT + SINUSOIDAL" SCHEME

5.1. Computational cost

Signal subspace methods are based on the the Singular Values Decomposition (SVD) of a structured $L \times L$ matrix with $L = N/2$ [7]. Now, in audio compression context, analysis windows of length 128 up to 2048 samples should be used to get sparse representations ($M \ll N$) of the modeled signal. This implies a high computational cost $O(L^3)$ for a straightforward SVD processing. Iterative fast algorithms can be used to compute the signal basis based on the Orthogonal Iterations algorithm and the Lanczos algorithm as well as exploiting the data matrix structure through the use of the FFT. Computational cost is then evaluated in terms of $O(LD^2)$ or $O(DL\log L)$. A panorama of these methods is presented in [7] and [12]. Note that, improvements are achieved in choosing a frequency transform that prevents any inversion and an EDS complex amplitudes processing proper strategy. The utilization of fast algorithms for our simulations has yielded a processing time to real time ratio of order 6 on a Pentium III under Matlab 5.3.

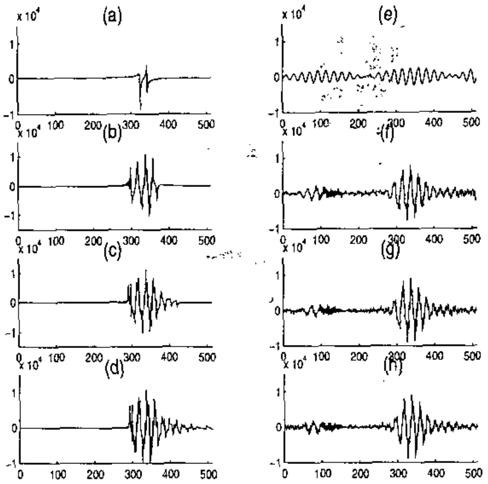

Fig. 4. castanets onset, progressive modeling ($M = 2; 10; 20; 35$), right part (a),(b),(c),(d) : FTSA-EDS, left part (e),(f),(g),(h) : SA-EDS

5.2. Quasi-stationary segments modeling

On figures 5 and 6, modeling of quasi-stationary speech and bells signals is shown. Note that, while FTSA-EDS is as efficient as SA-EDS on the speech signal (see figure 5), it presents limited performance on the bells signal (see figure 6). This is due to the fact that the signal is very oscillating, so it has a transient representation in the frequency domain which is not adapted to the EDS model.

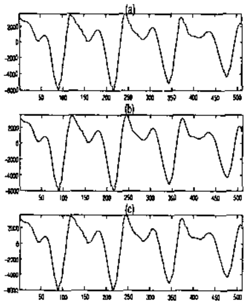

Fig. 5. speech segment, (a) original signal, (b) FTSA-EDS ($M = 35$), (c) SA-EDS ($M = 35$)

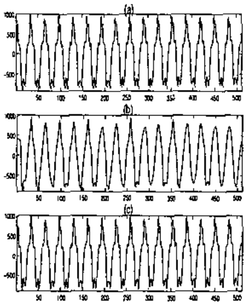

Fig. 6. bells segment, (a) original signal, (b) FTSA-EDS ($M = 35$), (c) SA-EDS ($M = 35$)

5.3. "Transient + sinusoidal" modeling scheme

In [6], [10] and [11], a "Sinusoidal+Transient+Noise" (STN) representation is developed. The FTSA-EDS which is appropriate for transient signals can be associated to sinusoidal modeling so as to provide a high performance and varied audio signal "Transient+Sinusoidal" modeling scheme. The sinusoidal analysis can be achieved by means of the SA-EDS algorithm presented in section (2.2) applied to the sinusoidal model[2], or spectral peak-picking techniques [1], or even Matching Pursuit with a Fourier waveforms dictionary [10]. The "Transient+Sinusoidal" modeling scheme is presented on figure 7. First, the audio signal $\boldsymbol{x}$ is analyzed with the FTSA-EDS model of order M_1 so as to represent its transient part. Second, the residual signal $\boldsymbol{r} = \boldsymbol{x} - \hat{\boldsymbol{x}}$ is computed which essentially consists of oscillating and noise components. Finally, SA-EDS modeling is run with order M_2 and $\forall m,\ d_m = 0$ on the signal $\boldsymbol{r}$ providing $\hat{\boldsymbol{r}}$. The transient and sinusoidal parts are then synthesized into $\hat{\boldsymbol{x}} + \hat{\boldsymbol{r}}$ with a model order $M_1 + M_2$. Note that, in contrast to previous work on STN systems, the transient part is modeled prior to the sinusoidal part. An example of a 1024-sample drums signal (32 ms) modeled with this approach is given on figure 8.

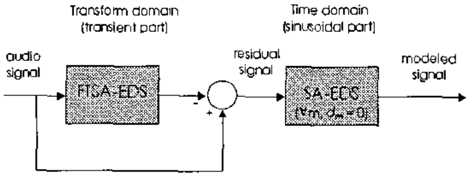

Fig. 7. "Transient+Sinusoidal" modeling bloc diagram

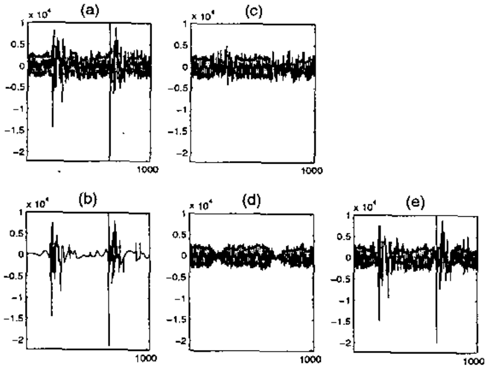

Fig. 8. (a) Drums signal: $\boldsymbol{x}$, (b) transient part : $\hat{\boldsymbol{x}}$ ($M_1 = 25$), (c) residual signal : $\boldsymbol{r}$, (d) sinusoidal part : $\hat{\boldsymbol{r}}$ ($M_2 = 15$), (e) modeled signal : $\hat{\boldsymbol{x}} + \hat{\boldsymbol{r}}$ ($M_1 + M_2 = 40$)

6. CONCLUSION

In this communication, we have presented a strong transient audio signals modeling scheme named FTSA-EDS (Frequency-Transform Subspace Algorithm with Exponentially Damped Sinusoids model). The algorithm is based on the parametric non-stationary model EDS and a signal subspace algorithm for determining the model parameters. The analysis is made in the DCT-IV frequency domain. Simulation results presented on a typical transient signal, a castanets signal, allow to work out the high performance of this approach. Pre-echo is completely canceled and an excellent reproduction of the onset dynamic is achieved. Finally, a "Transient+Sinusoidal" modeling scheme is presented that allows to represent the wider variety of audio signals which has been confirmed through informal listening tests.

[2] $\forall m,\ d_m = 0$ in expression (1).

Acknowledgements

Special acknowledgements go to Nicolas Moreau, professor at the Department of Signal and Image Processing of Telecom Paris, for supervising this work.

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] J. Nieuwenhuijse, R. Heusdens, E.F. Deprettere, "Robust Exponential Modeling of Audio Signal", *Proc. of IEEE ICASSP*. Vol. 6 , 1998

[3] J. Jensen, S.H. Jensen, E. Hansen, "Exponential sinusoidal modeling of transitional speech segments.", *Proc. of IEEE ICASSP*, Vol 1, 1999

[4] P. Lemmerling, I. Dologlou, S. Van Huffel, "Speech Compression based on exact modeling and Structured Total Least Norm optimization", *Proc. of IEEE ICASSP*, Seattle, US, May 1998

[5] R. Boyer, K. Abed-Meraim, "Audio transients modeling by Damped & Delayed Sinusoids (DDS)", *Proc. of IEEE ICASSP*, May 2002, *Accepted*

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

[7] A-J. Van Der Veen, ED. F. Deprettere, A. Lee Swindlehurst, "Subspace-Based Signal Analysis Using Singular Value Decomposition", *Proc of the IEEE*, Vol. 81, No 9, September 1993

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

[9] P. Stoica, H. Li, J. Li, "Amplitude Estimation of Sinusoidal Signals : Survey, New Results, and an Application", *IEEE Trans. on SP*, Vol. 48, No. 2, February 2000.

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

[11] S. Levine, *Audio Representations for Data Compression and Compressed Domain Processing*, PhD thesis, Stanford University, 1998.

[12] P. Comon, G.H. Golub, "Tracking a Few Extreme Singular Values and Vector in Signal Processing", *Proc. of the IEEE*, Vol. 78, No. 8, August 1990