AN EDS MODELLING TOOL FOR TRACKING AND MODIFYING MUSICAL SIGNALS

Bertrand DAVID, Gaël RICHARD and Roland BADEAU

ENST, Department of Signal and Image processing
École Nationale Supérieure des Télécommunications,
46, rue Barrault,
75634 PARIS cedex 13, FRANCE
bedavid,grichard,rbadeau@tsi.enst.fr

ABSTRACT

An analysis/synthesis scheme for musical signals is introduced in this paper. It is based on an adaptive subspace analysis and the Exponentially Damped Sinusoids model. This method leads to a new representation, called the HR-ogram, where the signal components are represented as points in the time-frequency plane. These points are gathered according to their frequency, phase and amplitude proximity from an analysis time-instant to the following one. This leads to an accurate deterministic/stochastic decomposition using a projection onto the noise subspace. The whole technique allows a separate processing for both components.

1. INTRODUCTION

Most of the analysis/synthesis schemes designed for musical sounds found in the literature are based on either a frequency-domain or a time-domain approach. Both families have evolved into a broad variety of algorithms from their very ancestors: the so-called phase-vocoder [1] and the OLA (overlap-add) method. These tools and their derivatives (Quatieri and Serra techniques [2, 3], synchronized OLA methods SOLA [4], PSOLA [5]) are widely used in the context of audio signal processing [6, 7].

The technique described in this paper is mostly related to the first class of methods: it relies on an Exponentially Damped Sinusoids (EDS) model and takes advantage of the signal decomposition into a deterministic part and a noise component. The EDS modelling allows an accurate representation of each signal frame in terms of the amplitudes, phases, damping factors and frequencies of the component sine waves while the modifications can be processed separately on both parts (deterministic and stochastic) of the signal decomposition.

This work follows earlier ones designed for estimating, tracking and modifying musical sounds [8, 9] and rely on the high resolution properties of the subspace analysis [10]. The main drawback of this approach is the computational cost of such algorithms but the newest versions have become adaptive and overcome the constraint of computing a Singular Value Decomposition at each time step [11].

The theoretical background presented in section 2 mainly recovers from [9], highlighting the relationship with the well-known quasi-stationary models [2] and [3] and the applicative context. Section 3 shows analysis results and the modification and synthesis techniques are discussed in section 4.

2. THEORETICAL BACKGROUND

2.1. Definitions and model

The discrete signal to be analysed and modified is assumed real valued and denoted $s(t)$. It is segmented in overlapping frames $x(t,u) = s(t + t_a(u))w_a(t)$ where $t_a(u)$ are the analysis marks indexed by the non-negative integer u and w_a is the analysis window assumed of finite length L_a. The time-instants $t_a(u)$ are usually regularly spaced, *i.e.* $t_a(u) = u\Delta_a,\ u \in \mathbb{N}$, where the interval Δ_a is a fixed increment such as $\Delta_a \leq L_a$. In this paper, w_a will always be the rectangular window of length L_a.

For each frame, an Exponentially Damped Sinusoids model is used:

$$x(t,u) = \sum_{k=1}^{M}(b_k z_k^t + b_k^* z_k^{*t}) \qquad (1)$$

where $b_k = \dfrac{A_k}{2}\exp(j\phi_k)$ is the complex amplitude of the k^{th} component (A_k is the amplitude of the corresponding real component and ϕ_k its initial phase), and $z_k = \exp(-\alpha_k + j2\pi f_k)$ is its complex pole. α_k and f_k denote its damping factor and frequency, ranging in $\mathbb{R}$ for the first one and in $[0\ 1/2]$ for the other. These parameters are implicit functions of $t_a(u)$, not reported in (1) for lightening purposes. Like the well-known sinusoidal decompositions found in [2, 3], the parameters are assumed to vary slowly, *i.e.* they are considered constant over the window duration. Anyway, it should be noted that, in contrast, the definition of the EDS model includes a possible variation of the components envelopes leading to a more accurate representation for each frame, yielding to a residual of lower energy than the one obtained by an undamped modelling for the same order M and window length L_a.

2.2. Analysis stage

The estimation of the $4M$ parameters for each frame $x(t)$ (for simplicity the u-dependency is removed) is performed using a subspace-based technique for the frequencies f_k and damping factors α_k. The b_k are jointly estimated by a least squares method. The subspace analysis takes into account the particular mathematical structure of the model, leading to a high resolution (HR) estimation: when the signal is noise-free the limit for separating two close components is only restrained by the limited computational capacity.

2.2.1. HR method outlines

Using the $L_a = 2N - 1$ samples of the frame, the square Hankel structured data matrix H is defined as

$$\mathbf{H} = \begin{pmatrix} x(0) & x(1) & \dots & x(N-1) \\ x(1) & x(2) & \dots & x(N) \\ \vdots & \vdots & \ddots & \vdots \\ x(N-1) & x(N) & \dots & x(L_a - 1) \end{pmatrix}. \qquad (2)$$

Assuming $2M \leq N$, the symmetric real matrix $\mathbf{H}$ is rank-deficient of dimension $2M$ [9]. Its eigendecomposition yields

$$\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \qquad (3)$$

where $\mathbf{U}$ is an $N \times 2M$ orthonormal real matrix. In presence of an additive white noise, $\mathbf{H}$ becomes full rank and the columns of $\mathbf{U}$ are defined as the $2M$ dominant eigenvectors, corresponding to the $2M$ eigenvalues of highest magnitude.

The signal poles $\{z_k, z_k^*\}_{k=1,\dots,M}$ are estimated by taking into account the rotational invariance property of the signal subspace, which is expressed in terms of a real $2M \times 2M$ matrix $\mathbf{\Phi}$ whose eigenvalues are the signal poles:

$$\mathbf{U}_\uparrow = \mathbf{U}_\downarrow \mathbf{\Phi} \qquad (4)$$

where $\mathbf{U}_\uparrow$ (*resp.* $\mathbf{U}_\downarrow$) is obtained by deleting the first (*resp.* the last) row of $\mathbf{U}$.

2.2.2. Estimation of the complex amplitudes

This estimation of the complex amplitudes $\{b_k, b_k^*\}_{k=1,\dots,M}$ is performed for each frame using a Least Squares (LS) method. The $N \times 2M$ Vandermonde matrix $\mathbf{Z}$ is defined by

$$\mathbf{Z} = \begin{pmatrix} 1 & \dots & 1 & 1 & \dots & 1 \\ z_1 & \dots & z_M & z_1^* & \dots & z_M^* \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ z_1^{N-1} & \dots & z_M^{N-1} & z_1^{*N-1} & \dots & z_M^{*N-1} \end{pmatrix}. \qquad (5)$$

When the frame is noise-free, the column vector $\mathbf{b} = [b_1, \dots, b_M, b_1^*, \dots, b_M^*]^T$ satisfies the relation

$$\mathbf{Z}\mathbf{b} = \mathbf{x} \qquad (6)$$

where $\mathbf{x} = [x(0), \dots, x(N-1)]^T$. The solution of (6) in the least squares sense is $\mathbf{b} = \mathbf{Z}^+\mathbf{b}$ where $\mathbf{Z}^+$ denotes the pseudo-inverse of the matrix $\mathbf{Z}$.

2.2.3. Subspace tracking

Since the HR-method relies on the eigenvalue decomposition (EVD) of the data matrix $\mathbf{H}$, without further modification the estimation of the frequencies and damping factors would require an EVD at each time step and thus would lead to a high computational cost (*cf.* [9] for more details). However this cost can be reduced by using an adaptive algorithm which avoids the computation of the EVD. It is based on an iterative algorithm called *Orthogonal Iteration* [12] and uses a two steps procedure which yields the matrix $\mathbf{U}$ when the convergence is reached. For our tracking purpose, this algorithm is applied in a sequential way, assuming that it converges much faster than the variations of the signal subspace. As shown in table 1, it involves two auxiliary matrices $\mathbf{A}(u)$ and $\mathbf{R}(u)$ and an economy size QR factorization since $\mathbf{A}(u)$ is of dimension $N \times 2M$.

Table 1: Sequential iteration for subspace tracking

```
Initialization of the signal subspace matrix
    $\mathbf{U}(0) = \begin{pmatrix} \mathbf{I}_{2M} \\ \mathbf{0}_{(N-2M)\times 2M} \end{pmatrix}$
For each analysis time-instant $t_a(u), u = 1, 2, \dots$ iterate:
    matrix product $\mathbf{A}(u) = \mathbf{H}(u)\mathbf{U}(u-1)$
    skinny QR factorization $\mathbf{A}(u) = \mathbf{U}(u)\mathbf{R}(u)$
```

3. ANALYSIS RESULTS: HR-OGRAM

In this section, the capabilities of the algorithm are demonstrated for both synthetic and real data (singing voice). A time-frequency representation is introduced, called the HR-ogram[1] which serves the same goals as the spectrogram in Fourier analysis. The HR-ogram represents the components as points $(t_a(u), f_k(u))$ in the time-frequency plane for each analysis time-instant $t_a(u)$. The energy ϵ_k of the k^{th} component is represented in decibels using gray levels and defined as

$$\epsilon_k = A_k^2 \frac{1 - \exp(-2\alpha_k L_a)}{1 - \exp(-2\alpha_k)}. \qquad (7)$$

This definition includes the effect of the damping factor in the graph. This avoids the overvaluing of spurious poles, often related to noise, which can be highly damped and estimated at very high magnitude values, resulting in a weak component.

3.1. Simulation example

The graphs of figure 1 show the analysis results for a three components signal with an additive white noise corresponding to a 27dB signal to noise ratio (SNR). All the components are undamped.

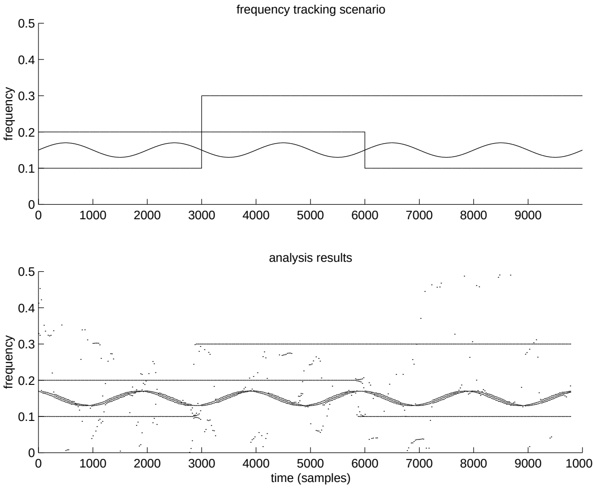

Figure 1: Analysis results for a synthetic signal

[1] standing for High Resolution Spectrogram

Two of them present a frequency jump at distinct time-instants while the third one is sinusoidally modulated at the period of 2000 samples and with a 0.05 frequency deviation.

The analysis parameters are set as follows: the window length is $L_a = 201$, the number of components is overestimated to $M = 6$ and the analysis is performed every $\Delta_a = 25$ samples.

The results show clearly the good tracking of the three components. The convergence of the algorithm is fast enough to handle the frequency jumps and when they occur for one component the estimation of the other ones remains stable and accurate. The modulated component is often represented by a few (2 or 3) poles. The signal model (1) is indeed not fully respected by the frequency modulated component. The spurious poles corresponding to the additive noise are easily identifiable for they do not aggregate in a specific shape and are widely spread in the whole frequency range.

3.2. Singing voice subspace tracking

This example is a female soprano singing voice, who realizes an up and down glissando between C5 and E5 . The signal is recorded at the 44100 Hz sampling rate. The analysis provides the HR-ogram shown in figure 2, which highlights the spectral lines associated to the harmonic structure but also the poles related to noise, gathered in formant-like shapes.

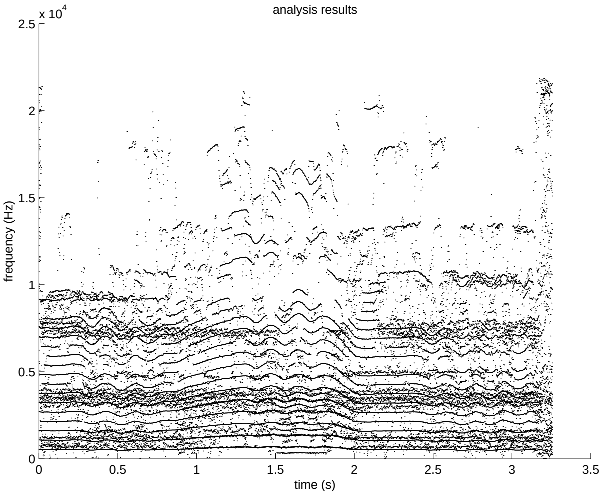

Figure 2: HR-ogram of a soprano singer

4. MODIFICATION AND SYNTHESIS

Each point of the HR-ogram is well localized both in frequency and time domains and can be modified individually. However, in order to achieve a high quality analysis/modification/synthesis the poles related to the sinusoidal components and those related to noise must be processed separately. Moreover, the EDS model does not represent accurately the stochastic part of the signal.

4.1. Deterministic/stochastic decomposition

4.1.1. Method

As in most of the analysis schemes, a noise component $w(t,u)$ is added to the model (1) leading to:

$$x(t,u) = \sum_{k=1}^{M} (b_k z_k^t + b_k^* z_k^{*t}) + w(t,u). \qquad (8)$$

This component is often expressed as a time-varying filtering of a white stochastic process [6] and will be referred to as the stochastic component of the signal while the noise-free EDS model will be referred to as the deterministic component.

A common technique to derive the noise component consists in subtracting to $x(t,u)$ the deterministic part, after its estimation. But this can lead to a significative amount of sinusoidal components introduced in $w(t,u)$. In order to avoid this effect, $w(t,u)$ is obtained by projecting the signal onto the noise subspace. For the noisy model (8), the matrix $\mathbf{H}$ is full rank. The signal subspace is the space spanned by the eigenvectors associated to the $N-2M$ smallest eigenvalues.

The number M of sinusoids is chosen accordingly to the stability of the spectral lines of their associated poles. M is first overestimated and the signal poles $z_k(u)_{k=1,\ldots,M}$ and corresponding complex amplitudes $b_k(u)$ are estimated at the analysis time-instant $t_a(u)$. $z_m(u+1)$ and $b_m(u+1)$ are estimated at the time-instant $t_a(u+1)$ and distances are computed to measure how these poles are close in terms of frequency, amplitude and phase. The corresponding distances d_f, d_A and d_ϕ are thus defined as:

$$d_f(k,m) = 4(f_k(u) - f_m(u+1))^2 \qquad (9)$$

$$d_A(k,m) = (A_k(u) - A_m(u+1))^2 \qquad (10)$$

$$d_\phi(k,m) = \frac{(\lfloor \phi_k(u) + 2\pi f_k(u)\Delta_a \rfloor - \lfloor \phi_m(u+1) \rfloor)^2}{4\pi^2} \qquad (11)$$

where $\lfloor \phi \rfloor$ denotes the principal determination of ϕ and the signal is normalized to a maximum magnitude equal to one. The spectral lines are formed according to the following steps:

1. for each k, $m_k = \arg_m \min d(k,m)$ is computed where $d(k,m) = d_A(k,m) + d_f(k,m) + d_\phi(k,m)$,
2. for each k, the poles $z_k(u)$ and $z_{m_k}(u+1)$ are connected if $|f_k(u) - f_{m_k}(u+1)|/f_k(u) < 1\%$ and $d_A < -20$dB and $d_\phi < -30$dB.

The number of components of the deterministic part $\hat{M} < M$ is defined as the number of poles which have been connected between $t_a(u)$ and $t_a(u+1)$. The stochastic component is obtained by the projection:

$$\mathbf{w} = (\mathbf{I}_{2\hat{M}} - \mathbf{U}_{2\hat{M}}^H \mathbf{U}_{2\hat{M}})\mathbf{x} \qquad (12)$$

where $\mathbf{U}_{2\hat{M}}$ is the subspace matrix whose columns are the $2\hat{M}$ dominant eigenvectors, $\mathbf{w} = [w(0,u),\ldots,w(L_a-1,u)]^T$ and $\mathbf{x} = [x(0,u),\ldots,x(L_a-1,u)]^T$.

4.1.2. Results

Figure 3 shows the spectral lines tracked by the technique described in section 4.1.1, leading to a time-frequency representation of the deterministic component of the preceding singing voice signal for the time indices ranging between 1s and 1.25s (*cf.* the

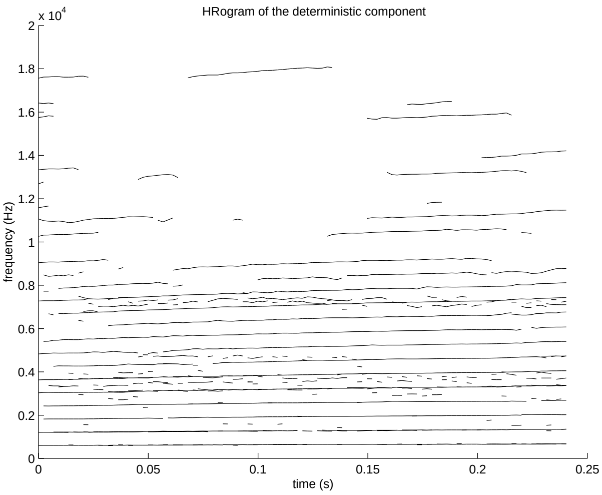

Figure 3: HR-ogram of the deterministic part

HR-ogram of figure 2). Besides a few lines related to noise components and easily identifiable because of their short time duration, the harmonic structure is highlighted and shows clearly the frequency modulation due to the *glissando* produced by the soprano singer. The short-term spectra (10ms) of the signal and its stochastic part are represented in figure 4. The sinusoidal part has been mostly removed by the projection while the formant-like reinforcements around 4000 Hz and 7500 Hz are emphasized.

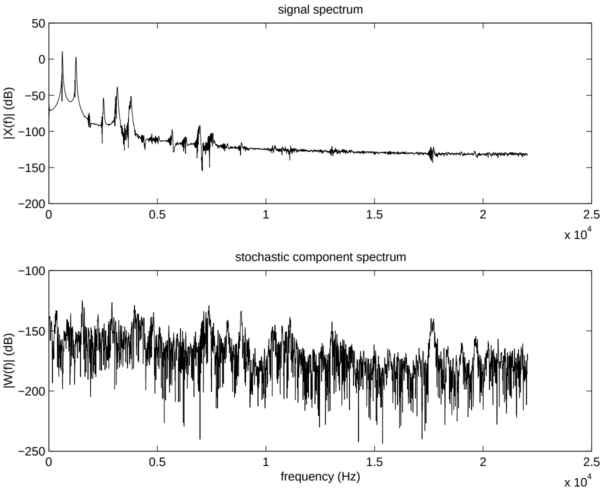

Figure 4: Short-term spectra of $x(t,u)$ and $w(t,u)$

4.2. Modifications

Coming along with the deterministic/stochastic decomposition, many audio effects can be processed. For example, pitch-shifting can be applied only on the deterministic part, eventually taking into account the spectral envelope. An interesting by-product of this decomposition is the capability of processing each spectral line to add or remove vibrato and tremolo, to adjust the pitch or the duration or to modify the ratio between the voiced and the unvoiced part of the sound.

5. CONCLUSIONS

In this paper an analysis/synthesis scheme has been proposed. It uses a High Resolution adaptive method which overcomes the Fourier resolution limit and achieves an accurate estimation of the sinusoidal components of the signal . The stochastic part is then derived by projecting the signal onto the noise subspace. The deterministic and stochastic parts can thus be processed separately, leading to high quality audio effects.

Acknowledgements. The authors are very grateful to Nathalie Henrich for having provided the musical signals.

6. REFERENCES

[1] M. R. Portnoff, “Implementation of the digital phase vocoder using the fast fourier transform,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, no. 3, pp. 243–248, Jun 1976.

[2] R. J. McAulay and T. F. Quatieri, “Speech analysis/synthesis based on a sinusoidal representation,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, no. 4, pp. 744–754, Aug 1986.

[3] X. Serra and J. Smith, “Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition,” *Computer Music J.*, vol. 14, no. 4, pp. 12–24, Winter 1990.

[4] S. Roucos and A. M. Wilgus, “High quality time-scale modification of speech,” *Proc. IEEE ICASSP-85, Tampa*, pp. 493–496, Apr 1985.

[5] E. Moulines and F. Charpentier, “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,” *Speech Communication*, vol. 9, no. 5/6, pp. 453–467, Dec 1990.

[6] M. Kahrs and K. Brandenbourg, *Applications of Digital Signal Processing to Audio and Acoustics*. Dortrecht, Netherland: Kluwer Academic Press, 1998.

[7] E. Moulines and J. Laroche, “Non parametric techniques for pitch-scale and time-scale modification of speech.” *Speech Communication*, vol. 16, pp. 175–205, Feb 1995.

[8] B. David, R. Badeau, and G. Richard, “Sintrack analysis for tracking components of musical signals,” in *Forum Acusticum*, Sevilla, Spain, Sep 2002.

[9] R. Badeau, R. Boyer, and B. David, “EDS parametric modeling and tracking of audio signals,” in *5th Int. Conf. on Digital Audio Effects (DAFx-02)*, Hamburg, Germany, Sep 2002, pp. 139–144.

[10] Y. Hua and T. K. Sarkar, “Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-38, no. 5, pp. 814–824, May 1990.

[11] R. Badeau, G. Richard, and B. David, “Sliding window adaptive SVD algorithms,” *IEEE Trans. Signal Processing*, to be published.

[12] G. Golub and C. V. Loan, *Matrix computations*, 3rd ed. Baltimore and London: Johns Hopkins University Press, 1996.