TIME-VARIANT HARMONIC AND TRANSIENT SIGNAL MODELING BY JOINT POLYNOMIAL AND PIECEWISE LINEAR APPROXIMATION

Miroslav Zivanovic[1] *and Johan Schoukens*[2]

[1] Dpt. IEE, Universidad Publica de Navarra, Campus Arrosadia, 31006, Pamplona, Spain
phone: + 34 948 16 90 24, fax: + 34 948 16 97 20, email: miro@unavarra.es

[2] Dpt. ELEC, Vrije Universiteit Brussel, Pleinlaan 2, B-1050, Brussels, Belgium
phone: +32 2 629 29 47, fax: +32 2 629 28 50, email: jschouk@vub.ac.be

ABSTRACT

We present a compact approach to simultaneous modeling of non-stochastic time-variant components in audio signals. We show that the harmonic energy can be properly described by a single polynomial, while short events are well captured by a continuous piecewise linear function. The proposed method is robust to fundamental frequency estimation errors and inharmonicities in the audio signal. The comparative results suggest that our method achieves the performance of the state-of-the-art by using much less parameters and thus yielding higher computational efficiency.

1. INTRODUCTION

Time-variant harmonic modelling is proven to be a useful analysis tool for characterizing audio signals like music and speech. It has been widely used in a number of application areas like audio synthesis, transformation, coding and enhancement.

The general audio signal model is often conceived as a finite sum of harmonically related sinusoids, whose instantaneous amplitudes $A_i(n)$ and frequencies $iF_0(n)$ vary slowly in the analysis window, plus additive noise $r(n)$:

$$s(n) = \sum_i A_i(n)\cos\left[2\pi\, iF_0(n)n + \varphi_i\right] + r(n)\,. \qquad (1)$$

A number of different approaches aim at describing the time variations of the harmonic parameters. Methods like [1, 2] estimate the harmonic parameters by fitting (1) to the original signal through maximum likelihood or least-squares method (LS). Another approach is based on the analysis of the signal's STFT, where the harmonic parameters are typically estimated in the spectral peak detection-description-classification framework [3, 4]. The statistical approach [5] implements a Bayesian network with a particular prior structure, built from the conditional probabilities which establish the relationship among the harmonic parameters.

These methods, however, often fail to correctly characterize short-time events like fast transients, which are common in audio signals e.g. sharp attacks at note onset/offset, transition between voiced/unvoiced regions in speech signals etc [6]. There are two general strategies to capture short-time events in audio signals. One is based on extending (1) by explicit modelling of transients [7-9]. This yields a simultaneous estimation of the harmonic and transient components, providing thus a compact audio signal description. A large number of parameters and high computational cost are however principal shortcomings of these methods. Regarding the second strategy, each signal component is estimated at a time by analyzing the signal's expansion in local trigonometric basis or some other conveniently chosen time-frequency or time-scale representation [10, 11]. Serious drawbacks in these methods consist in mutual bias produced by separate signal component estimation and difficulty to impose a clear harmonic/transient separating threshold.

The method we propose herein addresses the issue of joint harmonic and transient signal component modeling. The key idea is the expansion of the input signal onto a set of harmonically related sinusoids determined by the signal's mean instantaneous fundamental frequency. If the signal changes in a slow and continuous manner, the expansion coefficients define small-order time polynomials. In order to allow for fast energy changes during transient events, the expansion coefficients are the parameters of small-order time piecewise linear functions (PWL). The cue to using the same approximation basis is the fact that the polynomial and PWL approximations describe very different variation trends. Accordingly, a simple linear least-squares (LS) algorithm will properly mix the contributions from both approximations as a function of the "transientness" and "harmonicity" of the analyzed signal segment.

The present paper is organized as follows: in Section 2 we propose a signal model used to describe harmonic and transient components in audio signals. Section 3 explains the model behavior. In Section 4 we pose the estimation problem as linear LS. In Section 5 we present a comparative study among different methods together with an illustrative example. The conclusions appear in Section 6.

2. THE SIGNAL MODEL

Similar to [7-9] we are going to treat a realistic audio signal model represented as a sum of harmonic, transient and re-

This work is sponsored by the Fund for Scientific Research (FWO-Vlaanderen), the Flemish Government (Methusalem 1), and the Belgian Federal Government (IUAP VI/4).

 ISSN 2076-1465

sidual term. The residual is typically modelled as a Gaussian noise whose power spectral density is modified by a time-variant filter bank. The models for the harmonic and transient component respectively are described in the following subsections.

2.1 The harmonic model

Below is a brief summary of the harmonic model described in [12]. It has been developed based on the assumption that the harmonic parameters vary slowly and continuously in the analysis window. Starting from expression (1), a linear variation of the parameters accounts for the continuity constraint:

$$A_i(n)\sin\varphi_i = a_0^{(i)} + a_1^{(i)} n, \tag{2}$$

$$A_i(n)\cos\varphi_i = b_0^{(i)} + b_1^{(i)} n, \tag{3}$$

$$f_0(n) = f_0 + f_1 n. \tag{4}$$

The assumption about the slow variation is formulated through the following approximations:

$$\sin 2\pi f_1 n^2 \approx 2\pi f_1 n^2, \quad \cos 2\pi f_1 n^2 \approx 1. \tag{5}$$

By rewriting the harmonic term in (1) as a sum of angles and combining it with (2 - 5), we obtain the model $h(n)$ for the harmonic component in the audio signal:

$$h(n) = \sum_{i=1}^{I} p_s^{(i)}(n)\sin(2\pi i f_0 n) + p_c^{(i)}(n)\cos(2\pi i f_0 n), \tag{6}$$

$$p_s^{(i)}(n) = a_0^{(i)} + a_1^{(i)} n - 2\pi i b_0^{(i)} f_1 n^2 - 2\pi i b_1^{(i)} f_1 n^3, \tag{7.a}$$

$$p_c^{(i)}(n) = b_0^{(i)} + b_1^{(i)} n + 2\pi i a_0^{(i)} f_1 n^2 + 2\pi i a_1^{(i)} f_1 n^3. \tag{7.b}$$

The last expressions show that both amplitude and frequency time-variations are compactly captured by the signal expansion onto the f_0-harmonic basis with polynomial coefficients. In addition, (7.a) and (7.b) provide us with the variation trends in the signal, as well as a possibility to estimate the harmonic model parameters.

If a short-time event takes place in the analysis window, then the initial assumptions no longer hold. As a consequence, the model fails to follow the time-variations in spite of incrementing the approximation order. In absence of frequency modulation during the short-time event, the model may still be able to operate correctly. In opposite case, the coefficients in (7.a) and (7.b) get mutually coupled, not allowing for correct joint modelling of amplitude and frequency variation trends.

2.2 The transient model

It is not an easy task to assign a general model to a transient component in an audio signal. Although it can be argued that most transients are strongly related to the harmonic components (note onset/offset), their overall behaviour can be very complex. Continuous PWL approximation, however, turns out to be a very good candidate for this task, and the reason is twofold. On one hand, it allows fast and abrupt changes in the signal's amplitude. On the other, the modelling can be accomplished by using only a few linear segments in the analysis window.

A continuous PWL function can be efficiently represented by a linear combination of triangular ("hat") functions, also known as the second-order or linear B-splines. A set of hat functions shifted in time forms a basis for a PWL approximation of the audio signal. Such basis assures that the PWL function is continuous and smooth at the joint points i.e. the adjacent linear segments share the same breakpoint. As an illustration, a 5-element uniformly distributed basis is shown on Figure 1. For an arbitrary number of elements M of the basis and length of the analysis window N, we can define the k^{th} basis element $\psi_k(n)$ for the interior breakpoints $k \in [2, M\text{-}1]$:

$$\psi_k(n) = \begin{cases} 0, & n \notin [k-1, k+1]\dfrac{N}{M} \\ \dfrac{M}{N} n - k + 1, & n \in [k-1, k]\dfrac{N}{M} \\ -\dfrac{M}{N} n + k + 1, & n \in [k, k+1]\dfrac{N}{M}. \end{cases} \tag{8}$$

For the boundary breakpoints (k = 1, M) the expressions are slightly different but very similar to (8). According to the aforementioned discussion, the model $t(n)$ for the transient component will contain the basic harmonic structure modulated by an expansion of the audio signal onto a set of uniformly distributed basis functions:

$$t(n) = \sum_{i=1}^{I}\sum_{k=1}^{M} \alpha_k^{(i)} \psi_k(n)\sin(2\pi i f_0 n) + \beta_k^{(i)} \psi_k(n)\cos(2\pi i f_0 n), \tag{9}$$

where α_k and β_k are scalars.

Let us say a couple of words about the choice of decomposition basis in the context of present work. Higher-order basis functions are often used for piecewise polynomial approximation of smooth functions because of their properties of orthogonality and time-shift invariance. However, they are not adequate for describing transients because they often produce large overshoots between successive breakpoints. On the contrary, a PWL approximation follows the variation trend of the data without additional inflections, thus preserving the signal's shape. Another topic of interest is the choice of equidistant (uniform) distribution of the basis functions in the analysis window. Although irregular breakpoint spacing could yield more optimal signal approximation, the uniform distribution proves to be more practical because it is difficult to find globally optimal breakpoint spacing without any a priori knowledge about the transient energy evolution within the analysis window.

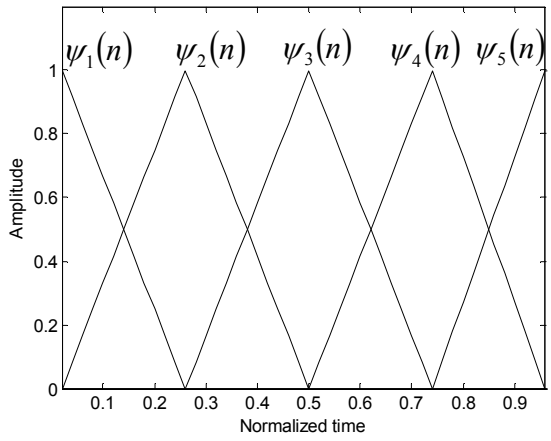

Figure 1 – Five-element basis for a PWL approximation.

3. SIGNAL MODEL BEHAVIOUR

A superposition of the harmonic and transient model is meant to describe the time-variations of the non-stochastic energy in an audio signal. In order to get a deeper insight, let us examine the joint action of the models within the time interval corresponding to a single linear segment of the PWL approximation for a single signal harmonic. According to (8) and (9) the transient model for the i^{th} harmonic within a k^{th} linear segment will be determined by the contributions from the basis functions $\psi_{k-1}(n)$ and $\psi_k(n)$:

$$t_k^{(i)}(n) = \left(c_0^{(i)} + c_1^{(i)} n\right)\sin\left(2\pi\, i f_0 n\right) + \\ + \left(d_0^{(i)} + d_1^{(i)} n\right)\cos\left(2\pi\, i f_0 n\right), \tag{10}$$

$$c_0^{(i)} = \alpha_k^{(i)}(k+1) + \alpha_{k-1}^{(i)} k\,,\ c_1^{(i)} = \frac{M}{N}\left(\alpha_k^{(i)} - \alpha_{k-1}^{(i)}\right), \tag{11.a}$$

$$d_0^{(i)} = \beta_k^{(i)}(k+1) + \beta_{k-1}^{(i)} k\,,\ d_1^{(i)} = \frac{M}{N}\left(\beta_k^{(i)} - \beta_{k-1}^{(i)}\right). \tag{11.b}$$

Combining (11) and (7) we arrive at the following expression for the signal's i^{th} harmonic within a k^{th} linear segment:

$$s_k^{(i)}(n) = P_s^{(i)}(n)\sin\left(2\pi\, i f_0 n\right) + P_c^{(i)}(n)\cos(2\pi\, i f_0 n) \tag{12}$$

$$P_s^{(i)}(n) = a_0^{(i)} + c_0^{(i)} + \left(a_1^{(i)} + c_1^{(i)}\right)n - \\ - 2\pi\, i b_0^{(i)} f_1 n^2 - 2\pi\, i b_1^{(i)} f_1 n^3\,, \tag{13.a}$$

$$P_c^{(i)}(n) = b_0^{(i)} + d_0^{(i)} + \left(b_1^{(i)} + d_1^{(i)}\right)n + \\ + 2\pi\, i a_0^{(i)} f_1 n^2 + 2\pi\, i a_1^{(i)} f_1 n^3\,. \tag{13.b}$$

Accordingly, the modified terms will follow coarse changes in the signal while the remaining terms will describe fine time-variations. Note that we are now capable of capturing correctly amplitude and frequency variations thanks to additional degrees of freedom in (13.a) and (13.b) provided by the parameters c and d. Due to the fact that the polynomial and PWL approximation describe different time-variation trends in the signal, a simple linear fit will do to adjust the contributions from the models. However, due to a certain linear dependency between the models, the system of equations is rank-deficient and has no unique solution. Accordingly, we use pseudo-inverse LS to choose the effective rank of the decomposition matrix and thus obtain a solution that has the minimum possible residual norm.

4. PARAMETER ESTIMATION

In order to apply linear LS, we first have to estimate f_0 which is roughly the mean instantaneous fundamental frequency in the analysis window. Correct instantaneous fundamental frequency estimation is crucial in many audio applications in order to avoid subharmonic errors. There are a huge number of strategies which allow us to detect and estimate the time-varying fundamental frequency in audio signals e.g. direct evaluation of signal's periodicity, frequency domain harmonic matching, spectral period evaluation, psychoacoustic methods etc.

It turns out, however, that the proposed method has a nice property of efficiently mitigating f_0 estimation errors. It is easily shown that an incorrectly estimated f_0 will produce a bias in all except the continuous term in (13). As this bias is linearly dependent on the f_0 estimation error, a simple readjustment of the coefficients will immediately improve the fit. Let us mention that this error compensation mechanism is also beneficial for properly capturing inharmonicities (deviation from theoretic harmonic frequencies) which are often present in audio signals.

Accordingly, we have chosen for our application a very simple and computationally efficient f_0 estimation method based on the Interpolated FFT [13]. Briefly, it consists of calculating the Discrete-time Fourier Transform of the signal around the fundamental frequency by using only two samples in the corresponding FFT. The main benefits of this method are efficient long and short-term leakage suppression and stability to additive noise and arithmetic roundoff errors. This method has been developed for stationary sinusoids but it has been proven heuristically to work well for most real-world audio signals which are inherently non-periodic.

Once f_0 has been estimated, the rest of the model parameters are obtained by solving a set of linear equations generated for each harmonic according to (12). The signal matrix expression and corresponding LS solution are:

$$\boldsymbol{s} = \boldsymbol{\Phi}\boldsymbol{p} + \boldsymbol{r}\,, \tag{14}$$

$$\hat{\boldsymbol{p}}^{LS} = \boldsymbol{\Phi}^{+}\boldsymbol{s}\,. \tag{15}$$

The vectors $\boldsymbol{s}$ and $\boldsymbol{r}$ contain the measurement data and additive noise respectively, while $\boldsymbol{\Phi}$ is built out of time-variant sine and cosine terms related to the estimated f_0. The matrix $\boldsymbol{\Phi}^{+}$ is the pseudo-inverse of $\boldsymbol{\Phi}$.

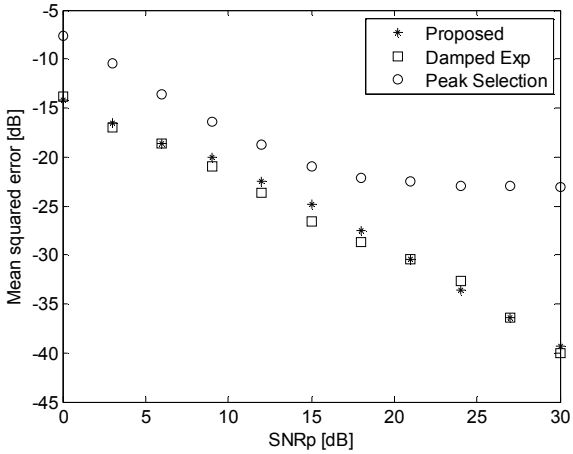

Figure 2 – Signal approximation root mean square error (RMSE). The SNR_p is calculated with reference to the highest harmonic.

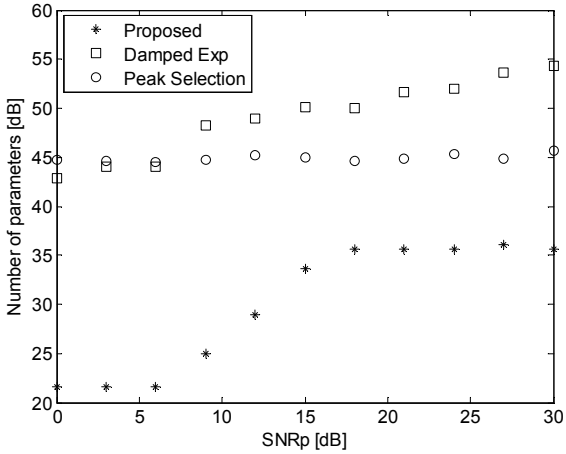

Figure 3 – Number of harmonic parameters to be estimated. The SNR_p is calculated with reference to the highest harmonic.

5. EXPERIMENTAL RESULTS

In this section we quantitatively and qualitatively evaluate the efficiency of the proposed signal modelling approach, through a comparative study and an illustrative example respectively.

For the comparative study we have chosen [7] (from now on the Exponential method) and [3] (from now on the Peak Selection method). The choice was motivated by the fact that both methods perform simultaneous audio signal component description by using very different analysis approaches. The Exponential method models the signal as a sum of stationary sinusoids modulated by a set of slowly time-varying damped exponentials. The Peak Selection method characterizes each peak in the spectrogram of the audio signal in terms of stochastic/non-stochastic.

The test signal $s_T(n)$ used for the comparative study has been carefully designed in order to best represent different real-world scenarios. For $n = 0,1, ..., N\text{-}1$ the signal $s_T(n)$ is:

$$s_T(n) = h_T(n) + t_T(n) + r(n) = \sum_{i=1}^{10} A_i(n)\cos\left[2\pi\, if_0 n + \varphi_i(n)\right] + r(n), \quad (16)$$

$$\varphi_i(n) = iA_{FM}\sin\left(2\pi f_{FM} n + \gamma\right) + \phi_i, \quad (17)$$

$$A_i(n) = \begin{cases} 0, & n < n_L \\ A_{AM}\dfrac{n - n_L}{n_T - n_L}, & n_L < n < n_T \\ \dfrac{A_0}{i}\left[1 + A_{AM}\cos\left(2\pi f_{AM} n + \delta\right)\right], & n_T < n < N - 1 \end{cases} \quad (18)$$

The first signal segment ($n < n_L$) is pure Gaussian noise with $n_L = 0.4N$. The energy time-evolution of the transient component is described by steep linearly-ascending amplitude with the following parameter values: $A_{AM} = 0.5$ and $n_T = 0.5N$. The harmonic component combines amplitude (AM) and frequency (FM) modulation through a sinusoidal law. As argued in [12] this kind of modulation allows a wide range of real-world AM and FM conditions to be covered. The harmonic parameters have been adjusted in such a way to assure the presence of the dominant mainlobe at the harmonic frequencies in the STFT. Correct operation of the Peak selection method depends on this condition, which is satisfied by letting $A_{FM} = 2$, $f_{FM} = (4N)^{-1}$ and $f_{AM} = 2F_{FM}$. The phase angles γ, δ and ϕ_i take arbitrary values in $[-\pi, \pi]$. The remaining parameters have been chosen as: $N = 1000$, $N_{FFT} = 2048$, $f_o =$ 1kHz, $f_s = 44.1$ kHz. The noise component is controlled through the *Peak Signal-to-Noise ratio (*SNR_p) which we define as the peak power of the highest harmonic above the neighbouring noise floor.

To evaluate the performance of the methods, we let the SNR_p vary in range [0, 30]dB and for each value we calculate the Root mean-squared error (RMSE) of the approximation over 100 realizations:

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}\left[h_T(n) + t_T(n) - \hat{h}_T(n) - \hat{t}_T(n)\right]^2}\ . \quad (19)$$

In case of The Exponential and proposed method, the model order selection can strongly influence the approximation accuracy. In order to establish a fair comparison, the model order that minimizes the RMSE for a given SNR_p has been determined experimentally for each method.

By examining the resulting curves on Figure 2 we observe that the Exponential and proposed method achieve very similar behaviour in presence of additive noise. The Peak Selection method, however, falls behind the Exponential and proposed method in the whole analysis range. For high SNR_p the difference of about 15dB is due primarily to failing to capture correctly the transient event. For middle and low SNR_p the difference is reduced to approximately 5dB because of the increased noise floor. At the same time, we calculate the number of harmonic parameters to be estimated as a function of the SNR_p and show the corresponding curves on Figure 3.

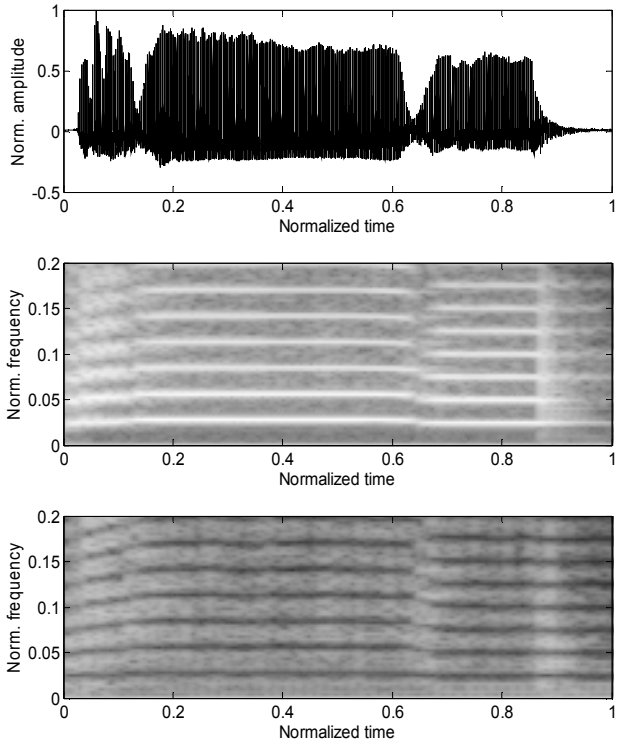

Figure 4 – (above) Trumpet signal; (middle) Signal spectrogram; (below) Residual spectrogram.

The proposed method is clearly the best, especially at low SNR_p where the gain with respect to the reference methods reaches 20dB. Under this condition the reference methods are comparable but as the noise impact gets smaller the Exponential method needs much more harmonic parameters to correctly represent the signal. Regarding the Peak Selection method, the number of parameters is roughly independent on the SNR_p because the total number of peaks in the STFT is similar from one experiment realization to another.

As an example of the qualitative behaviour of the proposed algorithm, we have considered a 250ms excerpt from a trumpet recording which contains short transients, as well as time-varying harmonic component. Figure 4 shows the time record, its spectrogram and spectrogram of the residual which is obtained by simply subtracting the estimated model (12) from the trumpet signal itself. We observe that the residual spectrogram exhibits the same line pattern as the signal spectrogram but in shades of gray, which means that most of the non-stochastic energy has been correctly removed. Note that the transient event at the beginning of the record is well captured by the model, in spite of the fact that it contains frequency modulation as well. By listening to the residual, we have detected no audible artefacts.

6. CONCLUSIONS

We have shown that both harmonic and transient behaviour in audio signals can be compactly described through a basic harmonic structure modified by the joint action of polynomial and piecewise linear approximation. The proposed model exhibits a high flexibility in reducing subharmonic errors and capturing inharmonicities in the signal. In addition, the parameters are easily computed as the model is linear-in-parameters. The experimental results show that the proposed method achieves a similar accuracy as one reference method, but using much less parameters and computational effort.

REFERENCES

[1] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis", *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 1, pp. 21-29, January 2001

[2] Q. Li, L. Atlas, "Time-variant least-squares harmonic modelling", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol.2, Hong-Kong, April 2003

[3] M. Lagrange et al, "Sinusoidal Parameter Estimation in a Non-Stationary Model", *Proceedings of the 5th International Conference on Digital Audio Effects*, Hamburg, Germany, September 2002

[4] J. J. Wells, D. T. Murphy, "High accuracy frame-by-frame non-stationary sinusoidal modelling", *Proceedings of the 9th International Conference on Digital Audio Effects,* Montreal, Canada, September 18-20, pp. 253-258, 2006

[5] C. Raphael and J. Stoddard, "Functional Harmonic Analysis Using Probabilistic Models", *Computer Music Journal*, 28(3), pp. 45–52, Fall 2004

[6] J. P. Bello, L. Daudet, S. Abdallah, C. Dixbury, M. Davies, M. B. Sandler, "A tutorial on onset detection in music signals", *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 5, pp.1035-1047, September 2005

[7] K. Hermus, W. Verhelst, P. Lemmerling, P. Wambacq, S. Van Huffel, "Perceptual audio modelling with exponentially damped sinusoids", *Signal Processing*, Vol. 85, pp. 163-176, 2005

[8] J. Jensen, R. Heusdens, S. H. Jensen, "A perceptual subspace approach for modeling of speech and audio signals with damped sinusoids", *IEEE Transactions on Speech and Audio Processing*, Vol. 12, No. 2, pp.121-132, March 2004

[9] S. Molla, B. Torresani, "Determining local transientness of audio signals", *IEEE Signal Processing Letters*, vol. 11, no. 7, July 2004.

[10] L. Daudet, B. Torresani, "Hybrid representations for audiophonic signal encoding", *Signal Processing*, Vol. 82, No. 11, pp. 1595-1617, 2002, Special Issue on Image and Video Coding Beyond Standards

[11] M. Vacher, D. Istrate, J. F. Serignat, "Sound detection and classification through transient models using wavelet coefficient trees", *Proceedings of the 12th Eusipco conference*, Vienna, Austria, September 2004

[12] M. Zivanovic, J. Schoukens, "Time-variant harmonic signal modeling by polynomial approximation and fully automated spectral analysis", *Proceedings of the 16th Eusipco conference,* Glasgow, UK, August 2009

[13] T. Grandke, "Interpolation algorithms for discrete Fourier transforms of weighted signals", *IEEE Transactions on Instrumentation and Measurement,* Vol.IM-32, No. 2, June 1983