16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

APPLICATION OF THE FAN-CHIRP TRANSFORM TO HYBRID SINUSOIDAL+NOISE MODELING OF POLYPHONIC AUDIO

Maciej Bartkowiak

Chair of Multimedia Telecommunications and Microelectronics, Poznan University of Technology
Polanka 3, 60-965, Poznan, Poland
phone: + (48-61) 6653850, fax: + (48-61) 6653899, email: mbartkow@multimedia.edu.pl
web: www.multimedia.edu.pl

ABSTRACT

Reliable classification of spectral peaks as tonal and noise-related is an important stage of hybrid sinusoidal+noise modeling. Spectral peaks of higher harmonics are often missed due to their wide frequency spread resulting from pitch variation. Recently introduced fan-chirp transform allows for compensating the changes of fundamental frequency in the process of spectral analysis of speech and harmonic sounds. In case of polyphonic audio the fundamental is often not unique and/or is hard to estimate. We propose a simple technique for estimation of chirp rates from already detected partials to improve the detection of higher harmonics through the application of frequency warping and fan-chirp analysis.

1. INTRODUCTION

Sinusoidal modeling is a well established signal processing framework applicable to speech and audio analysis, enhancement, restoration, source separation, automatic recognition, watermarking, compression, and synthesis [1]. Sinusoidal+noise (SN) modeling is an important member of the family of hybrid techniques that use different models to efficiently represent different classes of signal components. Within a SN model, a short segment of audio data is modeled as a sum of quasi-sinusoids with continuously varying magnitudes and frequencies (called the deterministic component), and a stochastic component (noise), whose short-time power spectra envelope changes over time,

$$\hat{x}(t) = \underbrace{\sum_{k=1}^{K} A_k(t) \sin\left(\varphi_k + 2\pi \int_0^t f_k(\tau)\, d\tau \right)}_{\text{deterministic component}} + \underbrace{h_n(t) * \xi(t)}_{\text{noise component}}. \quad (1)$$

In fact, this distinction is not as much critical from the perceptual point of view, as it is important due to the representation efficiency (in applications related to compression) and flexibility (in applications involving sound manipulations).

In general, the separation of the tonal (sinusoidal) and stochastic (noise) component is a difficult problem. First of all, the bulk of spectral components observed in natural audio exhibit only certain degree of coherence in time evolution of phase and instantaneous frequency. Consequently, most of them is neither purely sinusoidal nor purely random.

A common approach to the separation problem is to model the greater possible part of the signal energy by the deterministic component, under certain constraints (e.g. f_k being a harmonic series [2], that strongly narrows the range of applications). A residual signal is obtained by plain (time-domain) or spectral subtraction of the reconstructed sinusoids from the original signal. It is subsequently modeled as the stochastic component.

A more flexible approach is to perform a classification of spectral peaks (lobes surrounding local maxima of the magnitude short time spectrum) into tonal and non-tonal according to their shape. For example, Rodet [3] proposes a measure of sinusoidality based on complex cross-correlation of the short time spectra and the DFT of the analysis window. This approach is limited to stationary sinusoids, whereas time-varying components often exist in natural audio (fig.1).

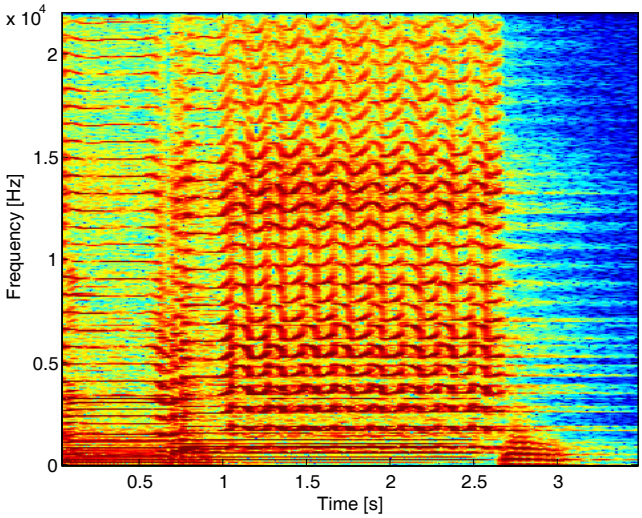

Figure 1 – Narrowband spectrogram of an example music excerpt showing a significant frequency spread of energy related to higher harmonics due to pitch variations. An analysis window of 4096 samples is necessary here to resolve low frequency partials

Lagrange et al [4] estimate the degree of local amplitude and frequency modulation using the time-frequency reassignment method of Auger and Flandrin [5]. Subsequently, individual spectral peaks are cross-correlated with a DFT of a distorted window function, and the degree of sinusoidality is determined and used in peak classification. Zivanovic et al [6,7] developed a peak classification system based on several local

spectrum descriptors: normalized bandwidth (NBD), normalized duration (NDD), frequency coherence (FCD). The distinction between sinusoidal peaks (main and side lobes) and noise is done upon the inspection of descriptor combined values.

The fundamental problem with all the approaches mentioned above is that they work under assumption that tonal energy manifests in the short time spectrum as a distinct peak, allowing a simple detection. In practice, such assumption hardly holds in case of instruments with free intonation (such as violin, trombone, etc), as shown in fig. 1, because variations of pitch cause the energy of higher partials to be spread over a wide frequency range and mutually overlap. The traditional DFT-based ML estimation method often fails at the task of musical spectrum analysis due to inappropriate underlying model that assumes local stationarity of partials.

Musical scales of many bass instruments start at 27Hz, the commonly used range begins at about 45Hz. High spectral resolution necessary for proper analysis of low pitched sounds requires the use of long DFT windows (60-100ms, i.e. 2^{11}-2^{12} samples if f_s = 44.1kHz) in order to reliably resolve individual partials (cf fig. 1). In a typical situation, instantaneous frequencies of partials change significantly during such a long period, thus they are no more observable as narrow spectral lines. Hence, it is reasonable to seek for locally-adaptive TF analysis methods [5,8,9] that commonly attempt at modeling the non-stationary spectral content on a chirp basis.

Among many chirp transforms and chirp estimation techniques proposed hitherto which often exhibit high computational complexity, the Fan-chirp transform (FChT) introduced by Kepesi and Weruaga [10,11] offers two fundamental advantages in the context of music analysis. It allows for simultaneous adapting to the pitch variations of all harmonics of given sound, and its computational complexity is very low, enabling online processing.

Developed primarily for the analysis of speech, FChT computes the spectrum of a signal on the set of basis functions with fan-like geometry in the time-frequency plane. The short-time fan-chirp transform (STFChT) is defined as

$$X(k,\alpha)=\sum_{n=0}^{N-1} x(n)\sqrt{\phi_\alpha'(n)}\,\exp\left(-\frac{j\,2\pi k\,\phi_\alpha(n)}{N}\right), \quad (2)$$

where $\phi_\alpha(n)$ is a time-frequency warping operator,

$$\phi_\alpha(n)=\left(1+0.5\,\alpha\,(n-N)\right)n, \quad (3)$$

and α is the skew parameter corresponding to the chirp rate.

In fact, the STFChT of a given signal is equivalent to the DFT of the same signal sampled on a non-uniform grid obtained by inverting the warping operator (3). Therefore a fast implementation is possible which requires just a resampling step followed by an FFT [11]. Since the mapping (3) is bijective in [0..*N*], the transform is reversible, provided no aliasing terms are introduced in the process of resampling. These aliasing terms may be avoided by appropriate upsampling of the original signal prior to warping.

2. MODELING OF POLYPHONIC MUSIC

2.1 The problem of fundamental frequency

STFChT is able to resolve harmonic partials whose frequency deviation within the analysis window is greater than spacing between corresponding mean frequencies. It is possible under the condition that an appropriate value of α is used, that corresponds to the rate of change of the fundamental frequency, and $|\alpha|< 2/N$. In the context of speech analysis, it may be approximated as (4)

$$\alpha=\frac{f_0'(t)}{f_0(t)}\cong\frac{f_0(n+1)-f_0(n-1)}{2f_0(n)}, \quad (4)$$

where $f_0(n)$ denotes a fundamental frequency estimated within a symmetric time window centered around *n*. Several techniques for the FChT-supported estimation of fundamental using either inter-frame or intra-frame approach are described in [10].

In the context of polyphonic music, f_0 is not unique due to the presence of multiple sounds of different pitch, often generated by different instruments. The issue of multiple pitch estimation from polyphonic audio has been addressed by many researchers (e.g. [12,13]) and is generally considered as a difficult task. Furthermore, some musical instruments (like bells, glockenspiel or Rhodes piano) exhibit strongly inharmonic spectra, therefore their fundamental is undefined. It is important to note however, that even without a strictly defined fundamental all the sinusoidal partials of pitched sounds follow a similar pattern in the time-frequency plane. Considering partials of a harmonically rich sound, their individual chirp rate estimates taken relative to their mean frequency estimates are strictly related to the pitch change rate. Therefore, instead of (4), α may be estimated from the statistics of individual chirp rates α_k of some lower frequency partials detected before calculating the FChT. It is a feasible solution, since low partials usually exhibit more stable frequencies and are relatively easy to detect.

2.2 Estimation of individual partials

Partials with a limited depth of frequency modulation may be often (but not always) modeled as linear chirps. It is possible to estimate their mean frequency and individual chirp rate by using one of several techniques developed for sinusoidal modeling. For example, Abe and Smith [14] demonstrated that for a chirp expressed as

$$x(t)=A_0\,\exp\left(\gamma_0 t+j(\varphi_0+\omega_0 t+\beta_0 t^2)\right), \quad (5)$$

weighted by a Gaussian window (as well as other windows), a non-zero frequency modulation term β_0 results in a quadratic shape of log amplitude and phase spectra. They proposed a quadratic-interpolated FFT method for estimating the ω_0 and β_0 ,

$$\hat{\omega}_0=\frac{2\pi}{N}\left(k_0-\frac{b}{2\,a}\right),\ \hat{\beta}_0=p\,\frac{d}{a}, \quad (6)$$

from the parameters of a parabola fitted to the log magnitude and phase spectrum surrounding peaks,

$$a = \left(\log\left|X_{k_0+1}\right| - 2\log\left|X_{k_0}\right| + \log\left|X_{k_0-1}\right|\right)/2$$
$$b = \left(\log\left|X_{k_0+1}\right| - \log\left|X_{k_0-1}\right|\right)/2 \quad , \qquad (7)$$
$$d = \left(\angle X_{k_0+1} - 2\angle X_{k_0} + \angle X_{k_0-1}\right)/2$$

where

$$p = -\frac{\pi^2}{N^2}\frac{d}{a^2+b^2} \qquad (8)$$

and k_0 is the index of FFT bin corresponding to local maximum of magnitude.

2.3 Chirp rate estimation for groups of partials

We propose a two-stage analysis procedure for sinusoidal modeling of polyphonic music. The main idea is to perform a standard analysis first, with the use of DFT for the detection of reliable low frequency partials and estimation of their parameters ω_0 and β_0. Subsequently, the non-stationary high frequency partials are detected and their parameters are estimated by the use of FChFT analysis, taking into account several most "interesting" values of chirp rate α, i.e. those values that most probably correspond to the local time-frequency skewness related to the underlying pitch modulation.

Let assume sounds coming from different instruments with different pitch variation are present simultaneously in the current analysis frame. Its spectrum shows a mixture of harmonic and inharmonic series of partials. We observe that the estimates of individual chirp rates $\alpha_k = \beta_k/(2\omega_k)$ of individual partials follow a multi-modal distribution that may be approximated by a Gaussian mixture model (GMM),

$$p_\alpha(\alpha) = \frac{\sum_m w_m p\left(\alpha|\phi_m\right)}{\sum_m w_m}, \text{ where} \qquad (9a)$$

$$p\left(\alpha|\phi_m\right) = \frac{1}{\sigma_m\sqrt{2\pi}}\exp\left(-\frac{(\alpha-\mu_m)^2}{2\sigma_m^2}\right), \text{ and} \qquad (9b)$$

ϕ_m denotes a certain "state" of the model representing a group of partials sharing a common chirp rate. The weights w_m are not explicitly known, but may be regarded as representing the bulk of partials exhibiting similar temporal evolution, thus they may be estimated from the heights of the empirical distribution modes. Our aim is to find the values of μ_m which are the interesting chirp rates that may reveal additional high frequency partials due to the time-frequency warping inherent in the FChFT.

We estimate the values of μ_m by employing an iterative algorithm based on the Expectation Maximization method. The algorithm starts with a classical sinusoidal analysis of a given audio frame with an optional peak verification in order to reject peaks induced by noise [6,7]. Initially, for each frame we gather the observed values of $\alpha_k=\beta_k/(2\omega_k)$ and form a pdf estimate (fig. 2) by the use of a histogram smoothing method [15]. Locations of peaks of this pdf estimate are the initial estimates of μ_m which may be iteratively improved as follows:

- For each sample of α calculate its distance to every μ_m.
- Calculate new estimations of μ_m through weighted averaging the values of α with the weights inversely proportional to the distances.
- Iterate until there is no significant change in μ_m.

Results of such iterations (fig. 2) are the values of the chirp rates that may be applied within the second stage employing FChFT analysis for enhanced estimation of non-stationary high-frequency partials. We have observed experimentally that for real world music the values of α are usually constrained in the range of <-1…1>, and most often do not exceed 0.5.

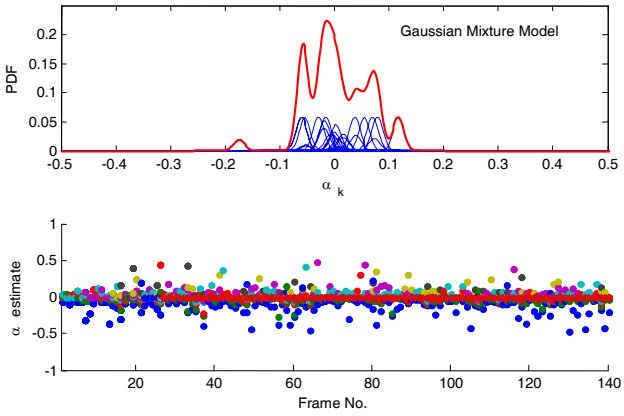

Figure 2 – Above: distribution of the estimated values of α_k for a single frame of the test signal (fig. 1). Below: estimated values of α in consecutive frames.

2.4 FChFT-based music analysis

Music spectrum analysis with the FChFT transform offers the possibility to reveal otherwise hidden spectral peaks related to non-stationary high frequency partials. It also offers an enhanced estimation of the parameters of lower partials due to the frequency deviations being compensated by the time-frequency skew inherent in the fan-chirp basis functions. Thanks to the chirp rates α being estimated in the first stage of the proposed technique (sec. 2.3), it is necessary to calculate the FChFT only for those few values of α, which is a computationaly feasible operation.

Our peak detection and estimation algorithm depends on the observation that for each sinusoidal partial with varying frequency the highest value of corresponding peak in the magnitude spectrum is offered by the output of the FChFT with such value of the α parameter that is closest to the first-order approximation of the real frequency variation function. In other words, the closest is the chirp rate used in the fan-chirp transform to the real frequency change rate, the more is the spectrum similar to a spectrum of a sinusoid.

The algorithm for the analysis is very straightforward:

1. For a data segment x of N samples, initialize a vector of peaks $P[k]$ with N/2 zeros. Also, insert peak values estimated from the DFT analysis in the first stage into the locations corresponding to the DFT bin numbers.
2. For the first candidate value of α estimated as described in 2.3, calculate the result of FChFT(x,α).

3. Find all sinusoidal peaks in the FChFT output, according to the chosen peak detection criteria.
4. For each of those peaks compare their magnitude to the magnitude of corresponding peak already gathered in the vector *P*. If the magnitude is higher, it means that a better approximation of corresponding partial is found. In such case, replace the existing peak with the new peak from the FChFT result. Also, collect the neighboring spectral data and write it to the entries of *P*. Label the peak with the current chirp rate, α.
5. Iterate steps 2..4 with subsequent values of α from the set.
6. For each of the peaks gathered in the vector *P*, calculate the corrected ω_0 and β_0 , according to (6-8). Correct the value of β_0 by taking into account the chirp rate α used for the particular peak.

Note that the above procedure does not guarantee that all hidden partials are detected. Unfortunately, some groups of highly non-stationary sinusoids may be missed if none of them have been detected in the first stage so that it could contribute to the estimation of optimal skew parameter α.

3. EXPERIMENTAL RESULTS

2.1 Synthetic signal

In order to verify the procedure proposed in section 2.3, a simple test has been set up. An artificial signal has been constructed by summing two inharmonic spectra of two bell sounds with deeply modulated pitch, synthesized using the FM synthesis technique (fig. 3). As it can be easily observed, the deep frequency deviation causes most of the high frequency partials to be blurred to a significant degree. Clearly, this signal spectrum contains at least two groups of partials and the distribution of α_k should reveal in each frame at least two modes of the pdf, corresponding to the different frequency modulation patterns.

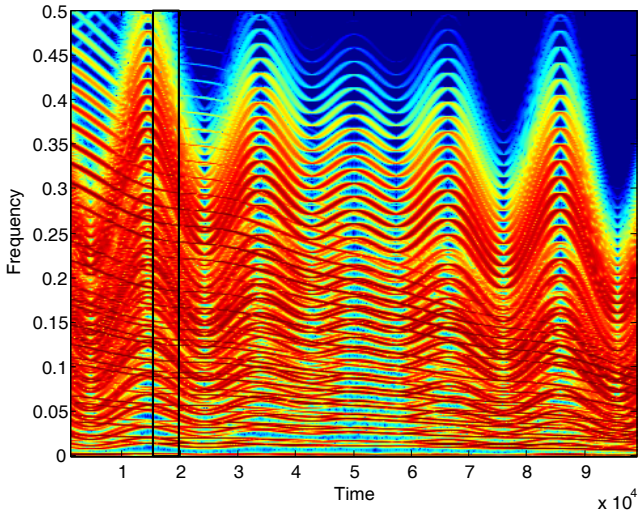

Figure 3 –Spectrogram of the synthetic signal. The black frame shows a data segment of N=4096 samples, further analyzed in fig. 4

Experiments show that for this synthetic signal about 20 lowest harmonics are detected reliably in the first stage of sinusoidal analysis. In fact, due to overlapping partials, the estimation of ω_k and β_k is not free of errors, therefore the actual values of α are slightly biased. Resulting chirp spectra are shown in fig. 4.

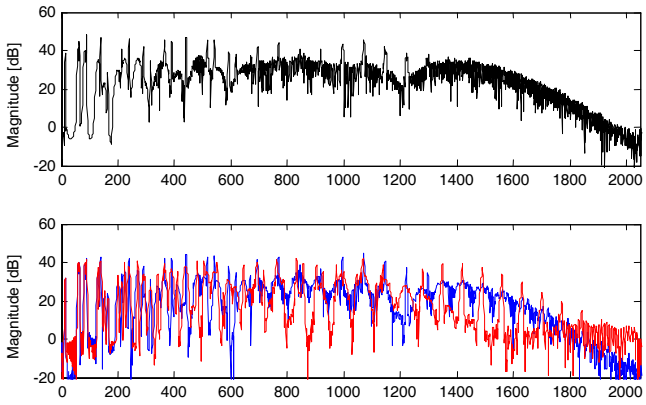

Figure 4 – Comparison of standard DFT (above, black) and FChFT analysis (below) with two values of automatically estimated α (shown in blue and red). In both cases, the analysis window is Hamming, 4096 samples.

As it can be clearly seen, most of the high-frequency partials that are entirely indiscernible in the DFT output become quite visible in the result of FChFT. It is important to note that fan-chirp analysis allowed to discriminate partials that are very close in frequency, but differ mostly in the chirp rate, α_k.

2.2 Analysis of real music

A series of experiments with various excerpts of popular and classic music from the EBU SQAM reference CD have been performed in order to verify the effectiveness of the new sinusoidal analysis technique in real-life applications. In each experiment, a benchmark was created from the results of standard sinusoidal analysis with an additional peak selection procedure based on spectral descriptors (NBD+FCD). Results of FChT-based analysis compared favorably with the benchmark, since many existing partials have been detected in the high frequency range. Moreover, more robust peak detection due to the chirp analysis allowed for changing the detection thresholds to more strict setting. Thanks to this, there was much less of false partials detected due to the spectral energy induced by noise.

A sample comparison from these experiments is shown in fig. 5. In the upper plot we show the sinusoidal partials detected with the standard DFT method followed by peak classification based on spectral descriptors. This typical result reveals serious deficiencies of the analysis technique. Most of the high frequency partials have not been properly detected due to the vibrato modulation, while there are many false multiple partials in the range of 12-15kHz induced by the irregular spectral peaks which are the side lobes of deeply modulated harmonics. It is worth mentioning, that a simpler analysis without application of spectral descriptors (not shown here) gives even worse results. In the lower plot, the results of FChFT-based analysis show much of the partials in

the high frequency range being properly detected, and also the number of false partials is significantly reduced.

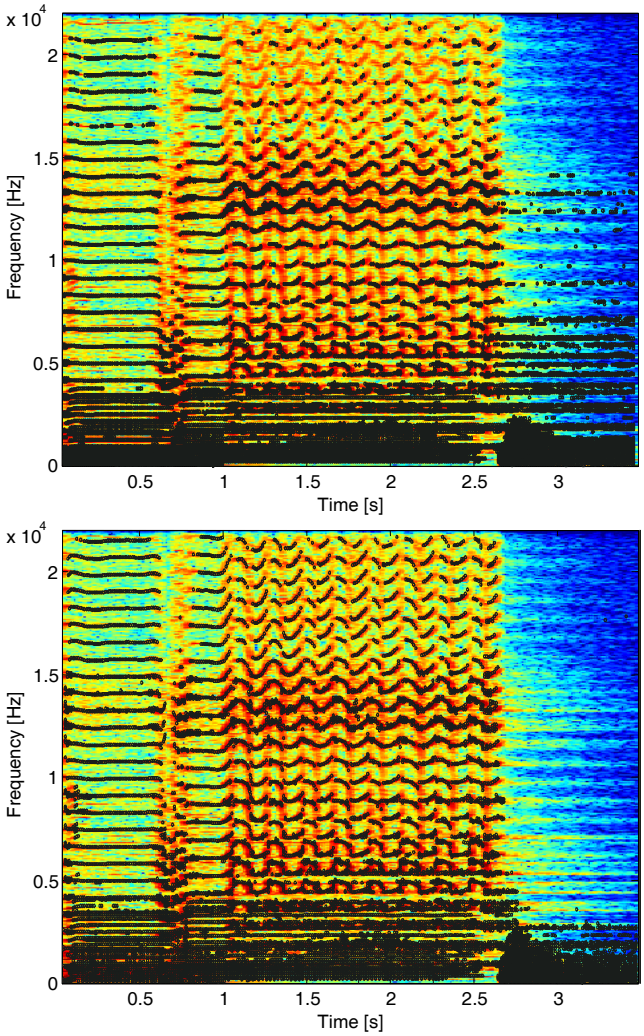

Figure 5 – Comparison of sinusoidal partial detection based on standard DFT technique (above) and the proposed technique exploiting fan-chirp transform analysis (below). These plots should be compared with figure 1.

One significant disadvantage of the proposed new technique for sinusoidal analysis is the additional computational burden related to the necessary calculations of several fan chirp transforms. However, detection and estimation are usually not very demanding in terms of computational complexity, compared to tracking, whose complexity is often data-dependant. Since our analysis results in much "cleaner" the data input to the tracking algorithm, the total operation speed of a modeling system may not increase significantly.

4. CONCLUSIONS

A computationally feasible application of the fan-chirp transform to hybrid sinusoidal+noise modeling of polyphonic music have been presented in the paper. A very simple technique has been proposed for estimation of the frequency warping parameter α that does not require pitch estimation. Experimental results confirm, that a substantial improvement in the detection of highly non-stationary partials is achieved, that enables a good quality modeling of wideband audio, without restrictions regarding harmonicity.

5. ACKNOWLEDGMENTS

This work was supported by the research grant 3 T11D 017 30 of the Polish Ministry of Science and Higher Education.

REFERENCES

[1] J.Beauchamp (red), *Analysis, Synthesis, and Perception of Musical Sounds: The Sound of Music*, Springer, 2006.

[2] X. Serra, J.O.Smith, "Spectral modelling synthesis: A sound analysis/synthesis system based on deterministic plus stochastic decomposition", *Computer Music Journal*, 14(4), 1990, pp. 12-14.

[3] X.Rodet, "Musical sound signal analysis/synthesis: Sinusoidal + residual and elementary waveform models", *IEEE Time-Frequency and Time-Scale Workshop*, *TFTS'97*, Coventry, UK, August 1997.

[4] M.Lagrange, S.Marchand, J-B.Rault, "Sinusoidal parameter extraction and component selection in a non-stationary model", *Proc. DAFx'02,* Hamburg, 2002, pp. 59-64.

[5] F. Auger, P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method", *Proc. ICASSP'95*, May 1995, vol. 4, pp. 1068-1089.

[6] A. Röbel, M.Zivanovic, X.Rodet, "Signal decomposition by means of classification of spectral peaks", *Proc. ICMC'04*, Miami, 2004.

[7] M.Zivanovic, A. Röbel, X.Rodet, "Adaptive threshold determination for spectral peak classification", *Proc. DAFx'07*, Bordeaux, 2007.

[8] S.Mann, S.Haykin, "Adaptive 'chirplet' transform: an adaptive generalization of the wavelet transform", *Optical Engineering*, vol.31, no.6, pp. 1243-1256, June 1992.

[9] X-G. Xia, "Discrete chirp-Fourier transform and its applications to chirp rate estimation", IEEE Trans. Sig. Proc, vol.48, no.11, pp. 3122-3133, November 2000.

[10] M. Kepesi, L. Weruaga, "Adaptive chirp-based time-frequency analysis of speech signals", *Speech Comm*., vol.48, pp. 474-492, 2006.

[11] L. Weruaga, M. Kepesi, "The fan-chirp transform for non-stationary harmonic sounds", *Signal Proc.*, vol. 87, pp. 1504-1522, 2007.

[12] P.J.Walmsley, S.J.Godsill, P.J.W.Rayner, "Polyphonic pitch tracking using joint Bayesian estimation of multiple frame parameters", *Proc. IEEE Workshop on Audio and Acoustics*, Mohonk, NY State, 1999

[13] Y. Chunghsin; A. Röbel, X. Rodet, "Multiple fundamental frequency estimation of polyphonic music signals", Proc. ICASSP '05, March 2005, vol.3 , pp. 225-228.

[14] M. Abe, J.O. Smith, "Design criteria for the quadratically interpolated FFT method (III): Bias due to amplitude and frequency modulation", *CCRMA Rep. STAN-M-116*, October, 2004.

[15] W. Hardle, *Smoothing Techniques: With Implementation in S*, Springer-Verlag, Berlin, 1990