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. 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 multiple voting of already detected partials to improve the detection of higher harmonics through fan-chirp frequency warping.

1. INTRODUCTION

Sinusoidal modeling is a well established signal processing tool 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 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.

The common approach to the separation 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, that strongly narrows the range of applications [2]). 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). 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.

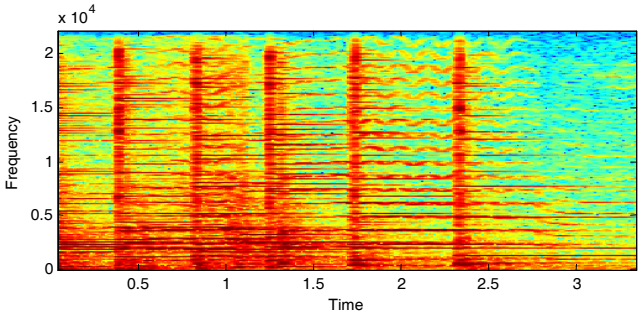

Figure 1 – Narrowband spectrogram of an example music excerpt showing a significant frequency spread of energy related to higher harmonics due to pitch variations.

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 assump-

tion 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.
High spectral resolution required for proper analysis of low-pitched sounds (sometimes below 65Hz) enforces the use of long analysis windows (60-100ms, i.e. 2^{11}-2^{12} samples if $f_s = 44.1$kHz) in order to reliably resolve individual partials. Classical DFT-based analysis often fails at this task, due to inappropriate model of local stationarity applied to music. It is thus reasonable to seek for locally-adaptive TF analysis methods [5,8,9].

Among many chirp-based transforms and chirp estimation techniques proposed hitherto for analysis of non-stationary signals, the fan-chirp transform (FChT) introduced by Kepesi and Weruaga [10,11] offers two fundamental advantages. It allows for simultaneous adapting to the pitch variations of all harmonics of given sound, and it's 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). Thus, 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

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

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 non-harmonic 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 are strictly related to the pitch change rate. Therefore, instead of (4), α may be estimated by multiple voting of individual chirp rates α_k of some 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 quadratically-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,

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

where

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

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

2.3 Chirp rate estimation for groups of partials

Let assume sounds coming from different instruments with different pitch variation are present simultaneously. The individual estimates of $\alpha_k = \beta_k / (2\omega_k)$ follow a multi-modal distribution that may be approximated by a mixture of Gaussians. The modes of this distribution correspond to rates of change of individual pitches and thus the candidate values of α may be estimated without knowing the actual pitch value.

The main assumption is that there are at least few partials belonging to a certain sound that can be detected without a chirp transform. The main idea is to perform a multi stage analysis in order to resolve distinct groups of partials one by one.

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]. The sinusoidal analysis involves ω_k and β_k estimation according to (6). Few stages of subsequent pitch change rate estimation follow. At every stage, the value of α that best fits to a dominant group of partials is determined by seeking the highest mode of the distribution of already gathered sinusoidal data (fig. 2). In order to make the statistical model more reliable we introduce additional weighting of the estimates coming from individual peaks. Thus,

$$\alpha = \arg\max \frac{\sum_k \Psi_k \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\left(\alpha - \hat{\beta}_k / (2\hat{\omega}_k)\right)^2 / (2\sigma^2)\right]}{\sum_k \Psi_k}, \quad (9)$$

where Ψ_k denotes weights of individual peaks, for example a measure of sinusoidality.

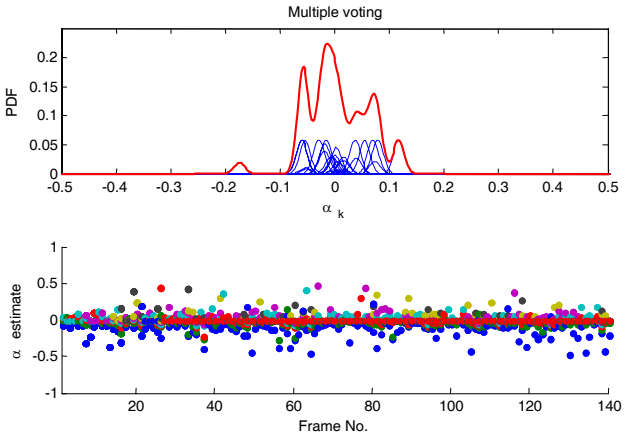

Figure 2 – Above: distribution of the estimated values of α_k for a single frame of the test signal (fig. 1). Below: estimated values of α in consecutive frames.

After performing a FChT-based analysis with the estimated α, all partials that have been successfully identified as sinusoidal (cf fig. 3) are subsequently removed from the chirp-spectrum. A residual signal is calculated through an inverse FChT. The whole procedure is repeated, until all modes of the distribution are examined, and no new partials are discovered. Note that this procedure does not guarantee that all sinusoids are detected. Unfortunately, some groups of highly nonstationary partials may be missed if none of them have been detected in subsequent stages so that it could contribute to the voting of optimal warping 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 non-harmonic spectra of two bell sounds synthesized using the FM synthesis technique with linearly gliding pitch at significantly different slopes (fig. 3). Clearly, this signal spectrum contains at least two groups of partials and the distribution of α_k should reveal two modes corresponding to the frequency sweep rate.

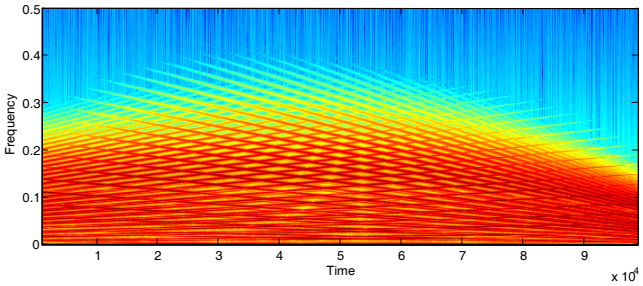

Figure 3 –Spectrogram of the synthetic benchmark signal.

Experiments show that for this synthetic signal at least ten lowest harmonics are detected reliably in the preliminary 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.

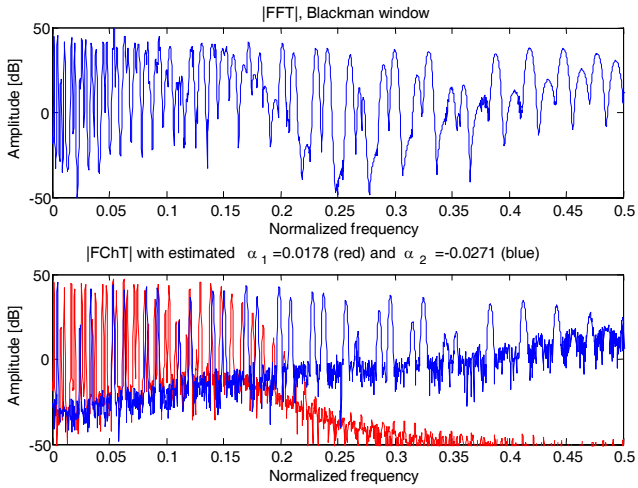

Figure 4 – Comparison of standard DFT (above) and STFChT with two values of α estimated by multiple voting of (9) - below.

Despite the observed estimation error of α, most of the higher harmonics that are missed by DFT-based detection have been detected correctly in case of FChT. 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, β_k.

2.2 Analysis of real music

A series of experiments with various excerpts of popular and classic music have been performed in order to verify the effectiveness of the new peak detection approach 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

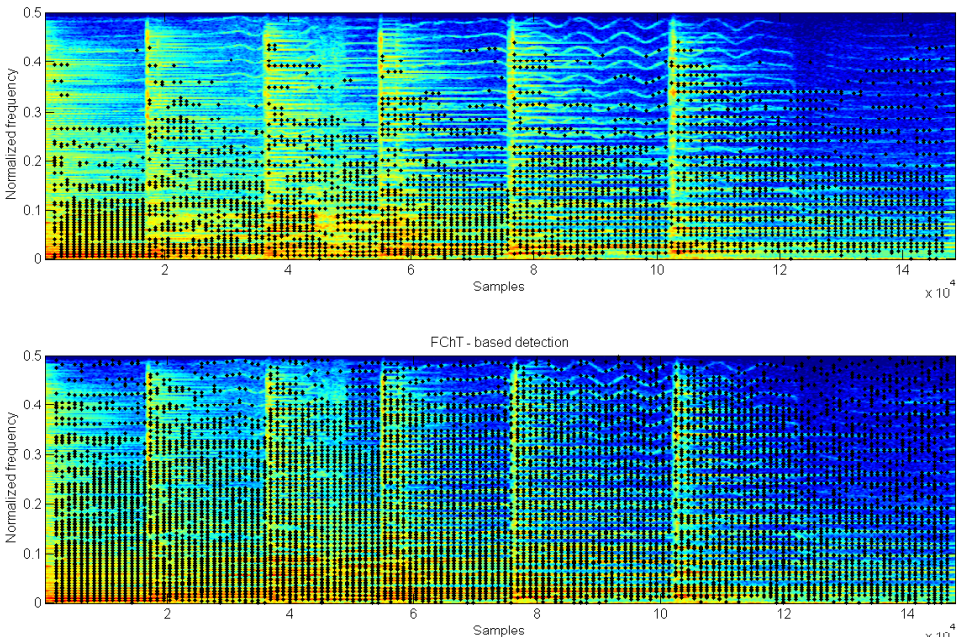

Figure 5 – Comparison of sinusoidal partial detection based on standard DFT technique (above) and the proposed technique exploiting fan-chirp transform analysis (below).

(NBD+FCD). Results of FChT-based analysis compared favorably with the benchmark, since many additional partials have been detected (fig. 5). Although still several partials are missed, the most noticeable improvement is in highly-nonstationary partials, especially in the high frequency range.

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 indicate, that a significant improvement in the detection of highly-nonstationary partials has been chieved, that enables a good quality modeling of wideband audio, without restrictions regarding harmonicity.

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.