Improving Spectral Analysis Precision with an Enhanced Phase Vocoder using Signal Derivatives

Sylvain Marchand

SCRIME, LaBRI, Bordeaux I University
351, cours de la Libération
F-33405 Talence Cedex, France

sm@labri.u-bordeaux.fr
http://dept-info.labri.u-bordeaux.fr/~sm

Abstract

The purpose of this presentation is to demonstrate the practical interest of an original improvement of the classic Fourier analysis. The n-th order short-time Fourier Transform (FT^n) extends the classic short-time Fourier transform by also considering the first n signal derivatives. This technique greatly improves Fourier analysis precision not only in frequency and amplitude but also in time, thus minimizing the well-known problem of the trade-off of time versus frequency. The implementation of this analysis method leads to an enhanced phase vocoder particularly well-suited for extracting spectral parameters from the sounds.

1 Introduction

In order to faithfully imitate and also transform existing sounds using a computer, a formal representation is needed for these sounds. Spectral models provide general representations in which such operations can be performed in a very natural and musically expressive way. Such models require an accurate analysis method to extract spectral parameters from sounds which were usually recorded in the temporal model, that is audio signal amplitude as a function of time. The accuracy of the analysis method is extremely important since the perceived quality of the resulting sounds depends mainly on it.

In spite of its many drawbacks, the short-time Fourier transform is often used in the very first step of the analysis process. The purpose of this presentation is to show the practical interest of an original improvement of the classic Fourier analysis. The n-th order short-time Fourier Transform (FT^n) takes advantage of the first n signal derivatives in order to improve the precision of the Fourier analysis not only in frequency and amplitude but also in time, thus minimizing the problem of the trade-off of time versus frequency in the classic short-time Fourier transform.

After introducing in section 2 the sound model which is considered in the rest of this paper, we summarize in section 3 the principles of the FT^n method and the way to implement it for $n = 1$, section 4 presents the precision enhancements achieved, and section 5 gives some results obtained using an enhanced phase vocoder based on this method.

2 Sound Model

Analyzing a sound means extracting parameters from it according to a certain mathematical representation of the sound: a sound model. Defining the model precisely is necessary before describing the analysis method itself.

Since spectral models parameterize sound at the basilar membrane of the ear, the resulting sound transformations are closely linked to the acoustic perception. Xavier Serra and Julius O. Smith III propose in [1] a spectral model based on a deterministic plus stochastic decomposition. This is the model mainly considered here, except that the equations have been reformulated for the purposes of homogeneity and some hypotheses have been modified. This sound model decomposes any audio signal in two parts: a deterministic part consisting of sinusoids, plus a stochastic part also called noise.

The present work focuses on a restriction of this model assuming that the noise component can be neglected. In practice, this restriction means that the considered sounds should have a low noise level, which is true for many clear natural sounds.

The remaining deterministic part consists of a sum of sinusoidal oscillators (*partials*) for which frequency and amplitude evolve in a slow time-varying manner. More formally, the expression of an audio signal a is given by the following equations:

$$a(t) = \sum_{p=1}^{P} a_p(t) \cos(\varphi_p(t)) \qquad (1)$$

with
$$\frac{d\varphi_p}{dt} = 2\pi \ f_p(t) \qquad (2.1)$$

i.e. $$\varphi_p(t) = \varphi_p(0) + 2\pi \int_0^t f_p(u)\, du \qquad (2.2)$$

where t is time expressed in seconds, P is the number of partials, f_p, a_p, and φ_p are the frequency, amplitude, and phase of the p-th partial respectively.

The initial phases (for $t = 0$) will be ignored during analysis and can arbitrary be set to 0 (zero) for resynthesis (this choice can be done according to psychoacoustic experiments which are beyond the scope of this presentation).

Another restriction is that the partials have to be sufficiently spaced in frequency, i.e. given any sound a there must exist a minimal distance $d > 0$ so that:

$$\min_{i \neq j, t} \{ |f_j(t) - f_i(t)| \} > d \qquad (3)$$

This condition, which also prevents two partials frequencies from « crossing », is a reasonable hypothesis verified for almost every monophonic natural sound. The reasons why it is needed will be discussed in the next section.

3 Principles of FT^n

An exhaustive theoretical presentation of the FT^n method is beyond the scope of this paper. This presentation can be found in a research report by Myriam Desainte-Catherine and Sylvain Marchand [2] submitted for publication. However this section summarizes its most important points for $n = 1$, since this is necessary for the understanding of the results exposed in the rest of this presentation. Basic knowledge of the classic short-time Fourier analysis is also required.

In the sound model which is considered in this paper both frequency and amplitude are slow time-varying parameters, so that during a single analysis window of the short-time Fourier transform the frequency and amplitude derivatives are close to 0 (zero). Under such conditions using the first signal derivative can help improving Fourier analysis precision both in frequency and amplitude. The idea behind this technique is extremely simple: derivating a sine gives a sine, with a different phase but the same frequency.

3.1 Signal Derivatives

Since a_p is slow time-varying, let us assume that its derivative is 0 (zero). From *Equations 1* and *2.1*:

$$\frac{da}{dt}(t) = \sum_{p=1}^{P} 2\pi\, f_p(t)\, a_p(t) \cos(\varphi_p(t) - \frac{\pi}{2}) \qquad (4)$$

This section summarizing the first order Fourier analysis method (FT^1), only the first signal derivative must be examined.

3.2 Spectral Parameters

In practice the audio signal a is uniformly sampled at rate R. Let us note DFT^k the amplitude spectrum of the Discrete Fourier Transform of the k-th signal derivative, computed using N consecutive samples from a certain location l. More formally:

$$DFT^k[m] = \frac{1}{N} \left| \sum_{n=0}^{N-1} w[n] \frac{d^k a}{dt^k}[l+n]\, e^{-j\frac{2\pi}{N}nm} \right| \qquad (5)$$

where w is an N-point analysis window.

A consequence of *Equation 1* and *Equation 4* is that for each partial p there is a maximum in both DFT^0 and DFT^1 spectra for a certain index m_p. Note that DFT^0 is the classic short-time Fourier analysis, and using only DFT^0 leads to the classic phase vocoder. Very good introductory texts on this subject can be found for example in [6], [7], or [8]. Approximate frequency and amplitude values are:

$$f_p^0 = m_p \frac{R}{N} \qquad (6)$$

$$a_p^0 = DFT^0[m_p] \qquad (7)$$

Taking advantage of DFT^1, a much more accurate frequency value can be obtained using the equation:

$$f_p^1 = \frac{1}{2\pi} \frac{DFT^1[m_p]}{DFT^0[m_p]} \qquad (8)$$

It is extremely important to note that the effects of any analysis window are the same on both DFT^0 and DFT^1 as soon as the same analysis window is used to compute these two spectra, and these effects are compensating thanks to the division in the preceding equation. The accurate partial amplitude is:

$$a_p^1 = \frac{a_p^0}{W(|f_p^1 - f_p^0|)} \qquad (9)$$

where $W(f)$ is the amplitude of the continuous spectrum of the analysis window w at frequency f. With the FT^1 method this window should be chosen as small as possible. The only condition is that two frequencies must lie in two different Fourier transform bins, which is always possible thanks to the model restriction defined in *Equation 3*. When this condition is not satisfied, a bin contamination occurs and makes *Equation 8* false.

4 Analysis Precision

The analysis window w has a great impact on the analysis precision in both frequency and amplitude. An exhaustive discussion about analysis windows is beyond the scope of this paper, and can be found in [3]. Before presenting some results of the FT^1 method on complex natural sounds, we point out some limitations and imprecisions of the classic Fourier analysis - fixed by the FT^1 analysis - using a single sinusoidal oscillator. Of course this example is synthetic but this is a real reference example for any analysis. Indeed many sounds consist of a sum of partials and the analysis process is a linear operation. *Figure 1* and *Figure 2* show the results of the short-time Fourier analysis on a single sinusoidal oscillator for which frequency is linearly increasing while its amplitude remains constant.

4.1 Frequency Precision

With the FT^1 analysis the evolutions of the partial frequency shown in *Figure 1* are almost perfectly recovered even with a very small N thanks to *Equation 8*. Such a result would have been impossible to achieve with the classic short-time Fourier analysis since a large analysis window is needed to have such a great frequency precision, in which case the time resolution is so bad that the evolution of the frequency with time can not be successfully recovered. That is the reason why the classic phase vocoder yields poor results when analyzing sounds with vibrato.

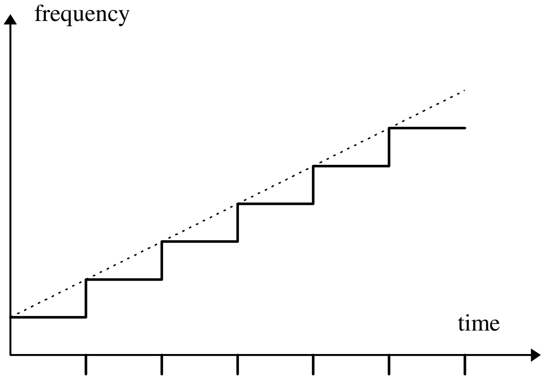

Figure 1: Original (dashed) versus Fourier analyzed (solid) frequency evolutions for a single sinusoidal oscillator for which frequency is linearly increasing while its amplitude remains constant. (The marks on the time axis indicate when the oscillator frequency goes from one bin to the other.) The analyzed frequency curve is not a line as it should be, but a sort of stairs, due to spectrum sampling.

4.2 Amplitude Precision

With respect to amplitude, again the FT^1 analysis accurately recovers the evolutions of the partial where the classic Fourier analysis has failed. The effects of windowing on the amplitude are almost completely cancelled thanks to *Equation 9*. With the classic short-time Fourier analysis, the analyzed amplitude curve is not flat, i.e. not a constant, but a succession of bumps, due to the shape of the analysis window mainlobe. This is why the classic phase vocoder can perform poorly when analyzing sounds with tremolo.

Of course such little deformations can not generally be heard, but they may become dramatically audible as soon as some transformations are performed.

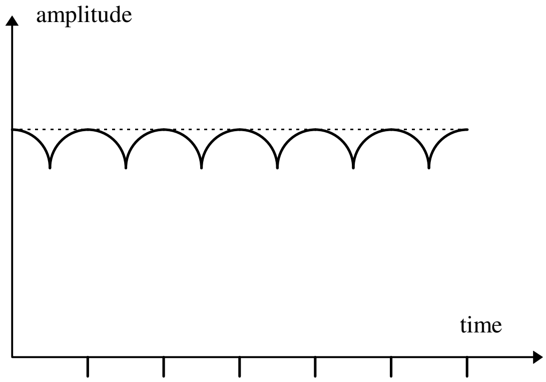

Figure 2: Original (dashed) versus Fourier analyzed (solid) amplitude evolutions for the same oscillator as in *Figure 1*. The analyzed amplitude curve of the oscillator is distorted as its frequency goes from bin to bin, due to the shape of the analysis window mainlobe.

5 Results on Natural Sounds

The short-time Fourier analysis is present in the very first step of many spectral analysis methods like the McAulay-Quatieri analysis [4] used in Lemur [5] and Spectral Modeling Synthesis (SMS) [1]. When no precautions are taken, the imprecisions pointed out in the previous section appear.
The FT^n analysis method has been implemented as a part of a sound analysis software package running on different platforms, called InSpect [9]. This program features accurate partial tracking for low-noise sound. It also performs basic sound operations (time compression and expansion, amplitude envelope extraction, etc.) and allows resynthesis according to the spectral model considered. In the first step of its analysis algorithm, the classic phase vocoder based on the short-time Fourier transform has been replaced by an enhanced version using the FT^n method, which is particularly well-suited for analyzing the deterministic parts of sounds. Our method has been successfully tested on both synthetic and natural sounds with a low noise level, and the same precision improvements as in section 4 happen when such a replacement is done.
For example it is well-known that sounds with vibrato are hard to analyze with the classic short-time Fourier transform. In order to analyze the voice with deep vibrato of a soprano singer, the FT^n method requires an analysis window 8 times smaller than the classic Fourier method does and a great quality improvement is achieved, as shown in *Figure 3*.
Of course the FT^n analysis succeeds with classic instruments like guitars, pianos, trumpets, etc. Samples are available on the Internet [9]. On most of high-pitched sounds (more than 180 Hz), excellent results have been achieved with very small analysis windows, down to 256 points with R = 44100 Hz, i.e.

less than 6 ms analysis time.

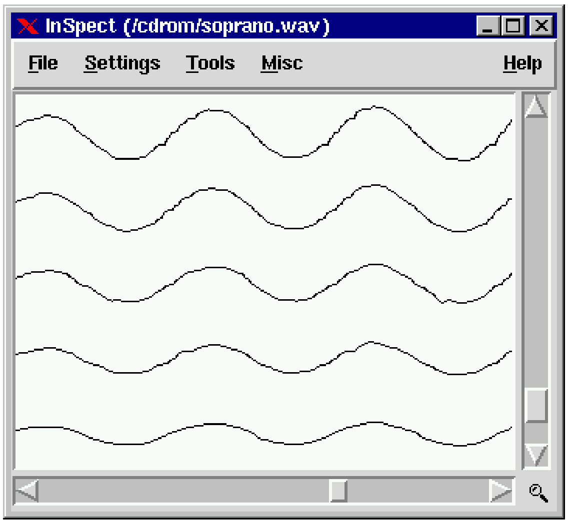

Figure 3: Snapshot of the InSpect program analyzing a soprano voice with deep vibrato. The analysis window is 512-point Hann and R = 44100 Hz. The curves represent the partial frequencies as functions of time (only 0.3 seconds are visible on the picture).

Originally designed for sounds with slow time-varying partials, our method has turned out to allow precise analysis of instruments with quite fast evolutions (even for the attack phase). This is indeed possible because small analysis windows are sufficient for high-pitched sounds.

6 Conclusions

In this presentation FT^n - n-th order (short-time) Fourier Transform - has been introduced. This method is an enhancement of the standard short-time Fourier transform, providing greater accuracy for both frequency and amplitude with small analysis windows, thus permitting greater time resolution. From the complexity point of view, this method is very interesting, since it requires the computation of two small discrete Fourier transforms instead of one much larger.

This method can be practically used during the analysis phase of spectral modeling synthesis, instead of a classic phase vocoder. Making a comparison between the FT^n method and the zero-padding plus interpolation method described in [10] and used in SMS should be an interesting research topic.

The main interest of this analysis method, providing precise spectral modeling parameters, is to allow ever deeper musical transformations on sounds by minimizing deformations due to analysis artifacts. The next step is to structure the sound model in such a way that musical operations can be simply expressed.

7 Acknowledgments

This research was done at the SCRIME (*Studio de Création et de Recherche en Informatique et Musique Electroacoustique*) and was supported by the *Conseil Régional d'Aquitaine*, the *Ministère de la Culture*, the *Direction Régionale des Actions Culturelles d' Aquitaine*, and the *Conseil Général de la Gironde*.

References

[1] Xavier Serra and Julius O. Smith. 1990. Spectral Modeling Synthesis: A Sound Analysis / Synthesis System Based on a Deterministic plus Stochastic Decomposition. *Computer Music Journal*, Volume 14, Number 4, pp. 12-24.

[2] M. Desainte-Catherine and S. Marchand. 1998. High Precision Fourier Analysis of Sounds using Signal Derivatives. LaBRI Research Report Number 120498, Online. <http://www.labri.u-bordeaux.fr/Publications/>

[3] Fredric J. Harris. 1978. On the use of windows for harmonic analysis with the discrete Fourier transform. In *Proceedings IEEE*, Volume 66, pp. 51-83.

[4] Robert J. McAulay and Thomas F. Quatieri. 1986. Speech Analysis / Synthesis Based on a Sinusoidal Representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*. Volume 34, Number 4, pp. 744-754.

[5] Kelly Fitz and Lippold Haken. 1996. Sinusoidal Modeling and Manipulation Using Lemur. *Computer Music Journal*, Volume 20, Number 4, pp. 44-59.

[6] James A. Moorer. 1978. The Use of the Phase Vocoder in Computer Music Applications. *Journal of the Audio Engineering Society*. Volume 26, Number 1/2, pp. 42-45.

[7] Mark B. Dolson. 1986. The Phase Vocoder: A Tutorial. *Computer Music Journal*. Volume 10, Number 4, pp. 14-27.

[8] Marie-Hélène Serra. 1997. « Introducing the Phase Vocoder » in C. Roads et al. Editors, *Musical Signal Processing*, Swets & Zeitlinger Publishers, pp. 31-90.

[9] Sylvain Marchand. "InSpect Software Package" <http://www.scrime.u-bordeaux.fr/InSpect.html>

[10] Xavier Serra. 1989. A System for Sound Analysis / Transformation / Synthesis Based on a Deterministic plus Stochastic Decomposition. PhD Thesis, CCRMA, Stanford University.