Proceedings of the 2nd COST G-6 Workshop on Digital Audio Effects (DAFx99), NTNU, Trondheim, December 9-11, 1999

A Hierarchical Constant Q Transform for Partial Tracking in Musical Signals

Ozgur Izmirli

Center for Arts and Technology
Box 5318, Connecticut College
New London, CT 06320, USA
oizm@conncoll.edu

ABSTRACT

This paper addresses a method for signal dependent time-frequency tiling of musical signals with respect to onsets and offsets of partials. The method is based on multi-level constant Q transforms where the calculation of bins in the higher levels of the transforms depend on the input signal content. The transform utilizes the signal energy in the subbands to determine whether the higher Q bins in the next level, that correspond roughly to the same frequency band in that level, will be calculated or not. At each higher level, the frequency resolution is increased by doubling the number of bins only for which there is significant energy in the previous level. The Q is adjusted accordingly for each level and is held constant within a level. Processing starts with a low Q that provides good time resolution and proceeds with higher levels until the desired maximum frequency resolution is achieved. The advantages of this method are two-fold: First, the time resolution depends on the spacing of the frequency components in the input signal, potentially leading to reduced time smearing, and second, although signal dependent, the conditional calculation of higher Q levels of the transform has a direct consequence of reducing the number of operations in calculating the final spectrum for regular harmonic monophonic sounds. Partial tracking is performed using conventional peak picking and a birth-death strategy of frequency tracks. Testing is being carried out by resynthesizing the input sound from the extracted parameters using a sum of sinusoids with cubic interpolation for phase unwrapping between frames.

1. INTRODUCTION

Audio compression, pitch or time modification, timbre classification/recognition and automatic transcription are some of the areas that constitute the motivation for musical signal analysis. The characteristics pertaining to the class of signals under scrutiny dictate the method of analysis to be performed. With resynthesis as an aim, effective methods are known that extract various components from the input signal and regenerate the sound from these components. For musical sounds that have definite pitch, a trivial attempt is to extract the sinusoidal parameters from the input signal. For example, the sound is modeled as a sum of sinusoids in Sinusoidal Modeling Synthesis (SMS) [1] and the information that is not captured by the sinusoids is conveyed by the residual part. An account of sinusoidal modeling and a comparison with other methods has been given in [2].

The possibility of incorporation of a constant Q approach into the sinusoidal modeling context is discussed in this paper. Although the constant Q transform is not as computationally efficient as the Fast Fourier Transform it has certain characteristics that are suitable for musical signals. It provides higher frequency resolution in the lower frequency portion and lower resolution at higher frequencies.

As with most fixed window size methods instead of implementing a fixed compromise between the frequency and time resolutions, it is possible to obtain a signal dependent resolution determination scheme that has a hierarchical nature. The hierarchical constant Q transform presented below virtually consists of layers of constant Q transforms with progressively increasing levels of Q and bin density at each level.

2. THE TRANSFORM

The constant Q transform used in this implementation is based on [3]. One of the things different from the original formulation is that here all bins of the transform have been aligned to the center time sample. The output of the transform in level d, bin k_d, is represented by $X_d[k_d,y]$. All analysis windows are centered around time sample y which corresponds to the frame time of the transform.

The processing for a fixed value of y starts with a low Q (level 0) that provides good time resolution and proceeds with higher levels until the predetermined maximum frequency resolution is achieved at level d_{max}.

The calculation of each higher level is conditional to the signal energy content in the corresponding bins in the level below it. As the number of bins is doubled at each level, two bins are calculated in level $d+1$ for one bin in level d. This condition can be written as:

$$\begin{aligned} &\textit{if} \quad X_d[k_d, y] > s \\ &\quad \textit{evaluate } X_{d+1}[2k_d, y] \textit{ and } X_{d+1}[2k_d + 1, y] \end{aligned}$$

in which the k indices in each level range from 0 to $\max(k_d)$ and $\max(k_{d+1}) = 2\max(k_d)+1$.

Calculation of the complete input signal is carried out by hopping the transform a fixed number of time samples and repeating the above procedure.

The amplitude, frequency and phase for each peak in the spectrum are then calculated. Currently the precise frequency is determined by looking at the phase changes of two transforms that are one-sample apart [4]. Due to the reason that all bins are centered around a single reference time sample and that the analysis windows have different lengths, the directly calculated phases lack a common reference. Accordingly, the phases for all bins are adjusted to reflect a value with respect to the center sample, enabling the method of frame based cubic interpolation for phase unwrapping to be applied [5]. The peak trajectories are determined using the conventional birth-death strategy of frequency tracks [1][5].

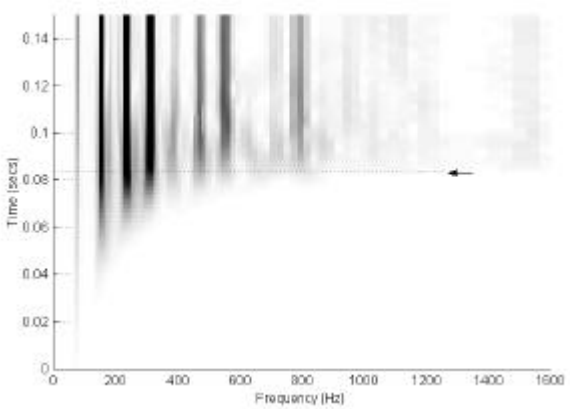

Figure 1. *The output of a single constant Q transform of a piano C3 sound.*

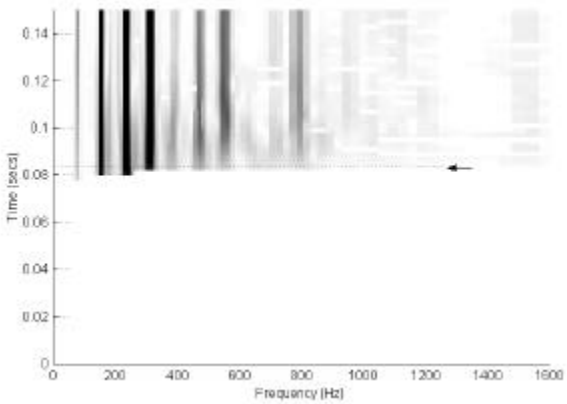

Figure 2. *The output of the hierarchical constant Q transform of a piano C3 sound.*

3. RESULTS

The constant Q transform as defined by Brown [3] can have high resolution at low frequencies that may help resolve closely spaced frequency components. Higher resolution inevitably implies longer filter lengths. The drawback of this feature, however, is that the time smearing of especially the lower frequency components becomes significant. This in turn, results in slow and premature buildup of those components in the resynthesized signal leading to pre-echo. The hierarchical constant Q transform presented here uses bandlimited energy information in frequency bands that become progressively smaller in order to obtain the best time resolution possible for that signal. This difference is demonstrated in figures 1 and 2. Figure 1 shows the spectrum of a single constant Q transform with bin spacing 24 bins per octave. Figure 2 shows the spectrum for the same input using the hierarchical constant Q transform. The time smearing can be seen to have improved in the hierarchical case. The arrows indicate the actual onset time.

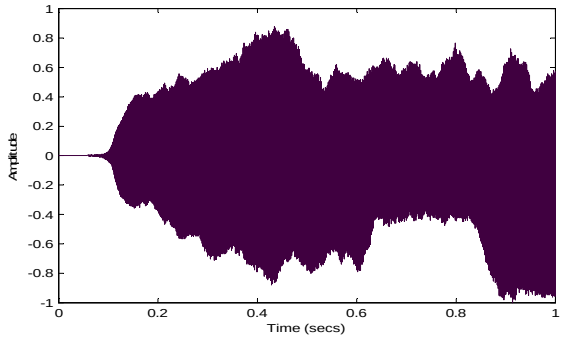

Figure 3. *The input signal : flute C5 sound.*

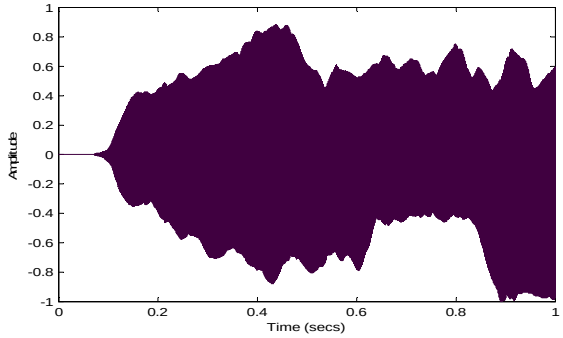

Figure 4. *The resynthesized signal.*

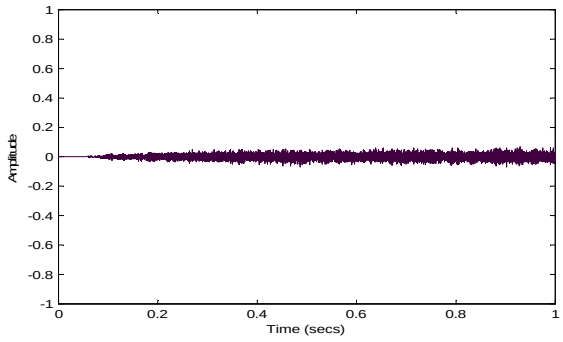

Figure 5. *The residual*

The overall resynthesis is demonstrated using the first second of a flute playing C5. The input waveform is shown in Figure 3. The synthesized signal from the extracted sinusoidal parameters can be seen in Figure 4. The corresponding residual is shown in Figure 5.

4. DISCUSSION

The effect of the conditional calculation on the time required to calculate the transform is more pronounced when the lower bins need not be calculated. Therefore, the frequency of the lowest partial in the input signal plays an important role in determining the time needed for calculation of the transform. The time spent for the calculation of the higher frequency bins are almost insignificant when compared to the lower frequency bins. For example, considering a 24 bin per octave spacing, the analysis window length for a bin at 110 Hz. is 310 mseconds as opposed to one at 1975 Hz. which is 17 mseconds.

The frequency spacing of the input signal plays an important role in the onset and offset time resolution obtained. More specifically, the larger the frequency spacing of a partial the better the time resolution with respect to onset and offset.

The tiling obtained with the hierarchical constant Q transform is conditioned through the energy in the bins in lower levels. That is to say, the amplitude, frequency and phase are all calculated in the highest level. Therefore, the Q is constant for all bins that are used to extract the sinusoidal parameters and is independent from the frequency spacing of the input signal. On account of this, the effect of the tiling is more pronounced with respect to onset and offset of the partial and does not address the adaptation to tracking of fast frequency changes. To track faster frequency changes, the resolution of the transform needs to be decreased.

The selection of the threshold, σ, determines the amount of spectral detail obtained from the transform. If the threshold is chosen to be too low then the hierarchical transform degenerates to a regular transform. On the other hand if it is chosen to be higher than necessary the consequence will be loss of dynamic range and abrupt onset and offset characteristics in the resynthesized signal.

5. CONCLUSIONS

The presented work in this article outlines a method to achieve signal dependent spectral calculation for partial tracking in musical signals. Although, the implementation has been presented using multi-level constant Q transforms, the approach is generic and can be applied to filterbanks that have arbitrary Q and center frequency spacing and also those that have been tailored for signals with known partial trajectories. The hierarchical constant Q transform presented here provides improved time resolution when compared to a single fixed-resolution transform with similar capabilities. The use of the method may also lead to a considerable amount of economy in computation time, which is of course dependent on the periodicity of the input signal and the sparsity of the peaks in its spectrum.

6. REFERENCES

[1] Serra X.. "A System for Sound Analysis/Transformation /Synthesis Based on a Deterministic Plus Stochastic Decomposition," Ph. D. Dissertation, Center for Computer Research in Music and Acoustics, Stanford, California., 1987.

[2] Rodet, X. "Musical Sound Signal Analysis/Synthesis: Sinusoidal+Residual and Elementary Waveform Models" IEEE Time-Frequency and Time-Scale workshop, Coventry, United Kingdom, 1997.

[3] Brown, J. C. "Calculation of a Constant Q Spectral Transform," J. Acoust. Soc. Am. 89, 425-434, 1991.

[4] Brown, J. C., M. S. Puckette "A High Resolution Fundamental Frequency Determination Based on the Phase Changes of the Fourier Transform," J. Acoust. Soc. Am. 94, 662-667, 1993.

[5] McAulay, R. J., T. F. Quatieri "Speech Analysis/Synthesis Based on a Sinusoidal Representation" IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 34, 744-754, 1986.