Audio Engineering Society

Convention Paper

Presented at the 120th Convention
2006 May 20–23 Paris, France

Towards an Inverse Constant Q Transform

Derry FitzGerald[1], Matt Cranitch[1], and Marcin T. Cychowski[1]

[1] *Cork Institute of Technology, Rossa Avenue, Bishopstown, Cork, Ireland*

Correspondence should be addressed to Derry FitzGerald (derry.fitzgerald@cit.ie)

ABSTRACT

The Constant Q transform has found use in the analysis of musical signals due to its logarithmic frequency resolution. Unfortunately, a considerable drawback of the Constant Q transform is that there is no inverse transform. Here we show it is possible to obtain a good quality approximate inverse to the Constant Q transform provided that the signal to be inverted has a sparse representation in the Discrete Fourier Transform domain. This inverse is obtained through the use of $\ell 0$ and $\ell 1$ minimisation approaches to project the signal from the constant Q domain back to the Discrete Fourier Transform domain. Once the signal has been projected back to the Discrete Fourier Transform domain, the signal can be recovered by performing an inverse Discrete Fourier Transform.

1. THE CONSTANT Q TRANSFORM

The Constant Q transform (CQT) was derived by Brown as a means of creating a log-frequency resolution spectrogram [1]. This has considerable advantages for the analysis of musical signals, as the frequency resolution can be set to match that of the equal tempered scale used in western music, where the frequencies are geometrically spaced, as opposed to the linear spacing that occurs in the discrete Fourier transform (DFT). The frequency components of the CQT have a constant ratio of center frequency to resolution, as opposed to the constant frequency difference and constant resolution of the DFT. This constant ratio results in a constant pattern for the spectral components making up notes played on a given instrument, and this has been used to attempt sound source separation of pitched instruments from both single channel and multi-channel mixtures of instruments[2],[3].

Given an inital minimum frequency f_0 for the CQT, the center frequencies for each band can be obtained from:

$$f_k = f_0 2^{\frac{k}{b}} \ (k = 0, 1, ...) \tag{1}$$

where b is the number of bins per octave. The fixed ratio of center frequency to bandwidth is then given by

$$Q = \left(2^{\frac{1}{b}} - 1\right)^{-1} \tag{2}$$

The desired bandwidth of each frequency band is

then obtained by choosing a window of length

$$N_k = Q\frac{f_s}{f_k} \tag{3}$$

where f_s is the sampling frequency. The CQT is defined as

$$X(k) = \frac{1}{N_k}\sum_{n=0}^{N_k-1} W_{N_k}(n)\, x(n)\, exp^{-j2\pi Qn/N_k} \tag{4}$$

where $x(n)$ is the time domain signal and W_{N_k} is a window function, such as the hanning window, of length N_k.

A more efficient implementation of the CQT is described in [4]. Using matrix multiplication the CQT of a column vector x can be defined as:

$$X = Tx \tag{5}$$

where

$$T_{nk} = \begin{cases} \frac{1}{N_k} W_{N_k}(n)\, exp^{-j2\pi Qn/N_k} & n < N_k \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Equivalently the transform can be written in the DFT domain as:

$$X = CY \tag{7}$$

where C is the matrix that results when a DFT is applied to each column of T, and where Y is the DFT of x.The advantage of carrying out the transform in this manner is that the windowed complex exponentials of the temporal transform matrix have DFTs that are close to zero everywhere except in the vicinity of the corresponding frequency in the DFT transform matrix. Therefore setting all elements of C with an absolute value below a low threshold to zero results in a sparse matrix. The computational cost of applying a DFT to the signal followed by a sparse matrix multiplication with C is considerably lower than applying T directly. Further, only the DFT coefficients below the Nyquist frequency from both C and Y need to be used in calculating the CQT. Therefore, once C has been calculated, the tranform can be applied efficiently to subsequent windows of the signal.

Despite the advantages the CQT offers for the analysis of musical signals, a considerable drawback of using the CQT is that there is no inverse for the transform. This is due to the fact that the transform matrix is no longer square, as is the case for the DFT. An approximate inverse has been made available by Brown, which uses a simple extension of the inverse DFT to the log- frequency domain [5]. However, using this approach leads to an inverse which is degraded considerably in comparison to the original signal. It can therefore be seen that a method of obtaining a high quality inverse for the CQT is desirable to further its use in the analysis of musical signals.

2. OBTAINING AN INVERSE CQT

As noted above, the CQT is not invertible due to the fact that the transform matrix is not square. A possible approach to obtaining an inverse CQT is by obtaining the pseudoinverse of the transform matrix. Using this approach does not result in a good quality inverse, and results in a poorer quality inverse to that proposed by Brown. However, in the case of signals containing only pitched instruments, there is additional information about the signal which can be taken advantage of in an effort to obtain a good quality inverse. This is that the DFT of a signal containing only pitched instruments tends to be sparse in nature. This sparsity has been used to attempt sound source separation and automatic music transcription through the use of sparse coding and matrix factorisation approaches [6],[7].

Rather than attempting to obtain an inverse CQT by inverting directly back to the time domain, it was decided to attempt to invert from the CQT domain back to the DFT domain in order to take advantage of signal sparseness. Once the DFT has been obtained, the signal can then be inverted back to the time domain. By separating the real and imaginary parts of the transform, equation 7 can be rewritten as:

$$B = AS \tag{8}$$

where

$$B = \begin{bmatrix} real(X) \\ imag(X) \end{bmatrix} \tag{9}$$

$$A = \begin{bmatrix} real(C) & -imag(C) \\ imag(C) & real(C) \end{bmatrix} \tag{10}$$

and

$$S = \begin{bmatrix} real(Y) \\ imag(Y) \end{bmatrix} \tag{11}$$

Once written in this form, the inversion of the CQT can be posed as the following problem, given B, the CQT of a signal, and A, the transform matrix, which we will now regard as an overcomplete signal dictionary, find S. In other words, given the CQT 'signal' decompose the 'signal' into an optimal superposition of dictionary elements. In this case, we want to ensure a sparse decomposition to reflect the fact that signals containing only pitched instruments typically have sparse representations in the DFT domain. The sparsest possible solution to the set of equations described above can be obtained by solving the following minimisation problem:

$$\text{Minimise} \quad \|S\|_0 \quad \text{subject to} \quad B = AS \tag{12}$$

where $\|S\|_0$ denotes the $\ell 0$ norm of S. This is a highly non-convex optimisation problem and so has proved difficult to solve. However, it has been shown that for many cases the solution to $\ell 0$ optimisation problem is also the solution to the $\ell 1$ optimisation problem:

$$\text{Minimise} \quad \|S\|_1 \quad \text{subject to} \quad B = AS \tag{13}$$

where $\|S\|_1$ denotes the $\ell 1$ norm of S [8]. This problem can be solved using standard linear programming techniques and has been extensively studied under the heading of Basis Pursuit [9] which solves the problem through the use of an interior-point linear programming method. More recently, a Sparse Bayesian approach to solving the $\ell 0$ minimisation problem has been developed by Wipf and Rao [10]. This algorithm is defined by the following iterative update equations.

$$S = \Gamma^{\frac{1}{2}} \left(A\Gamma^{\frac{1}{2}} \right)^{+} B \tag{14}$$

$$\gamma = \text{diag} \left(SS^T + \left[I - \Gamma^{\frac{1}{2}} \left(A\Gamma^{\frac{1}{2}} \right)^{+} A \right] \Gamma \right) \tag{15}$$

where $^+$ denotes the Moore-Penrose pseudoinverse, and where Γ is defined as

$$\Gamma \triangleq \text{diag}\left(\gamma\right) \tag{16}$$

Both of these methods have been investigated as a means of obtaining an inverse CQT.

The problem of inverting a CQT spectrogram then becomes that of solving an $\ell 0$ or $\ell 1$ minimisation problem for each frame of the CQT spectrogram, inverting the recovered DFT of each frame to the time domain, and then performing add-overlap on the frames. As windowing in the CQT is performed inside in the transform matrix C, the DFT signal recovered has had no window applied, and so a hamming window is applied to each time-domain frame before performing the add-overlap. Results obtained using the two methods described above are discussed in the following section.

3. RESULTS

Both the $\ell 1$ and $\ell 0$ minimisation methods for obtaining an inverse CQT were implemented in Matlab. The $\ell 1$ minimisation was carried out using a version of the Basis Pursuit algorithm available at [11], modified to include the CQT dictionary A, while the $\ell 0$ method was implemented based on the algorithm described in [10]. The tests were implemented using a CQT which covered the frequency range 110 Hz to 10000 Hz. Trials were carried out using 12, 24 and 48 bins per octave, corresponding to resolutions of semi, quarter and eighth tones respectively. These resulted in CQTs with 78, 156 and 312 frequency bins respectively.

Fig 1 shows, from top to bottom, the original waveform, the waveform obtained via Brown's inverse CQT method, the waveform obtained via $\ell 0$ inversion, and that obtained via $\ell 1$ inversion. The CQT used had a resolution of 24 bins per octave. Fig 2 shows, again from top to bottom, the DFT of the original waveform, the DFT obtained from Browns's method,the DFT obtained using the $\ell 0$ method, and the DFT obtained using the $\ell 1$ method. It can be seen that the waveforms and DFTs obtained by both the $\ell 0$ and the $\ell 1$ method are quite close to the original waveform and it's DFT, demonstrating the effectiveness of both methods as a means of obtaining an inverse CQT. Fig 3 shows an original waveform of an excerpt of music containing violin accompanied by piano, along with the inverses obtained using Brown's method,the $\ell 0$ method and the $\ell 1$ method. It can be seen that in general both the $\ell 0$ and $\ell 1$ method result in waveforms close to the original, though regions where the $\ell 1$ failed to converge can clearly be seen as bursts of high amplitude noise.

When allowed to run until convergence both the $\ell 0$ and $\ell 1$ methods took similar times to run. As an

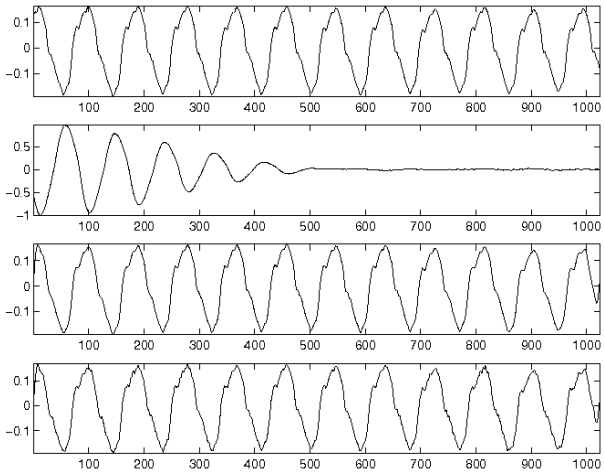

Fig. 1: Original waveform, inverted waveform (Brown's method), inverted waveform ($\ell 0$ method) and inverted waveform ($\ell 1$ method) respectively

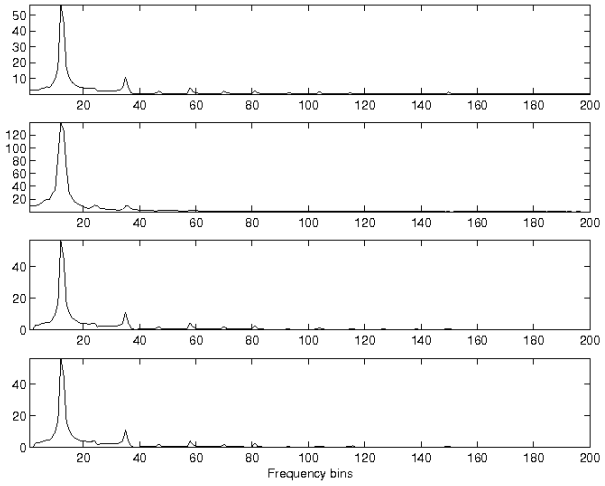

Fig. 2: DFT of Original waveform, DFT from Brown's method, DFT from $\ell 0$ method and DFT from $\ell 1$ method respectively

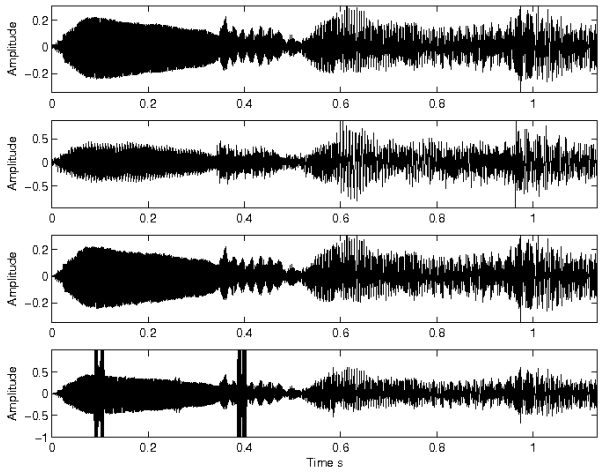

Fig. 3: Original excerpt of violin and piano, inverted excerpt via Brown's method, inverted excerpt via $\ell 0$ method and inverted excerpt via $\ell 1$ method respectively

example, running on a computer using a 2.6 Ghz Pentium IV processor, with 1 GB of RAM, it took approximately an hour to invert the CQT spectrogram of a 4.5 second audio file, using 12 bins per octave, a window size of 1024 samples and a hop-size of 156 samples. It has been observed that the $\ell 0$ minimisation technique converges more reliably than the $\ell 1$ minimisation method, though the $\ell 1$ approach does converge the majority of the time. However, when the $\ell 1$ minimisation approach does converge, it gives slightly better results than the $\ell 0$ method, which suffers from a slight distortion of the original sound. In all cases, the sound quality of the inverse increases with the number of bins per octave, and is always considerably greater than that of the method proposed by Brown.

The principal problem with both methods is the length of time taken to obtain the inverse CQT. Fortunately, it was noted that a single iteration of the $\ell 0$ method resulted in an approximate inverse which was quite close to the true inverse, but had a small amount of extra distortion in the recovered waveform. Using this approximation, the time taken to calculate the inverse in the example given above was reduced to 6.5 minutes, which is a considerable reduction in the time taken to calculate the inverse, though at the expense of a slight increase in distor-

tion of the recovered waveform. This fact, combined with it's greater stability means that the $\ell 0$ method is prefered as means of obtaining an inverse CQT.

A further problem with both methods is that they are only valid for signals containing only pitched instruments, as both methods required that the signal has a sparse representation in the DFT domain. Therefore, signals containing broadband noise such as drum sounds will not be inverted correctly. Tests on signals containing both pitched instruments and drums have confirmed this, resulting in a very distorted inverse, though the pitched instruments can still be heard.

4. CONCLUSIONS

A means of obtaining an inverse CQT has been presented and demonstrated. The approach used take advantage of the fact that signals containing only pitched instruments have sparse representations in the DFT, and obtains an inverse by reformulating the problem of obtaining an inverse CQT as that of decomposing a signal in the CQT domain into a sparse representation in the DFT domain. This problem has been attempted using both $\ell 0$ and $\ell 1$ minimisation approaches, and while both methods work well, the $\ell 0$ method has better convergence and can be inverted faster than the $\ell 1$ method. This research demonstrates that it is possible to obtain a high quality inverse CQT, provided that the signal to be inverted has a sparse representation in the DFT domain.

5. ACKNOWLEDGEMENTS

This research has been supported by funding from the Irish Research Council for Science, Engineering and Technology.

6. REFERENCES

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

[2] FitzGerald, D., Cranitch, M., and Coyle, E., "Shifted Non-negative Matrix Factorisation for Sound Source Separation", Proceedings of the IEEE conference on Statistics in Signal Processing, Bordeaux, France, July 2005.

[3] FitzGerald, D., Cranitch, M., and Coyle, E., "Sound Source Separation using shifted Non-negative Tensor Factorisation", IEEE International Conference on Acoustics, Speech and Signal Processing, 2006 (ICASSP2006), Toulouse France.

[4] Brown, J.C. and Puckette, M.S. (1992). "An Efficient Algorithm for the Calculation of a Constant Q Transform", J. Acoust. Soc. Am. 92, 2698-2701. (1992)

[5] http://web.media.mit.edu/ brown/cqtrans.htm

[6] Smaragdis, P.; Brown, J.C., "Non-negative Matrix Factorization for Polyphonic Music Transcription", IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 177-180, October 2003

[7] Abdallah, S.A. and M.D. Plumbley, "Polyphonic transcription by non-negative sparse coding of power spectra", in Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004), Barcelona, Spain, October 10-14, 2004.

[8] Donoho, D. L. and Elad, M.,"Maximal sparsity Representation via l1 Minimization", the Proc. Nat. Aca. Sci., Vol. 100, pp. 2197-2202, March 2003.

[9] Chen, S. S., Donoho, D. L., and Saunders, M. A., "Atomic Decomposition by Basis Pursuit", SIAM Journal on Scientific Computing, vol. 20-1, pp 33–61, 1999,

[10] Wipf, D., and Rao, B."$\ell 0$ Minimization for Basis Selection",Advances in Neural Information Processing Systems, vol. 17, MIT Press, 2005.

[11] http://www-stat.stanford.edu/∼atomizer/