Compression in the Transform Domain

Henry D. Pfister

October 16, 2015

1 Lossy Compression

When compressing audio and images, it is common for systems to tolerate small (perceptual) errors in the reconstructed signal. This is known as lossy compression. The question of how to quantify perceptual error in a problem is quite challenging and important. However, this will addressed only tangentially. For now, assume that error between the signal $x[n]$ and its reconstruction $\hat{x}[n]$ is measured using squared Euclidean distance

$$d_E^2 = \sum_{i=0}^{N-1} \left(x[n] - \hat{x}[n]\right)^2 .$$

Our study of quantization has shown that uniform quantization with step-size Δ results in an approximate quantization energy of $\Delta^2/12$ per dimension. Assume that the average signal power is $1/3$ per dimension and the signal range is $[-1, 1]$. In this case, then m bit quantization results in $\Delta_0 = 2^{-m+1}$. Thus, quantizing all coefficients to m bits requires Nm bits and results in $d_E^2 \approx N\Delta_0^2/12$.

Now, suppose that there is an orthogonal transformation that rotates the entire signal energy into a single dimension. Since the transform is orthogonal, that coefficient must have average power $N/3$ and its range must be roughly $[-\sqrt{N}, \sqrt{N}]$. Using m bits spread over this larger range, we get a step-size of $\Delta_1 = 2^{-m+1+\log_2 \sqrt{N}}$. The quantization noise power in this case is $\Delta_1^2/12 = 2^{-2m+2+2\log_2 \sqrt{N}} = N\Delta_0^2/12$. Since the transform is orthogonal, the squared distance is preserved by the inverse transform. Thus, we have achieved roughly the same quantization noise power using m bits instead of Nm bits.

Of course, the example was designed to illustrate a point. Still, if the transform is chosen well, then the number coefficients with large magnitude is often much smaller than N. Perceptual models can also allow one to ignore a larger set of coefficients (e.g., with moderate magnitude) whose perceptual impact is small. This can lead high-quality reconstruction with large compression ratios (e.g., as achieved by JPEG, MPEG, and MP3 audio).

2 The Discrete Cosine Transform (DCT)

The DCT is an orthogonal transform that is closely related to the discrete Fourier transform (DFT). The DCT is typically preferred over the DFT for transform-domain compression of signals because it typically compacts the energy in the signal onto a smaller number of transform-domain coefficients. There are four standard DCT flavors denoted type-I through type-IV. The simplest form of the DCT is type-I and the transform is defined by

$$X[k] = \frac{1}{2}\left(x[0] + (-1)^k x[N-1]\right) + \sum_{n=1}^{N-2} x[n] \cos\left(\frac{\pi k n}{N-1}\right)$$

$$x[n] = \frac{1}{N-1}\left(x[0] + (-1)^n x[N-1]\right) + \frac{2}{N-1} \sum_{k=1}^{N-2} X[k] \cos\left(\frac{\pi k n}{N-1}\right).$$

We observe here that $\cos\left(\frac{\pi(2N-2-k)n}{N-1}\right) = \cos\left(-\frac{\pi k n}{N-1}\right) = \cos\left(\frac{\pi k n}{N-1}\right)$ implies that $X[2N-2-k] = X[k]$.

2.1 Connection to the Discrete Fourier Transform

It turns out that the length-N type-I DCT of $x[n]$ is essentially equal to the DFT of the length-$2N-2$ sequence

$$\begin{aligned} y[n] &= \Big\{\underline{x[0]}, x[1], \ldots, x[N-2], x[N-1], x[N-2], x[N-3], \ldots, x[1]\Big\} \\ &= \begin{cases} x[n] & \text{if } 0 \le n \le N-2 \\ x[2N-2-n] & \text{if } N-1 \le n \le 2N-3. \end{cases} \end{aligned}$$

To see the connection with the DFT, we observe that

$$\begin{aligned} X[k] &= \frac{1}{2}\left(x[0] + (-1)^k x[N-1]\right) + \sum_{n=1}^{N-2} x[n] \cos\left(\frac{\pi k n}{N-1}\right) \\ &= \frac{1}{2}\left(x[0] + (-1)^k x[N-1]\right) + \frac{1}{2}\sum_{n=1}^{N-2} x[n]\left(e^{-j\pi kn/(N-1)} + e^{j\pi kn/(N-1)}\right) \\ &= \frac{1}{2}\sum_{n=0}^{N-2} x[n] e^{-j\pi kn/(N-1)} + \frac{1}{2}\sum_{n=1}^{N-1} x[n] e^{j\pi kn/(N-1)} \\ &= \frac{1}{2}\sum_{n=0}^{N-2} x[n] e^{-j\pi kn/(N-1)} + \frac{1}{2}\sum_{m=N-1}^{2N-3} x[2N-2-m] e^{j\pi k(2N-2-m)/(N-1)} \\ &= \frac{1}{2}\sum_{n=0}^{N-2} x[n] e^{-2\pi jkn/(2N-2)} + \frac{1}{2}\sum_{m=N-1}^{2N-3} x[2N-2-m] e^{-2\pi jkm/(2N-2)} \\ &= \frac{1}{2}\sum_{n=0}^{2N-1} y[n] e^{-2\pi jkn/(2N-2)}, \end{aligned}$$

where $k = 0, 1, \ldots, N-1$ runs only over the first N indices.

The energy compaction of the DFT is suboptimal because the length-N DFT is closely related to the DTFT of the periodic extension of $x[n]$ with period N. The reason is that the block boundary introduces an artificial discontinuity in the periodic extension at times $n = pN$ when the signal jumps from $x[N-1]$ to $x[0]$. In particular, the value $|x[0] - x[N-1]|$ can be large even if the rest of the signal is quite smooth (e.g., first order differences are small). From Fourier transform theory, we know that jumps in the signal introduce high frequency components.

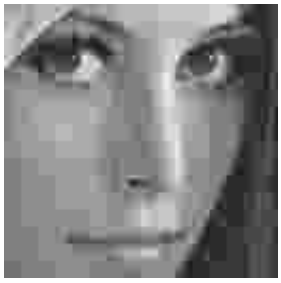
The improved energy compaction of the DCT occurs because the length-N DCT is closely related to the length-$2N-2$ DFT of the signal $y[n]$ and the periodic extension of $y[n]$ with period $2N-2$ does not contain an artificial discontinuity. In particular, the absolute first-order differences of the periodic extension lie in the set $\{|x[n+1] - x[n]| \mid n = 0, 1, \ldots, N-2\}$.

To compute the inverse DCT, we note that the length-$2N-2$ DFT inverse formula implies so that, for $n = 0, 1, \ldots, N-1$, we have

$$\begin{aligned}
x[n] &= y[n] \\
&= \frac{2}{2N-2} \sum_{k=0}^{2N-1} X[k] e^{2\pi jkn/(2N-2)} \\
&= \frac{1}{N-1} \sum_{k=0}^{N-2} X[k] e^{\pi jkn/(N-1)} + \frac{1}{N-1} \sum_{k=N-1}^{2N-3} X[k] e^{\pi jkn/(N-1)} \\
&= \frac{1}{N-1} \sum_{k=0}^{N-2} X[k] e^{\pi jkn/(N-1)} + \frac{1}{N-1} \sum_{k=N-1}^{2N-3} X[2N-2-k] e^{\pi jkn/(N-1)} \\
&= \frac{1}{N-1} \sum_{k=0}^{N-2} X[k] e^{\pi jkn/(N-1)} + \frac{1}{N-1} \sum_{k'=1}^{N-1} X[k'] e^{\pi j(2N-2-k')n/(N-1)} \\
&= \frac{1}{N-1} \sum_{k=0}^{N-2} X[k] e^{\pi jkn/(N-1)} + \frac{1}{N-1} \sum_{k'=1}^{N-1} X[k'] e^{-\pi jk'n/(N-1)} \\
&= \frac{1}{N-1} \left(X[0] + (-1)^n X[N-1] \right) + \frac{2}{N-1} \sum_{k=1}^{N-2} X[k] \cos\left(\frac{\pi kn}{N-1} \right).
\end{aligned}$$

3 The Modified Discrete Cosine Transform (MDCT)

In transform-domain compression of both audio and video, some of the most noticeable artifacts are due to blocking effects. For example, low bit-rate JPEG pictures and MPEG movies show noticeable artifacts at the boundaries of the 8×8 DCT blocks. For example, the blocking is quite noticeable in this piece of a JPEG image:



In audio, this artifact is often mitigated by overlapping blocks and applying a window function before computing the DCT. The problem with naively overlapping blocks is that twice as many transform coefficients are generated. This makes the compression problem twice as hard! The MDCT is clever algorithm that operates on length-$2N$ blocks of samples that overlap each other by N samples. The key property is that each new window only generates N new coefficients and still achieves perfect reconstruction. It also allows the blocks to be windowed. These properties make it ideally suited for audio compression.

The MDCT transform and its inverse are given by

$$\begin{aligned}
X[k] &= \sum_{n=0}^{2N-1} x[n] \cos\left(\frac{\pi}{N} \left(n + \frac{1}{2} + \frac{N}{2} \right) \left(k + \frac{1}{2} \right) \right) \\
x[n] &= \frac{1}{2N} \sum_{k=0}^{N-1} X[k] \cos\left(\frac{\pi}{N} \left(n + \frac{1}{2} + \frac{N}{2} \right) \left(k + \frac{1}{2} \right) \right).
\end{aligned}$$

In applications, length-$2N$ blocks are overlapped by N and transformed to get a set of coefficients $X[0], \ldots, X[N-1]$ for each new block of N inputs. For the inverse transform, length-$2N$ blocks are generated by the $x[n]$ equations and then they are added together in overlapping sections. In practice, a window function is typically used. This will be discussed later.

If $x[n]$ is real, then it is clear that the MDCT coefficients are also real. In addition, the extended MDCT coefficients (i.e., $k = N, N+1, \ldots, N$) satisfy the symmetry condition $X[2N-1-k] = (-1)^{N+1}X[k]$. This can be seen from identity

$$\begin{aligned}\cos\left(\frac{\pi}{N}\left(n+\frac{1}{2}+\frac{N}{2}\right)\left(2N-1-k+\frac{1}{2}\right)\right) &= \cos\left(2\pi\left(n+\frac{1}{2}+\frac{N}{2}\right)-\frac{\pi}{N}\left(n+\frac{1}{2}+\frac{N}{2}\right)\left(k+\frac{1}{2}\right)\right)\\ &= (-1)^{N+1}\cos\left(\frac{\pi}{N}\left(n+\frac{1}{2}+\frac{N}{2}\right)\left(k+\frac{1}{2}\right)\right).\end{aligned}$$

3.1 Connection to the Shifted Discrete Fourier Transform

The MDCT can be understood in terms of the shifted DFT [1]. The (a,b)-shifted DFT is defined by

$$\begin{aligned}X[k] &= \sum_{n=0}^{N-1} x[n]e^{-2\pi j(k+b)(n+a)/N}\\ x[n] &= \frac{1}{N}\sum_{k=0}^{N-1} X[k]e^{2\pi j(k+b)(n+a)/N}.\end{aligned}$$

To see that these transforms are orthogonal for all (a,b), we compute the inner product between any two basis vectors with

$$\begin{aligned}\chi_{k,k'} &= \sum_{n=0}^{N-1} e^{-2\pi j(k+b)(n+a)/N}e^{2\pi j(k'+b)(n+a)/N}\\ &= \sum_{n=0}^{N-1} e^{-2\pi j(k'-k)(n+a)/N}\\ &= e^{-2\pi j(k'-k)a/N}\sum_{n=0}^{N-1} e^{-2\pi j(k'-k)n/N}\\ &= N\delta[k-k'].\end{aligned}$$

The MDCT can be written in terms of the $(\frac{N+1}{2}, \frac{1}{2})$-shifted DFT with

$$\begin{aligned}
X[k] &= \sum_{n=0}^{2N-1} x[n] \cos\left(\frac{\pi}{N}\left(n+\frac{1}{2}+\frac{N}{2}\right)\left(k+\frac{1}{2}\right)\right) \\
&= \frac{1}{2}\sum_{n=0}^{2N-1} x[n]\left(e^{-\frac{j\pi}{N}\left(n+\frac{1}{2}+\frac{N}{2}\right)\left(k+\frac{1}{2}\right)} + e^{\frac{j\pi}{N}\left(n+\frac{1}{2}+\frac{N}{2}\right)\left(k+\frac{1}{2}\right)}\right) \\
&= \frac{1}{2}\sum_{n=0}^{N-1} x[n]e^{-\frac{j\pi}{N}\left(n+\frac{N+1}{2}\right)\left(k+\frac{1}{2}\right)} + \frac{1}{2}\sum_{n=0}^{N-1} x[n]e^{\frac{j\pi}{N}\left(n+\frac{N+1}{2}\right)\left(k+\frac{1}{2}\right)} \\
&\quad + \frac{1}{2}\sum_{n=N}^{2N-1} x[n]e^{-\frac{j\pi}{N}\left(n+\frac{N+1}{2}\right)\left(k+\frac{1}{2}\right)} + \frac{1}{2}\sum_{n=N}^{2N-1} x[n]e^{\frac{j\pi}{N}\left(n+\frac{N+1}{2}\right)\left(k+\frac{1}{2}\right)} \\
&= \frac{1}{2}\sum_{n=0}^{N-1} x[n]e^{-\frac{j\pi}{N}\left(n+\frac{N+1}{2}\right)\left(k+\frac{1}{2}\right)} + \frac{1}{2}\sum_{n=0}^{N-1} x[N-1-n]e^{\frac{j\pi}{N}\left(N-1-n+\frac{N+1}{2}\right)\left(k+\frac{1}{2}\right)} \\
&\quad + \frac{1}{2}\sum_{n=N}^{2N-1} x[n]e^{-\frac{j\pi}{N}\left(n+\frac{N+1}{2}\right)\left(k+\frac{1}{2}\right)} + \frac{1}{2}\sum_{n=N}^{2N-1} x[3N-n-1]e^{\frac{j\pi}{N}\left(3N-1-n+\frac{N+1}{2}\right)\left(k+\frac{1}{2}\right)} \\
&= \frac{1}{2}\sum_{n=0}^{N-1} x[n]e^{-\frac{j\pi}{N}\left(n+\frac{N+1}{2}\right)\left(k+\frac{1}{2}\right)} + \frac{1}{2}\sum_{n=0}^{N-1} x[N-1-n]\underbrace{e^{2\pi j\left(k+\frac{1}{2}\right)}}_{-1}e^{-\frac{j\pi}{N}\left(n+\frac{N+1}{2}\right)\left(k+\frac{1}{2}\right)} \\
&\quad + \frac{1}{2}\sum_{n=N}^{2N-1} x[n]e^{-\frac{j\pi}{N}\left(n+\frac{N+1}{2}\right)\left(k+\frac{1}{2}\right)} + \frac{1}{2}\sum_{n=N}^{2N-1} x[3N-n-1]\underbrace{e^{4\pi j\left(k+\frac{1}{2}\right)}}_{1}e^{-\frac{j\pi}{N}\left(n+\frac{N+1}{2}\right)\left(k+\frac{1}{2}\right)} \\
&= \sum_{n=0}^{2N-1} y[n]e^{-\frac{2\pi j}{2N}\left(n+\frac{N+1}{2}\right)\left(k+\frac{1}{2}\right)},
\end{aligned}$$

where the y-sequence associated with the input $x[n]$ is given by

$$y[n] = \begin{cases} \frac{1}{2}\left(x[n] - x[N-1-n]\right) & \text{if } 0 \le n \le N-1 \\ \frac{1}{2}\left(x[n] + x[3N-1-n]\right) & \text{if } N \le n \le 2N-1. \end{cases}$$

Thus, the MDCT $X[k]$ equals the first N coefficients of the $(\frac{N+1}{2}, \frac{1}{2})$-shifted length-$2N$ DFT of the signal $y[n]$. Let $Y[k]$ the $(\frac{N+1}{2}, \frac{1}{2})$-shifted length-$2N$ DFT of $y[n]$. For the inverse MDCT, it follows from the inverse shifted DFT that y-sequence satisfies

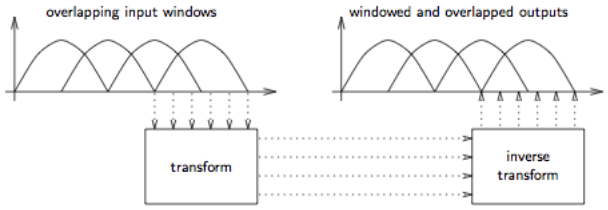
$$\begin{aligned}
y[n] &= \frac{1}{2N}\sum_{k=0}^{2N-1} X[k]e^{\frac{2\pi j}{2N}\left(n+\frac{N+1}{2}\right)\left(k+\frac{1}{2}\right)} \\
&= \frac{1}{N}\sum_{k=0}^{N-1} X[k]e^{\frac{2\pi j}{2N}\left(n+\frac{N+1}{2}\right)\left(k+\frac{1}{2}\right)} + \frac{1}{N}\sum_{k=N}^{2N-1} X[k]e^{\frac{2\pi j}{2N}\left(n+\frac{N+1}{2}\right)\left(k+\frac{1}{2}\right)} \\
&= \frac{1}{N}\sum_{k=0}^{N-1} X[k]e^{\frac{2\pi j}{2N}\left(n+\frac{N+1}{2}\right)\left(k+\frac{1}{2}\right)} + \frac{1}{N}\sum_{k=0}^{N-1} X[2N-1-k]e^{\frac{2\pi j}{2N}\left(n+\frac{N+1}{2}\right)\left(2N-1-k+\frac{1}{2}\right)} \\
&= \frac{1}{N}\sum_{k=0}^{N-1} X[k]e^{\frac{2\pi j}{2N}\left(n+\frac{N+1}{2}\right)\left(k+\frac{1}{2}\right)} + (-1)^{N+1}\frac{1}{N}\sum_{k=0}^{N-1} X[k]\underbrace{e^{2\pi j\left(n+\frac{N+1}{2}\right)}}_{(-1)^{N+1}}e^{-\frac{2\pi j}{2N}\left(n+\frac{N+1}{2}\right)\left(k+\frac{1}{2}\right)} \\
&= \frac{2}{N}\sum_{k=0}^{N-1} X[k]\cos\left(\frac{\pi}{N}\left(n+\frac{1}{2}+\frac{N}{2}\right)\left(k+\frac{1}{2}\right)\right).
\end{aligned}$$

Thus, the inverse MDCT of $X[k]$ computes the sequence $y[n]$ for $n = 0, 1, \ldots, 2N-1$.

It is important to note that the inverse MDCT of one block does not return the original signal $x[n]$ but instead the distorted signal $y[n]$. The mixing of the $x[n]$ coefficients in $y[n]$ is the result of something called time-domain aliasing. The MDCT experiences time-domain aliasing because it essentially applies a length-N type-IV DCT to an input block of length-$2N$ simply by extending the length of the sum. But, the length-N type-IV DCT is not orthogonal for inputs of length $2N$. Moreover, only N coefficients are generated so the inverse transform cannot possibly recover the original signal without distortion.

3.2 Overlap and Add Reconstruction

The final step is the inverse transform is the overlap and add process. The amazing part of this process is that the overlapp-add process causes the time-domain aliasing in one block to cancel time-domain aliasing in the next block. This phenomenon is known as time-domain aliasing cancellation. For completeness, we consider the entire process with windowing. Let $w[n]$ be a length-$2N$ window function satisfying $w[n] = w[2N-1-n]$ and $w[n]^2 + w[n+N]^2 = 2$.



For an input sequence $x[n]$, let $x_m[n] = w[n]x[n+mN]$ denote the windowed m-th input block of length $2N$. Likewise, the y-sequence associated with the windowed m-th input block is given

$$\begin{aligned} y_m[n] &= \begin{cases} \frac{1}{2}\left(w[n]x_m[n] - w[N-1-n]x_m[N-1-n]\right) & \text{if } 0 \le n \le N-1 \\ \frac{1}{2}\left(w[n]x_m[n] + w[3N-1-n]x_m[3N-1-n]\right) & \text{if } N \le n \le 2N-1 \end{cases} \\ &= \begin{cases} \frac{1}{2}\left(w[n]x[n+mN] - w[N-1-n]x[(m+1)N-1-n]\right) & \text{if } 0 \le n \le N-1 \\ \frac{1}{2}\left(w[n]x[n+mN] + w[3N-1-n]x[(m+3)N-1-n]\right) & \text{if } N \le n \le 2N-1. \end{cases} \end{aligned}$$

The MDCT of the m-th windowed input block is given by

$$X_m[k] = \sum_{n=0}^{2N-1} x_m[n] \cos\left(\frac{\pi}{N}\left(n + \frac{1}{2} + \frac{N}{2}\right)\left(k + \frac{1}{2}\right)\right).$$

As we saw in the last section, the inverse MDCT of $X_m[k]$ results in the sequence

$$y_m[n] = \sum_{k=0}^{N-1} X_m[k] \cos\left(\frac{\pi}{N}\left(n + \frac{1}{2} + \frac{N}{2}\right)\left(k + \frac{1}{2}\right)\right).$$

The window, overlap, and add reconstruction process for samples $n = 0, 1, \ldots, N-1$ of the m-th block

produces the output

$$\begin{aligned}
\hat{x}_m[n] &= w[n]y_m[n] + w[n+N]y_{m-1}[n+N] \\
&= \frac{w[n]}{2}\left(w[n]x[n+mN] - w[N-1-n]x[(m+1)N-1-n]\right) \\
&\quad + \frac{w[n+N]}{2}\left(w[n+N]x[n+mN] + w[2N-1-n]x[(m+1)N-1-n]\right) \\
&= \frac{1}{2}\underbrace{\left(w^2[n] + w^2[n+N]\right)}_{2}x[n+mN] \\
&\quad - \frac{1}{2}\underbrace{\left(w[n+N]w[2N-1-n] - w[n]w[N-1-n]\right)}_{0}x[(m+1)N-1-n] \\
&= x[n+mN] = x_m[n],
\end{aligned}$$

where $w[n+N]w[2N-1-n] = w[n]w[N-1-n]$ because the symmetry of the window implies that $w[2N-1-n] = w[n]$ and $w[n+N] = w[N-1-n]$. Thus, the overall process achieves perfect reconstruction.

References

[1] Y. Wang, L. Yaroslavsky, and M. Vilermo, "On the relationship between MDCT, SDPT and DFT," in *Proc. Intl. Conf. on Signal Processing (ICSP)*, vol. 1, pp. 44–47, IEEE, 2000.