For a finite number of cepstral coefficients the bilinear transform in Figure 6.27 results in an infinite number of warped cepstral coefficients. Since truncation is usually done in practice, the bilinear transform is equivalent to a matrix multiplication, where the matrix is a function of the warping parameter α. Shikano [43] showed these warped cepstral coefficients were beneficial for speech recognition.

6.5.2. Mel-Frequency Cepstrum

The *Mel-Frequency Cepstrum Coefficients* (MFCC) is a representation defined as the real cepstrum of a windowed short-time signal derived from the FFT of that signal. The difference from the real cepstrum is that a nonlinear frequency scale is used, which approximates the behavior of the auditory system. Davis and Mermelstein [8] showed the MFCC representation to be beneficial for speech recognition.

Given the DFT of the input signal

$$X_a[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi nk/N} \quad , \qquad 0 \le k < N \tag{6.139}$$

we define a filterbank with M filters ($m = 1, 2, \cdots, M$), where filter m is triangular filter given by:

$$H_m[k] = \begin{cases} 0 & k < f[m-1] \\ \dfrac{2\big(k - f[m-1]\big)}{\big(f[m+1] - f[m-1]\big)\big(f[m] - f[m-1]\big)} & f[m-1] \le k \le f[m] \\ \dfrac{2\big(f[m+1] - k\big)}{\big(f[m+1] - f[m-1]\big)\big(f[m+1] - f[m]\big)} & f[m] \le k \le f[m+1] \\ 0 & k > f[m+1] \end{cases} \tag{6.140}$$

Such filters compute the average spectrum around each center frequency with increasing bandwidths, and they are displayed in Figure 6.28.

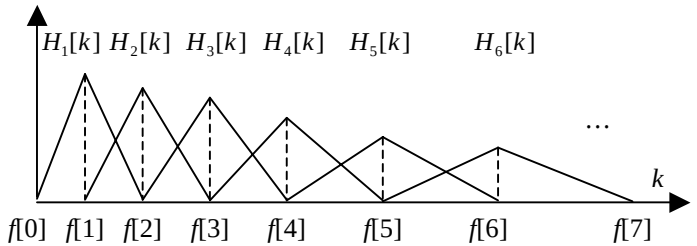

Figure 6.28 Triangular filters used in the computation of the mel-cepstrum using Eq. (6.140).

Alternatively, the filters can be chosen as

$$H_m'[k] = \begin{cases} 0 & k < f[m-1] \\ \dfrac{(k - f[m-1])}{(f[m] - f[m-1])} & f[m-1] \le k \le f[m] \\ \dfrac{(f[m+1] - k)}{(f[m+1] - f[m])} & f[m] \le k \le f[m+1] \\ 0 & k > f[m+1] \end{cases} \tag{6.141}$$

which satisfies $\sum_{m=0}^{M-1} H_m'[k] = 1$. The mel-cepstrum computed with $H_m[k]$ or $H_m'[k]$ will differ by a constant vector for all inputs, so the choice becomes unimportant when used in a speech recognition system that has trained with the same filters.

Let's define f_l and f_h to be the lowest and highest frequencies of the filterbank in Hz, F_s the sampling frequency in Hz, M the number of filters, and N the size of the FFT. The boundary points $f[m]$ are uniformly spaced in the mel-scale:

$$f[m] = \left(\frac{N}{F_s}\right) B^{-1}\left(B(f_l) + m\frac{B(f_h) - B(f_l)}{M+1}\right) \tag{6.142}$$

where the mel-scale B is given by Eq. (2.6), and B^{-1} is its inverse

$$B^{-1}(b) = 700\left(\exp(b/1125) - 1\right) \tag{6.143}$$

We then compute the log-energy at the output of each filter as

$$S[m] = \ln\left[\sum_{k=0}^{N-1} |X_a[k]|^2 H_m[k]\right], \qquad 0 \le m < M \tag{6.144}$$

The mel frequency cepstrum is then the discrete cosine transform of the M filter outputs:

$$c[n] = \sum_{m=0}^{M-1} S[m]\cos\left(\pi n(m+1/2)/M\right) \qquad 0 \le n < M \tag{6.145}$$

where M varies for different implementations from 24 to 40. For speech recognition, typically only the first 13 cepstrum coefficients are used. It is important to note that the MFCC representation is no longer a homomorphic transformation. It would be if the order of summation and logarithms in Eq. (6.144) were reversed:

$$S[m] = \sum_{k=0}^{N-1} \ln\left(|X_a[k]|^2 H_m[k]\right) \qquad 0 \le m < M \tag{6.146}$$

In practice, however, the MFCC representation is approximately homomorphic for filters that have a smooth transfer function. The advantage of the MFCC representation using