Mel Cepstrum Frequency Coefficients

Physiology of the Ear [1]

Cochlea is an inner structure that communicates directly with the auditory nerve, conducting a representation of sound to the brain. The cochlea is a spiral tube about 3.5 cm long, which coils about 2.6 times. The spiral is divided, primarily by the basilar membrane running lengthwise, into two fluid-filled chambers. The cochlea can be roughly regarded as a filter bank, whose outputs are ordered by location, so that a frequency-to-place transformation is accomplished. The filters closest to the cochlear base respond to the higher frequencies, and those closest to its apex respond to the lower.

Physical vs Perceptual Attributes

In psychoacoustics, a basic distinction is made between the perceptual attributes of a sound, especially a speech sound, and the measurable physical properties that characterize it.

Table 1: Relation between perceptual and physical attributes of sound.

Physical Quantity	Perceptual Quality
Intensity	Loudness
Fundamental Frequency	Pitch
Spectral shape	Timbre
Onset/Offset time	Timing
Phase difference in binaural hearing	Location

Frequency Analysis

Researchers have attempted derive a frequency scale to model the natural response of the human perceptual system, since the cochlea of the inner ear acts as a spectrum analyzer. The scale is not simple or linear. The cochlea acts as if it were made up of overlapping filters having bandwidths equal to the critical bandwidth.

Mel scale:

Mel scale is linear below 1kHz, and logarithmic above with equal numbers of samples taken below and above 1 kHz. The mel scale is based on experiments with simple tones (sinusoids) in which subjects were required to divide given frequency ranges into four perceptually equal intervals or to adjust the frequency of a stimulus tone to be half as high as that of a comparison tone. One mel is defined as one thousandth of the pitch of a 1 kHz tone. As with all such attempts, it is hoped that

the mel scale more closely models the sensitivity of the human ear than a purely linear scale and provides for greater discriminatory capability between speech segments. It can be approximated by:

$$B(f) = 1127\, ln\left(1 + \frac{f}{700}\right) \tag{1}$$

Spectral Analysis

- **From a finite record of a stationary data sequence, estimate how the total power is distributed over frequency.**
- Problem of determining the spectral content (i.e., the distribution of power over frequency) of a time series.
- In speech analysis, spectral models of voice signals are useful in better understanding the speech production process, and in addition can be used for both speech synthesis (or compression) and speech recognition.

Cepstrum Analysis:

The cepstrum of a signal is the Inverse Discrete-Time Fourier Transform (IDTFT) of the logarithm of the magnitude of the DTFT of the signal.

$$c[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|X(e^{i\omega})|e^{i\omega n}d\omega, \tag{2}$$

where the DTFT of the signal is defined as,

$$X(e^{i\omega}) = \sum_{n=-\infty}^{\infty} x[n]e^{-i\omega n}. \tag{3}$$

Mel-Frequency Cepstrum Coefficients (MFCC):

The term cepstrum is obtained by swapping the order of the letters in the word spectrum.

- Ceptrum transform:

```
Fourier Transform → Complex logarithm → Inverse Fourier Transform
```

- MFCC:

```
FT → Square of Magnitude → Mel Filter Bank → Real Logarithm → DCT
```

[2] In signal processing, a signal is viewed as a function of time. The term "size of a signal" is used to represent "strength of the signal". There are multiple ways top measure the signal size. Given a mathematical function, the area under the curve is a good measure of size of a signal. But a signal can have both positive and negative values and that leads to value cancelling each other out totally or partially. So we are left with two other options for defining the "size" of a signal,

1. Computation of the area under the absolute value of the function.
2. Computation of the area under the square of the function

The latter is favorable due to its mathematical tractability and its similarity to Euclidean Norm. Computation of the area under the square of the function, the energy of a continuous-time complex signal $x(t)$ is defined as

$$E_x = \int_{-\infty}^{\infty} |x(t)|^2 \, dt \tag{4}$$

If the signal x(t) is real, the modulus operator in the above equation does not matter. This is called "Energy" in signal processing terms. This is also a measure of signal strength.

The idea is to compute a frequency analysis based upon a filter bank with approximately critical band spacing of the filters and bandwidths. For 4 kHz bandwidth, approximately 20 filters are used. In most implementations, a short-time Fourier analysis is done first, resulting in a DFT $X_m[k]$ for the m^{th} frame. Then the DFT values are grouped together in critical bands and weighted by triangular weighting functions.

Calculating MFCCs:

- Splitting the signal into smaller frames (10 - 20ms)
- Compute the power spectrum for each frame
- Apply the mel filterbank to the power spectra, sum the energy in each filter.
- Take the logarithm of all filterbank energies.
- Apply DCT.
- Keep only the DCT coefficients 2-13.

Mel Scale

Filterbank Analysis:

Following equation is used to compute a Mel corresponding to a given frequency in Hertz,

$$F_{mel} = \frac{1000}{\log(2)} \cdot \left[1 + \frac{F_{hz}}{1000}\right] \tag{5}$$

where, F_{mel} is the equivalent mel scale frequency and F_{hz} is the normal frequency in Hertz.

Mel Scale to Frequency is:

$$M(f) = 1127 \ln(1 + \frac{f}{700}) \tag{6}$$

From frequency to Mel-scale is:

$$M^{-1}(m) = 700(e^{(\frac{m}{1127})} - 1) \tag{7}$$

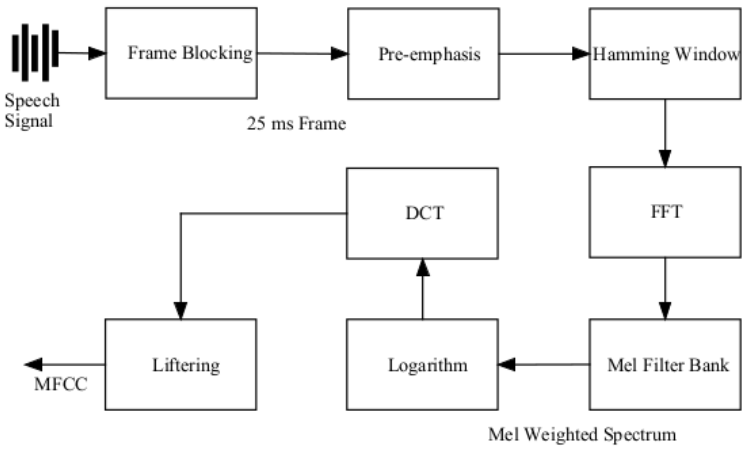

Figure 1: MFCC Block Diagram (**Credits:** Audio Processing and Speech Recognition Concepts, Techniques and Research Overviews by Soumya Sen, Anjan Dutta, Nilanjan Dey)

To change the log base,

$$\log_e(x) = \frac{log_{10}(x)}{log_{10}(e)} \tag{8}$$

Filterbank with M filters (m = 1, 2,. . . , M), where filter m is triangular filter given by:

$$H_m[k] = \begin{cases} 0 & k < f[m-1] \\ \frac{2(k-f[m-1])}{(f[m+1]-f[m-1])(f[m]-f[m-1])} & f[m-1] \leq k \leq f[m] \\ \frac{2(f[m+1]-k)}{(f[m+1]-f[m-1])(f[m+1]-f[m])} & f[m] \leq k \leq f[m+1] \\ 0 & k > f[m+1] \end{cases} \tag{9}$$

Such filters compute the average spectrum around each center frequency with increasing bandwidths as shown in the figure 2.

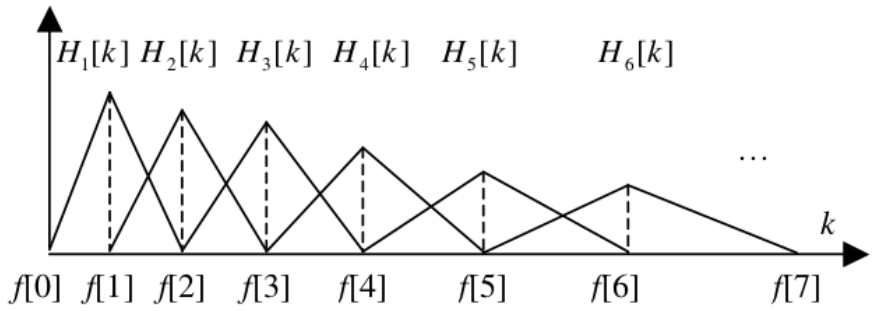

Figure 2: Triangular filters used in the computation of the mel-ceptrum.

Alternatively, the can be chosen as,

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \leq k < f(m) \\ 1 & k = f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & f(m) < k \leq f(m+1) \\ 0 & k > f(m+1) \end{cases} \tag{10}$$

Let's define f_l and f_h to be the lowest and highest frequencies of the filterbank in the Hz, Fs 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:

$$\frac{N}{F_s}\, B^{-1}\left(B(f_l) + m\frac{B(f_h) - B(f_l)}{M+1}\right) \tag{11}$$

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

$$B^{-1}(b) = 700\left(e^{\left(\frac{b}{1127}\right)} - 1\right) \tag{12}$$

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

$$S[m] = ln\left[\sum_{k=0}^{N-1} |X_a[k]|^2\ H_m[k]\right], \quad 0 \leq m \leq M \tag{13}$$

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

$$c[n] = \sum_{m=0}^{M-1} S[m]\, cos\left(\frac{\pi\, n}{N}(m + \frac{1}{2})\right), \quad 0 \leq n \leq M \tag{14}$$

First 13 cepstrum coefficients are used in speech recognition.

Implementation

***For more details, refer this wonderful article [3]**

- Split the sampled signal into 20-40ms frames (25ms is standard). Sampling rate of the signal is 16kHz and frame length is 0.025 * 16000 = 400 samples. Frame step is 10ms (Overlap to the frames). The first 400 sample frame starts at sample 0, the next 400 sample frame starts at sample 160 etc. until the end of the speech file is reached. If the speech file does not divide into an even number of frames, pad it with zeros so that it does.
- Apply DFT on the frames (0 - 399 samples frame). Then we compute the power spectrum of frame. We would generally perform a 512 point FFT and keep only the first 257 coefficients.
- Compute the Mel-spaced filterbank. 20-40 triangular filters (26 is standard) is applied to the power spectrum from step 2. Our filterbank comes in the form of 26 vectors of length 257 (assuming the FFT settings fom step 2). Each vector is mostly zeros, but is non-zero for a certain section of the spectrum. To calculate filterbank energies we multiply each filterbank with the power spectrum, then add up the coefficents. Once this is performed we are left with 26 numbers that give us an indication of how much energy was in each filterbank.
- Take the logarithm of each of the 26 energies from step 3, resulting with 26 log filterbank energies.
- Take the DCT of the 26 log filterbank energies to get 26 cepstral coefficients. For ASR, only the lower 12-13 of the 26 coefficients are kept.

Reference

[1]. Spoken Language Processing: A Guide to Theory, Algorithm and System Development

[2]. `https://www.gaussianwaves.com/2013/12/power-and-energy-of-a-signal/`

[3]. `http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral`

[4]. Audio Processing and Speech Recognition Concepts, Techniques and Research Overviews by Soumya Sen, Anjan Dutta, Nilanjan Dey