Feature Extraction: MFCC Vectors

In Feature Extraction, input waveform is transformed into a sequence of acoustic **feature vectors**, each vector representing the information in a small time window of the signal. One of the most common features is the **Mel Frequency Cepstral Coefficients (MFCC).**

The first step in speech processing is the Analog-to-Digital Conversion, which involves two steps: sampling and quantization. In order to efficiently represent the analog speech signal by its samples, the sampling frequency should be equal to atleast twice the maximum frequency of the speech signal. As most information in human speech is in frequencies below 10,000 Hz, a 20,000 Hz sampling rate would be necessary for complete accuracy. The samples are usually stored as integers, either 8-bit (values from-128 to 127) or 16 bit (values from -32768 to 32767). This process of representing real-valued numbers as integers is called **quantization** because there is a minimum granularity (the quantum size) and all values which are closer together than this quantum size are represented identically. The processes involved in extracting MFCC features are shown in Fig 5.4.1.

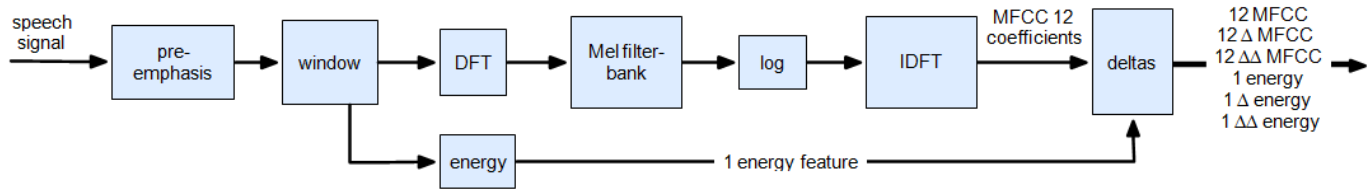

Fig 5.4.1 Extraction of MFCC Features from Quantized Digitized Waveform

5.4.1 Preemphasis

From the spectrum of voiced segments like vowels, it is observed that there is more energy at the lower frequencies than the higher frequencies. This drop in energy across frequencies called Spectral Tilt is caused by the nature of glottal pulse. Boosting the high frequency energy makes information from these higher formants more available to the acoustic model and improves phone detection accuracy. For example, the spectrum of the pronunciation of vowel [aa] before and after preemphasis is given in Fig 5.4.2.

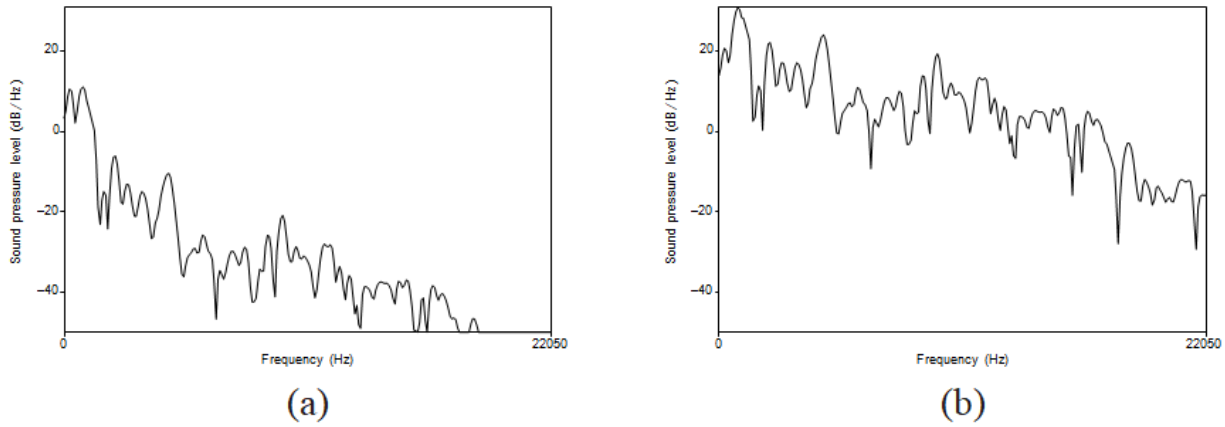

Fig 5.4.2 Spectrum of the pronunciation of vowel [aa] before (a) and after (b) preemphasis

5.4.2 Windowing

Although speech is a non-stationary signal, i.e. its statistical properties are not constant across time; it is assumed to be stationary in a small window that characterizes a particular subphone. This is done by running a window, which is non-zero inside some region and zero elsewhere, across the speech signal, and extracting the waveform inside this window. This windowing process is characterized by:

- **Width** of the window (in millisecond)
- **Offset** between successive windows, and
- **Shape** of the window.

The speech extracted from each window is said to be a **frame**, width of the frame is called **frame size** and the offset is said to be **frame shift**.

This windowing process can be represented as

$$y[n] = s[n]w[n]$$

where $s[n]$ is the speech signal and $w[n]$ is the window function of length L.

From Fig 5.4.3, using a rectangular window, $w[n] = \{1, \quad 0 \leq n \leq L-1 \; 0, \quad otherwise$, the extracted windowed signal looks just like the original signal. However the abrupt cuts at the boundaries would result in discontinuities in the spectrum. Hence, a Hamming window is most commonly used, which shrinks the values of the signal toward zero at the window boundaries, avoiding discontinuities.

$$w[n] = \{0.54 - 0.46 \cos cos\left(\frac{2\pi n}{L}\right) \quad 0 \leq n \leq L-1 \; 0 \quad otherwise$$

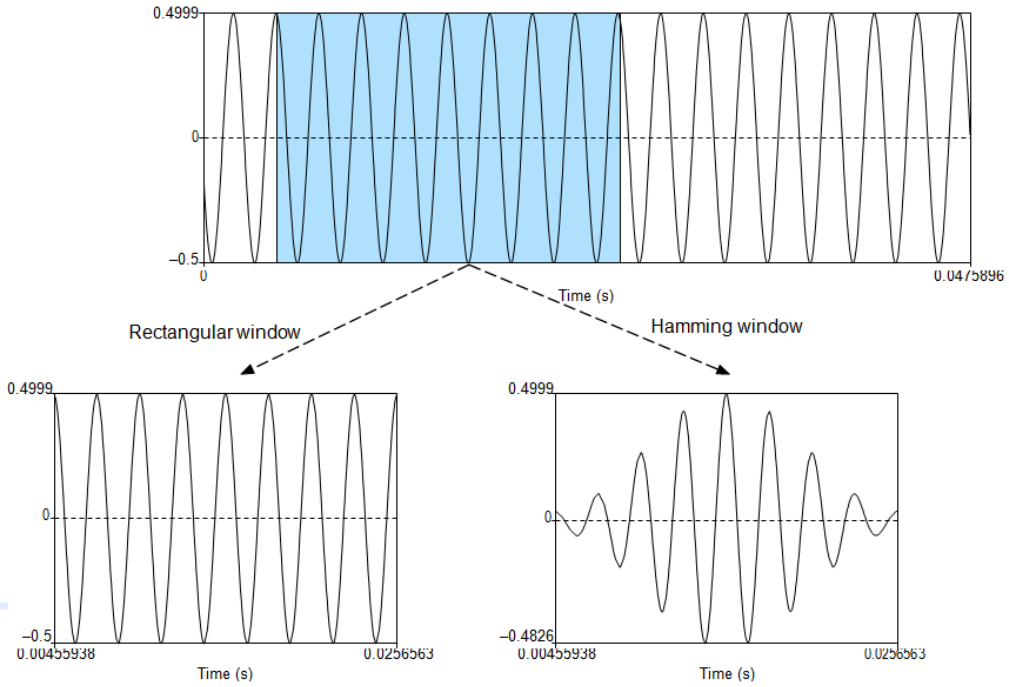

Fig 5.4.3 Windowing a sine wave with Rectangular and Hamming Window

5.4.3 Discrete Fourier Transform

Discrete Fourier Transform (DFT) is used to extract spectral information i.e. energy contained at different frequency bands of the discrete windowed signal.

$$X[k] = \sum_{n=0}^{N-1} x[n]e - j2\frac{\pi}{N}kn$$

where $X[k]$, $k = 0, 1, \ldots N - 1$ is the complex number representing the magnitude and phase of the frequency component in the original signal.

5.4.4 Mel filter bank and log

As the human hearing is less sensitive at higher frequencies, roughly above 1000 Hertz, modifying the output of DFT during feature extraction improves speech recognition performance. In MFCC, the frequencies output by the DFT are warped onto the mel scale. A **mel** is a unit of pitch defined so that pairs of sounds which are perceptually equidistant in pitch are separated by an equal number of mels. The mapping between frequency in Hertz and the mel scale is linear below 1000 Hz and the logarithmic above 1000 Hz. The mel frequency m can be computed from the raw acoustic frequency as follows:

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

As shown in Fig 5.4.4, this is implemented using a bank of filters which collect energy from each frequency band, with 10 filters spaced linearly below 1000 Hz, and the remaining filters spread logarithmically above 1000 Hz.

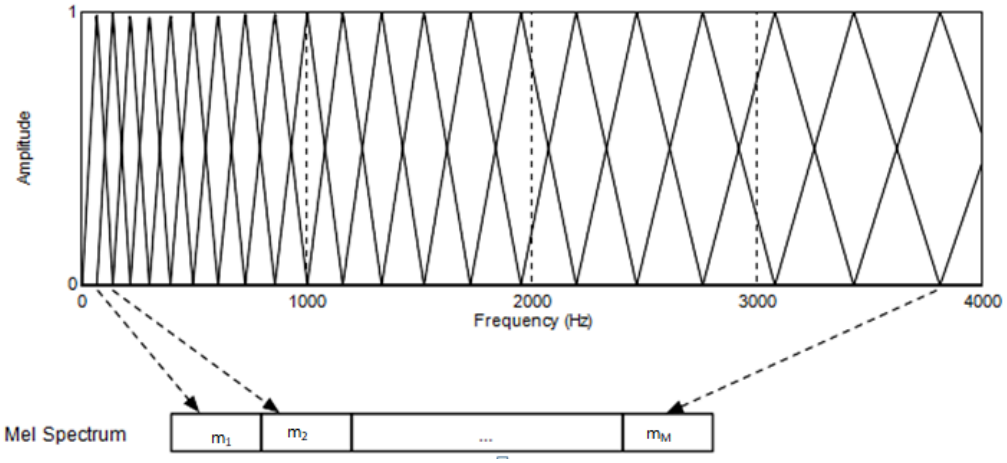

Fig 5.4.4 The Mel Filter Bank

Finally, log is applied to each mel spectrum values. This is because the human response to signal level is logarithmic; humans are less sensitive to slight differences in amplitude at high amplitudes than at low amplitudes.

5.4.5 The Cepstrum: Inverse Discrete Fourier Transform

The next step in MFCC feature extraction is the computation of the **cepstrum**. The cepstrum has a number of useful processing advantages and also significantly improves phone recognition performance. It is known that speech waveform is created when a glottal source waveform of a particular fundamental frequency is passed through the vocal tract, which has a particular filtering characteristic. Thus the generation of speech has a **source** and **filter**. However, the most useful information for phone detection is the **filter**, i.e. the exact position of the vocal tract. Cepstrum is used to deconvolve the source and filter and reveal only the vocal tract filter.

The cepstrum can be defined as the *spectrum of the log of the spectrum*.

- Consider a standard magnitude spectrum of a phone.
- Replace each amplitude value in the magnitude spectrum with its log.
- Visualize the log spectrum as if itself were a waveform (pseudo-signal)

- Apply inverse transform and obtain the spectrum of this pseudo-signal, to obtain the cepstrum (the word **cepstrum** is formed by reversing the first letters of **spectrum**)
- By taking the spectrum of the log spectrum, the signal is transformed from frequency domain to the time domain. Hence the unit of cepstrum is sample.

In the cepstrum, the glottal pulse can be viewed as a peak and lower cepstral values represent the vocal tract filter. The cepstral coefficients hold the property that the variances of the different coefficients are uncorrelated, whereas for the spectrum, spectral coefficients at different frequency bands are correlated. Cepstrum is defined as the **inverse DFT of the log magnitude of the DFT of a signal**. For a windowed frame of speech $x[n]$, the cepstum is given by:

$$c[n] = \sum_{n=0}^{N-1} log\left(\left|\sum_{n=0}^{N-1} x[n]e^{-j\frac{2\pi}{N}kn}\right|\right) e^{j\frac{2\pi}{N}kn}$$

5.4.6 Deltas and Energy

The **Energy** in a frame is the sum over time of the power of the samples in the frame. For a signal x in a window from time sample t_1 to time sample t_2, the energy is:

$$Energy = \sum_{t=t_1}^{t_2} x^2[t]$$

The Delta is a feature related to the change in cepstral features over time. The **delta** or **velocity** feature, and a **double delta** or **acceleration** feature are added to each of 13 features (12 cepstral features and 1 energy feature) of a frame. The 13 delta features represents the change between frames in the corresponding cepstral/energy feature, while the 13 double delta features represents the change between frames in the corresponding delta features. The delta value $d(t)$ for a particular cepstral value $c(t)$ at time t can be estimated as:

$$d(t) = \frac{c(t+1) - c(t-1)}{2}$$

Thus, a speech frame consists of 39 MFCC features i.e. 12 cepstral coefficients, 1 energy coefficients, 13 delta coefficients and 13 double delta coefficients.

5.5 Computing Acoustic Likelihoods

The output likelihood of the MFCC features vectors is computed by the B probability function of the HMM. Given an individual state q_i and an observation o_t, the observation likelihoods in B matrix is given by $p(o_t|q_i)$, called $b_t(i)$.

5.5.1 Vector Quantization (VQ)

The mapping of input MFCC vectors to discrete quantized symbols is called **vector quantization** or **VQ.** In vector quantization, each training feature vector is mapped into a small number of classes, and each class is represented by a discrete symbol. A vector quantization system is characterized by a **codebook**, a **clustering algorithm**, and a **distance metric**.

- A **codebook** is a list of possible classes, a set of symbols constituting a vocabulary $V = \{v_1, v_2, ..., v_n\}$. For each symbol v_k in the codebook, there is a **prototype vector**, known as a **codeword**, which is a specific feature vector. For example, for a codebook with 256 codeword, each vector is represented by a value from 0 to 255. This is referred to as 8-bit VQ, since we can represent each vector by a single 8-bit value. Each of these 256 values would be associated with a prototype feature vector.
- The codebook is created by using a **clustering** algorithm to cluster all the feature vectors in the training set into the 256 classes. Then, a representative feature vector from the cluster is chosen as the prototype vector or codeword for that cluster.
- Each incoming feature vector is compared to each of the 256 prototype vectors and the one which is closest (by some **distance metric**) is selected. The input vector is replaced by the index of this prototype vector.

A schematic of this process is shown in Fig. 5.5.1.

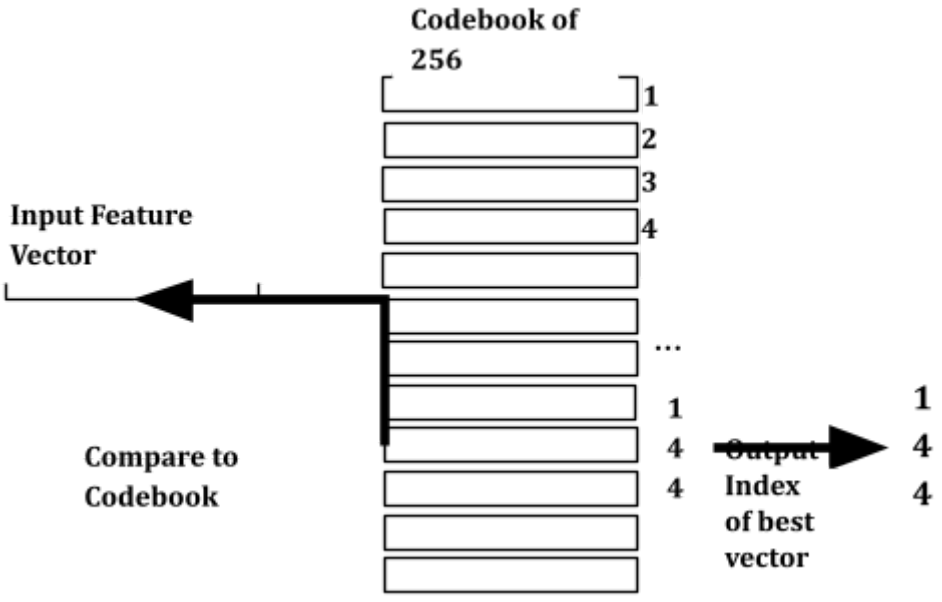

Fig 5.5.1 Schematic Architecture of Vector Quantization

The advantage of VQ is that since there are a finite number of classes, for each classv_k, the probability that it is generated can be computed by a given HMM state/sub-phone by simply counting the number of times it occurs in some training set when labeled by that state, and normalizing. Both the clustering process and the decoding process require a **distance metric** or **distortion** metric, that specifies how similar two acoustic feature vectors are. The distance metric is used to build clusters, to find a prototype vector for each cluster, and to compare incoming vectors to the prototypes.

The distance metric for acoustic feature vectors is **Euclidean distance**. Euclidean distance is the distance in N-dimensional space between the two points defined by the two vectors. Given a vector x and a vector y of length D, the square of the Euclidean distance between them is defined as:

$$d_{euclidean}(x, y) = \sum_{i=1}^{D} \left(x_i - y_i\right)^2 = (x - y)^T (x - y)$$

The Euclidean distance metric assumes that each of the dimensions of a feature vector are equally important. But actually each of the dimensions has very different variances. A large difference in a dimension with low variance should count more than a large difference in a dimension with high variance. A slightly more complex distance metric, the **Mahalanobis distance**, takes into account the different variances of each of the dimensions. If each dimension i of the acoustic feature vectors is assumed to have variance σ_i^2, then the Mahalanobis distance is:

$$d_{mahalanobis}(x, y) = \sum_{i=1}^{D} \frac{\left(x_i - y_i\right)^2}{\sigma_i^2}$$

Process Summary:

When decoding a speech signal, to compute an acoustic likelihood of a feature vector o_t given an HMM state q_j:

- compute the Euclidean or Mahalanobis distance between the feature vector and each of the N codewords
- choose the closest codeword, getting the codeword index v_k
- look up the likelihood of the codeword index v_k given the HMM state j in the pre-computed B likelihood matrix defined by the HMM:

$$\hat{b}_j(o_t) = b_j(v_k) \textit{ such that } v_k \textit{is codeword of closest vector to } o_t$$

Advantage of VQ: Extremely easy to compute and requires very little storage.

Drawback of VQ: Small number of codewords is insufficient to capture the wide variability in the speech signal.