# Feature Extraction for ASR: Preprocessing

Wantee Wang

2015-03-14 16:51:02 +0800

Audio signal is constantly changing, so to simplify analysis we need first frame the signal into short frames. Then we assume the signal within the short time is statistically stationary. Typically we choose the time of 25ms, and the frames are overlapped with shift of 10ms. If the frame is much shorter we don't have enough samples to get a reliable spectral estimate, if it is longer the signal changes too much throughout the frame.

## 1 DC Offset Removal

The first processing we do is to remove the *DC offset* of the signal. The DC offset is the mean value of the waveform. The term originated in electronics, where it refers to a direct current voltage. For a real sound wave propagated in the air, the mean value should equal to zero. Thus we remove the DC offset by subtracting the mean value from the original signal, i.e.,

$$x'[n] = x[n] - \frac{1}{N}\sum_i x[i]$$

## 2 Pre-emphasis

Pre-emphasis is performed for flattening the magnitude spectrum and balancing the high and low frequency components. It boosts the high frequencies component, thereby improving the signal-to-noise ratio, before they are transmitted or recorded onto a storage medium. Upon playback, a de-emphasis filter is applied to reverse the process.

The reason for using pre-emphasis in speech processing, is due to the rapid decaying spectrum of speech, when one deals with music signals , it is may not need to apply the filter. This decay in high-frequency part is seen to be suppressed during the sound production mechanism of humans. Moreover, it can also amplify the importance of high-frequency formants.

The formula for pre-emphasis filter is

$$x'[n] = x[n] - kx[n-1]$$

where $k$ is the pre-emphasis coefficient which should be in the range $0 \leq k < 1$, typical value is $k = 0.97$.

Take the $z$ transform for both sides,

$$X'(z) = X(z) - kX(z)z^{-1}$$

Therefore, $H(z) = \frac{X'(z)}{X(z)} = 1 - kz^{-1}$, the weight for low frequency is smaller than high frequency.

# 3 Hamming windowing

Hamming windowing is given by

$$x'[n] = \left\{ \alpha - \beta \cos(\frac{2\pi(n-1)}{N-1}) \right\} x[n]$$

where $\alpha = 0.54$ and $\beta = 0.46$.

It is used to deal with the finite Fourier transform problem. If the start and end of the finite samples don't match then that will look just like a discontinuity in the signal, and show up as lots of high-frequency nonsense in the Fourier transform. And if the samples happen to be a beautiful sinusoid but an integer number of periods don't happen to fit exactly into the finite sample, your FT will show appreciable energy in all sorts of places nowhere near the real frequency.

Windowing the data makes sure that the ends match up while keeping everything reasonably smooth, this greatly reduces the sort of spectral leakage. Detail explanation is in this link.