Notes on Automatic Speech Recognition

Speech Features

1 Auditory System as a Filter Bank

1.1 Critical-band Experiments

The critical band is determined experimentally (due to Fletcher) as follows.

1. A test listener is presented with a tone plus a wide-band noise.

2. In the beginning, the tone has low enough intensity that it is not perceived by the listener.

3. The intensity of the tone is increased until it is barely perceived. This intensity is called the *threshold intensity*.

4. The bandwidth of the noise is decreased and the corresponding threshold intensity is recorded.

5. When the threshold intensity starts to decrease, the noise bandwidth is the *critical band* of the frequency of the tone.

6. Repeat this experiment for several frequencies, one can obtain the cortical-band as a function of frequency.

It is found that the critical band of a high frequency is higher than the critical band of a low frequency. The Bark-scale is a good approximation to the critical band, which is

$$\Omega(\omega) = 6 \log \left\{ \frac{\omega}{1200\pi} + \left[\left(\frac{\omega}{1200\pi} \right)^2 + 1 \right]^{0.5} \right\}.$$

Basically, it is linear when ω is small and logarithmic when $\omega \geq 1000$ Hz.

1.2 Filter Shapes

The critical-band experiments only tell the bandwidths of the auditory filters but not the shapes. This can be determined experimentally (due to Patterson) as follows.

1. Subjects are presented with a tone plus a band noise. Note that the band of the noise does not cover the tone frequency.

2. Again we measure the threshold intensity as we vary the bandwidth of the noise. Assume that the threshold intensity and the noise power are proportional, we have

   $$P = K \int_0^\infty N(f)|H(f)|^2 df,$$

   where P is the threshold intensity, $N(f)$ is the noise power density, and $H(f)$ is the frequency response of the auditory filter. With the additional assumption that noise spectrum is rectangular with density N_0, one has

   $$|H(f)|^2 = \frac{1}{KN_0}\frac{dP}{df}.$$

3. In addition to low-pass noise, experiments are also carried out with high-pass noise.

4. Additionally, experiments are also carried out with band notched noise.

5. The empirical results are represented by a symmetric filter

   $$|H(f)|^2 = \frac{1}{[(\delta f/\alpha)^2 + 1]^2}.$$

2 Cepstrum

2.1 Source-Filter Model for Speech Generation

The speech signal can be modeled as the resonator (vocal tract) output excited by the source (the vocal cords). Here the source are related to pitch while the resonator is related to phone identity. Often we want to separate the resonator and the source in the signal analysis. This is called the deconvolution of the signal. In Mandarin, pitch is used for tone recognition while the resonator characteristics is used for base-syllable recognition.

2.2 The Real Cepstrum

Let E be excitation and V be vocal tract resonator, then

$$X(z) = E(z)V(z),$$

where X is the convolved signal. It follows that

$$\log |X(z)| = \log |E(z)| + \log |V(z)|.$$

As a function of the frequency (on the unit circle of z-plane, E is relatively fast-changing while V is slow-changing. Using this property, if one can separate the fast-changing and slow-changing (*high-time* and *low-time*) components in X, then E and V can be separated.

Note that this is very similar to the ideas of high-pass and low-pass filters. Only that we are now in the frequency domain rather than the time domain. There are a few terms invented to emphasize the difference: *lifter* vs. filter, *quefrecy* vs. frequency, and *cepstrum* vs. spectrum.

The cepstum is the inverse z-transform of $\log |X(z)|$, using the contour of unit circle

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

2.3 The Complex cepstrum

The complex cepstum is the inverse z-transform of $\log X(\omega)$,

$$\hat{x}[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log X(\omega) e^{j\omega n} d\omega.$$

It provide the phase information in addition to the magnitude information.

3 Linear Prediction

3.1 Predictive Model

We have seen that poles of the z-transform produce resonance frequencies on the unit circle. Since a formant corresponds to a resonance frequency of the vocal tract, we can approximate each formant by a pair of poles (cf. Section 6.8 of the textbook), with transfer function

$$H(z) = \frac{1}{1 - bz^{-1} - cz^{-2}}.$$

Generalizing this to multiple poles, we have

$$H(z) = \frac{1}{1 - \sum_{i=1}^{P} a_i z^{-i}}.$$

This is the all-pole model for the vocal tract. This transfer function actually corresponds to the time-domain specification

$$y[n] = x[n] + \sum_{i=1}^{P} a_i y[n-i].$$

The second term on the right-hand side is the P-th order *linear predictor* of $y[n]$ by its most recent P samples. This relates the linear prediction to the spectrum. The problem now is to determine the $a_i's$, called linear prediction coefficients (LPC), such that the prediction error is minimized.

Given a windowed signal $y[n], n = 0, \dots, N-1$, let the error signal be defined as

$$e[n] = y[n] - \tilde{y}[n] = y[n] - \sum_{i=1}^{P} a_i y[n-i].$$

To minimize $\sum_{n=0}^{N-1} e^2[n]$, the $a_i's$ must satisfy

$$\sum_{i=1}^{P} a_i \phi(j,i) = \phi(j,0),$$

where $\phi(j,i) = \sum_{n=0}^{N-1} y[n-j]y[n-i]$. There are standard approaches to to solve $a_i's$ for the above equation, such as the Levinson/Durbin recursions.