1 Introduction

The pitch determination is very important for many speech processing algorithms. For example, the concatenative speech synthesis methods require pitch tracking on the desired speech segments if prosody modification is to be done. Chinese speech recognition systems use pitch tracking for tone recognition, which is important in disambiguating the myriad of homophones. Pitch is also crucial for prosodic variation in text-to-speech systems and spoken language systems.

In this project, pitch detection methods via autocorrelation method, cepstrum method, harmonic product spectrum (HPS), and linear predictive coding (LPC) are examined.

2 Pitch Detection via Autocorrelation Method

2.1 Autocorrelation Method

A commonly used method to estimate pitch (fundamental frequency) is based on detecting the highest value of the autocorrelation function in the region of interest. Our perception of pitch is strongly related to periodicity in the waveform in the time domain. A method to estimate fundamental frequency from the waveform directly is to use autocorrelation.

The statistical autocorrelation of a sinusoidal random process

\[ x[n] = \cos(w_0 n + \phi) \] (1)

is given by

\[ R[m] = E\{x^*[n]x[n+m]\} = \frac{1}{2} \cos(w_0 m) \] (2)

which has maxima for \( m = lT_0 \), the pitch period and its harmonics, so that we can find the pitch period by computing the highest value of the autocorrelation. Similarly, it can be shown that any WSS periodic process with period \( T_0 \) also has an autocorrelation which exhibits its maxima at \( m = lT_0 \).

In practice, we need to obtain an estimate \( \hat{R}[m] \) from knowledge of only \( N \) samples. The empirical autocorrelation function is given by

\[ \hat{R}[m] = \frac{1}{N} \sum_{n=0}^{N-1-|m|} (w[n]x[n]w[n+m]x[n+m]) \] (3)

where \( w[n] \) is a window function of length \( N \). For the random process in Eq. (1), results in an expected value of

\[ E\{\hat{R}[m]\} = \left( 1 - \frac{|m|}{N} \right) \frac{\cos(w_0 m)}{2}, |m| < N \] (4)

whose maximum coincides with the pitch period for \( m > m_0 \) [1].

Since pitch periods can be as low as 40Hz (for a very low-pitched male voice) or as high as 600 Hz (for a very high-pitched female or child’s voice), the search for the maximum is conducted within a region.
2. PITCH DETECTION VIA AUTOCORRELATION METHOD

2.2 Experiments

Figure 1: Waveform (My voice recording “bee”) and unsmoothed pitch track with autocorrelation method. Notice that the pitch values in the weakly voiced or unvoiced regions are essentially random.

Figure 2: Waveform and autocorrelation function for frame 5 in Fig. 1. The estimated pitch is 156.863Hz.
3 Pitch Detection via Cepstral Method

3.1 Cepstral Method

Cepstral analysis provides a way for the estimation of pitch. If we assume that a sequence of voiced speech is the result of convoluting the glottal excitation sequence $e[n]$ with the vocal tract’s discrete impulse response $\theta[n]$. In frequency domain, the convolution relationship becomes a multiplication relationship. Then, using property of log function $\log AB = \log A + \log B$, the multiplication relationship can be transformed into an additive relationship. Finally, the real cepstrum of a signal $s[n] = e[n] * \theta[n]$ is defined as

$$c[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log|S(w)|e^{jn}dw$$

where

$$S(w) = \sum_{n=-\infty}^{\infty} s[n]e^{-jwn}.$$  

That is, the cepstrum is a Fourier analysis of the logarithmic amplitude spectrum of the signal. If the log amplitude spectrum contains many regularly spaced harmonics, then the Fourier analysis of the spectrum will show a peak corresponding to the spacing between the harmonics: i.e. the fundamental frequency. Effectively we are treating the signal spectrum as another signal, then looking for periodicity in the spectrum itself.

The cepstrum is so-called because it turns the spectrum inside-out. The x-axis of the cepstrum has units of quefrency, and peaks in the cepstrum (which relate to periodicities in the spectrum) are called rahmonics. To obtain an estimate of the fundamental frequency from the cepstrum we look for a peak in the quefrency region corresponding to typical speech fundamental frequencies ($1$/quefrency).

3.2 Experiments

![Figure 3: Waveform and unsmoothed pitch track with cepstral method.](image)
4. Pitch Detection via HPS method

4.1 Harmonic Product Spectrum (HPS)

If the input signal is a musical note, then its spectrum should consist of a series of peaks, corresponding to fundamental frequency with harmonic components at integer multiples of the fundamental frequency. Hence when we compress the spectrum a number of times (downsampling), and compare it with the original spectrum, we can see that the strongest harmonic peaks line up. The first peak in the original spectrum coincides with the second peak in the spectrum compressed by a factor of two, which coincides with the third peak in the spectrum compressed by a factor of three. Hence, when the various spectrums are multiplied together, the result will form clear peak at the fundamental frequency.

4.2 Experiments
4. PITCH DETECTION VIA HPS METHOD

Figure 5: Waveform and unsmoothed pitch track with HPS method.

Figure 6: Waveform and Cepstrum for frame 4 in Fig. 5. The estimated pitch is 155.3562Hz.
5 Formants Detection via LPC Method

5.1 Linear Predictive Coding (LPC)

A very powerful method for speech analysis is based on linear predictive coding (LPC), also known as auto-regressive (AR) modeling. This method is widely used because it is fast and simple, yet an effective way of estimating the main parameters of speech signals.

An all-pole filter with a sufficient number of poles is a good approximation for speech signals. Thus, we could model the filter \( H(z) \) as

\[
H(z) = \frac{X(z)}{E(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{1}{A(z)}
\]

where \( p \) is the order of the LPC analysis. Taking inverse z-transform in Eq. 7 results in

\[
x[n] = \sum_{k=1}^{p} a_k x[n-k] + e[n].
\]

Linear predictive coding gets its name from the fact that it predicts the current sample as a linear combination of its past \( p \) samples.

If we plot \( H(e^{jw}) \), we expect to see peaks at the roots of the denominator. From this fact, we would be able to detect formant frequencies [1].

5.2 Experiments

The number of LPC coefficients \( p \) is created by the next formula

\[
p = 2 + \text{sampling frequency}/1000
\]

as a rule of thumb. For example, 44.1kHz gives 46.

![Waveform and unsmoothed formants track with LPC method](image-url)
Figure 8: Waveform and Fourier Analysis of LPC filter for frame 5 in Fig. 7. The estimated formants are 380.4Hz, 1070.5Hz, 3215.7Hz, 5662.0Hz, 6454.8Hz.

6 Conclusion

Four methods for pitch determination, autocorrelation method, cepstrum method, HPS method, and LPC method, were examined. The first three methods obtained the similar pitch frequency, whereas, the LPC method obtained a different pitch frequency. The LPC method would be used to detect not only pitch but also formants.

Bibliography


