Notes on pitch detection and linear prediction

Lucio Bianchi

January 14, 2014

1 Pitch detection

1.1 Zero-Crossing Rate

Zero-Crossing Rate is the rate of sign-changes along a waveform $x(n)$ of duration N, i.e. the number of times that the signal crosses the zero level reference:

$$\text{zcr} = \frac{1}{N}\sum_{n=1}^{N-1}|\text{sign}(x(n)) - \text{sign}(x(n-1))|. \tag{1}$$

Starting from the zero-crossing rate, we can estimate the pitch of a sound signal as

$$\text{pitch} = \text{zcr} \cdot \frac{Fs}{2}. \tag{2}$$

1.2 Autocorrelation

The autocorrelation function is defined as

$$r(l) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)x(n-\tau). \tag{3}$$

The autocorrelation function has the following properties.

- For a pure tone, the ACF exhibits peaks at $L, 2L, 3L, \ldots,$ where L is the period of the tone. The peak at lag L will be higher than peaks at $2L, 3L, \ldots$.
- For a real signal:
  - the fundamental frequency component will behave like a pure tone, with the highest peak at lag L
  - other harmonics will produce one peak (not the highest) at lag L.

Thanks to these observations we can conclude that a large peak in $r(l)$ will occour at the lag L corresponding to the period of the fundamental frequency of the signal, resulting from the sum of the contribustions due to all harmonic components.

1.3 Cepstrum

A speech signal $y(n)$ can be modeled as the superposition of an excitation $x(n)$ (possibly containing the pitch) and resonances $h(n)$ (due to vocal tract, ...):

$$Y(\omega) = H(\omega)X(\omega). \tag{4}$$

Since $X(\omega$ will contain the pitch, we aim at separating these two components. The separation is not trivial, but in the literature have been proposed to exploit properties of the logarithms:

$$\log|Y(\omega)| = \log|H(\omega)X(\omega)| = \log|H(\omega)| + \log|X(\omega)|. \tag{5}$$

By observing the signal $\mathcal{F}^{-1}\{\log|Y(\omega)|\}$, we can notice two components:

- a quick oscillation, due to the harmonic structure of the speech;
- a slow behavior, related to resonances.

Hence, we can separate the two components in time domain:

$$\mathcal{F}^{-1}\{\log|Y(\omega)|\} = \mathcal{F}^{-1}\{\log|X(\omega)|\} + \mathcal{F}^{-1}\{\log|H(\omega)|\}, \tag{6}$$

where

- the part of $\mathcal{F}^{-1}\{\log|Y(\omega)|\}$ towards the origin describes the spectral envelope, while
- the part of $\mathcal{F}^{-1}\{\log|Y(\omega)|\}$ far from the origin describes the excitation.

2 Linear prediction

The idea behind linear prediction is to approximate a voiced speech signal as the superposition of a linear combination of past samples of the signal and an excitation signal:

$$S(z) = \sum_{p=1}^{P} a_p z^{-p} S(z) + gX(z). \tag{7}$$

By defining

$$A(z) = \sum_{p=1}^{P} a_p z^{-p} \tag{8}$$

we can write

$$S(z) = A(z)S(z) + gX(z), \tag{9}$$

$$S(z)\left(1 - A(z)\right) = gX(z), \tag{10}$$

$$S(z) = \frac{g}{1 - A(z)} X(z). \tag{11}$$

Consider the linear combination of past samples as a predictor for the signal $s(n)$

$$\hat{s}(n) = \sum_{p=1}^{P} a_p s(n-p). \tag{12}$$

We can define the prediction error as

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{p=1}^{P} a_p s(n-p), \tag{13}$$

hence

$$E(z) = (1 - A(z))\, S(z). \tag{14}$$

The parameters a_p are usually chosen by minimizing the expected value of the squared prediction error, i.e.

$$\text{minimize}_{a_p} \mathrm{E}[e^2(n)], \tag{15}$$

which leads to Yule-Walker equations

$$\sum_{p=1}^{P} a_p r(l-k) = -r(l), \text{ for } 1 \le l \le P, \tag{16}$$

where $r(l) = \mathrm{E}[s(n)s(n-l)]$ is the autocorrelation function.

In order to solve numerically the problem of computing the parameters a_p, we can rewrite (16) in matrix form by defining the parameter vector

$$\mathbf{a} = \begin{bmatrix} a_1 & \dots & a_P \end{bmatrix}^T, \tag{17}$$

the autocorrelation matrix (which has the topology of a Toeplitz matrix)

$$[\mathbf{R}]_k^l = r(l-k),$$

$$\mathbf{R} = \begin{bmatrix} r(0) & r(1) & \dots & r(P-1) \\ r(1) & r(2) & \dots & r(P-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(P-1) & r(P-2) & \dots & r(0) \end{bmatrix}, \tag{18}$$

and the vector

$$\phi = \begin{bmatrix} r(1) & \dots & r(P) \end{bmatrix}^T. \tag{19}$$

Hence (16) written in matrix form is

$$\mathbf{Ra} = \phi, \tag{20}$$

which can be inverted as

$$\mathbf{a} = \mathbf{R}^{-1}\phi. \tag{21}$$

3 Voiced/Unvoiced classification

A simple but effective idea to discriminate between voiced and unvoiced speech is based on three classification criteria:

- Cepstrum intensity,
- Zero-Crossing Rate,
- Short-time energy.

Cepstrum intensity The use of Cepstrum intensity to identify a voiced signal is based on the consideration that voiced segments have strong and sharp peaks due to periodicity. Hence we look for the maximum of the cepstrum in the ith segment of the speech signal $s_i(n)$

$$C_i = \max(\mathrm{Re}\{\mathcal{F}^{-1}\{\log|s_i(n)|\}\}). \tag{22}$$

The ith segment is likely to be voiced if

$$C_i > \tau_{\mathrm{cep}}, \text{ where } \tau_{\mathrm{cep}} = \mathrm{median}(C). \tag{23}$$

Zero-Crossing Rate The use of Zero-Crossing Rate is motivated by the fact that zcr is higher for unvoiced rather than unvoiced segments. The ith segment is likely to be voiced if

$$\mathrm{zcr}_i < \tau_{\mathrm{zcr}}, \text{ where } \tau_{\mathrm{zcr}} = \mathrm{median}(\mathrm{zcr}). \tag{24}$$

Short-time Energy The use of Short-Time Energy is motivated by the fact that voiced segments have higher energy than unvoiced segments. The short-time energy is defined as the energy of the ith frame, i.e.

$$\mathrm{ste}_i = \sum_{n=1}^{N} |s(n)|^2. \tag{25}$$

The ith segment is likely to be voiced if

$$\mathrm{ste}_i > \tau_{\mathrm{ste}}, \text{ where } \tau_{\mathrm{ste}} = \mathrm{median}(\mathrm{ste}). \tag{26}$$