Frequency-Domain Pitch Estimation

Guy Narkiss

Source: Discrete-Time Speech Signal Processing \ Quatieri, chapter 10

Introduction

- Reliable estimation of the fundamental frequency is of high importance and has many applications to speech coding, speech synthesis and speech recognition.
- Different approaches to speech analysis/synthesis naturally lead to differnet methods for pitch and voicing estimation.
- Pitch and voicing estimation algorithms based on the sinusoidal model will be introduced.

Presentation outline

- Background:
  - Common pitch estimation methods.
  - Sinusoidal Speech model.
- Pitch estimation based on sinusoidal model.
- Voicing detection.
- Multi band pitch and voicing estimation.

The pitch

- Duration: 2.5-16msec (20-128 samples at 8Khz).
- Frequency: 62.5-400Hz.

Voiced speech /o/

Unvoiced plosive /k/

Unvoiced fricative /f/

Voiced plosive /g/

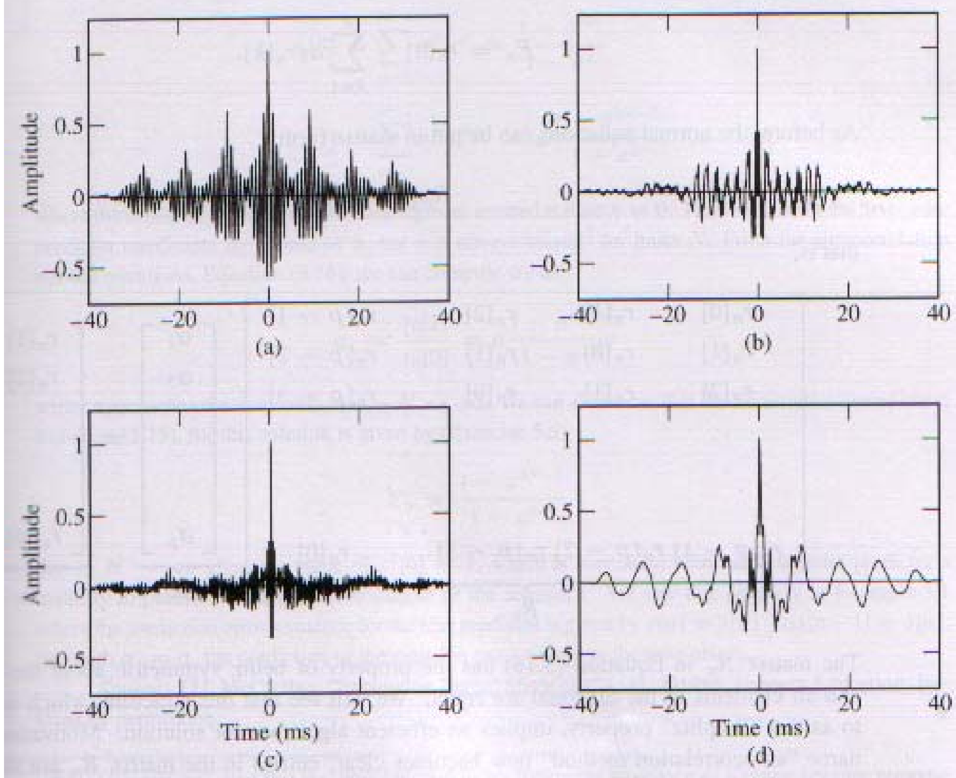


Figure 5.6 Illustration of autocorrelation functions of speech: (a) vowel /o/ in "pop"; (b) unvoiced plosive /k/ in "baker"; (c) unvoiced fricative /f/ in "father"; (d) voiced plosive /g/ in "go."

Common pitch estimation methods

- Autocorrelation method:

Define the short-time sequence $s_n[m]$

$$s_n[m] \triangleq s[m]w[n-m]$$

$$\left(S_n[m]=0 \quad \forall m \notin \left[n-\tfrac{N_w-1}{2}, n+\tfrac{N_w-1}{2}\right]\right)$$

$$w[n] - Analysis\ window, length\ N_w\ (odd)$$

$$s[m] - Speech\ signal$$

Define the short-time Autocorrelation $r_n[\tau]$

$$r_n[\tau] = s_n[\tau] * s_n[-\tau] = \sum_{m=-\infty}^{\infty} s_n[m] s_n[m+\tau]$$

Autocorrelation method

In order to minimize the MSE:

$$E[P] = \sum_{m=-\infty}^{\infty} \left(s_n[m] - s_n[m+P] \right)^2$$

We will choose:

$$\hat{P} = \arg\max_{p} \sum_{m=-\infty}^{\infty} s_n[m] s_n[m+P]$$

$$\hat{P} > \varepsilon$$

$$r_n[\tau] > Threshold$$

Common pitch estimation methods

- AMDF (Average Magnitude difference Function)

$$A_n[P] = \sum_{k=n-N_w+P}^{n} \left| s[k] - s[k-P] \right|, \quad {\scriptstyle 20 \le P \le 128}$$

$$\hat{P} = \arg\min_{P} A_n[P], \quad {\scriptstyle (A_n[P] < Threshold)}$$

- Cepstral Method (CEP)

$$\mathrm{c[n]} \triangleq \mathrm{IDFT}\left\{ \log \left| \mathrm{DFT}\left\{ \mathrm{s[n]} \right\} \right| \right\}$$

$$\hat{P} = \arg\max_{P} \left(Re\left\{ c[P] \right\} \right), \quad {\scriptstyle c[P] > Threshold;\, P_{\min} < P < P_{\max}}$$

Common pitch estimation methods

- LPC based pitch detector (SIFT- Simplified Inverse Filtering Technique)
  1. Calculate the All-pole filter $\frac{1}{A(Z)}$.
  2. Filter the speech with A(z) to produce the residual signal (=excitation) u[n].
  3. $\hat{P}$ = lag between peaks of u[n].

Typical drawbacks of pitch estimators

- Pitch doubling.
- Pitch halving.
- Mixed V/UV segments.
- Adaptive threshold.
- First formant.
- Speech corruption by noise.

Solutions:

- Non-linear transformation of original speech signal.
- Pitch tracking from frame to frame (recursive smoothing, median filter, limit change rate…).

Sinusoidal Speech model

$$s(t) = (\mathrm{Re}) \sum_{k=1}^{K} A_k e^{j(\omega_k t + \theta_k)}$$

$$K(t) - Number\ of\ frequencies.$$

$$A_k(t), \omega_k(t), \theta_k(t) - Amp., Freq.\ phases.$$

- Model parameters are calculated using STFT.
- Analysis window length must be at least 2.5 times the average pitch.

Pitch estimation based on sinusoidal model

- The sinusoidal model representation is used to extrapolate the short-time sequence $s_n[m]$.

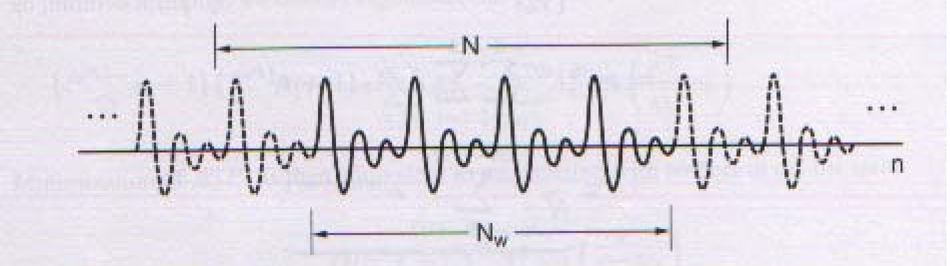


Figure 10.1 Extrapolation of a short-time segment from a periodic sequence.

- The same MSE criterion as in the autocorrelation method is used, but this time with the signal:
$\tilde{s}[m] = \sum_{k=1}^{K} A_k e^{j(\omega_k m + \theta_k)}$ instead of $s_n[m]$.

The MSE:

$$E[P]=\frac{1}{N}\sum_{m=-(N-1)/2}^{(N-1)/2}\left|\sum_{k=1}^{K}A_k\left[e^{j(mw_k+\theta_k)}-e^{j((m+P)w_k+\theta_k)}\right]\right|^2$$

Substituting $\left|\sum_{k=1}^{K}[\cdot]\right|^2$ by $\sum_{k=1}^{K}[\cdot]\sum_{l=1}^{K}[\cdot]$ and letting the extrapolation interval N go to infinity yields:

$$E[P]\approx\sum_{k=1}^{K}A_k^2\left(1-\cos\left(P\omega_k\right)\right)$$

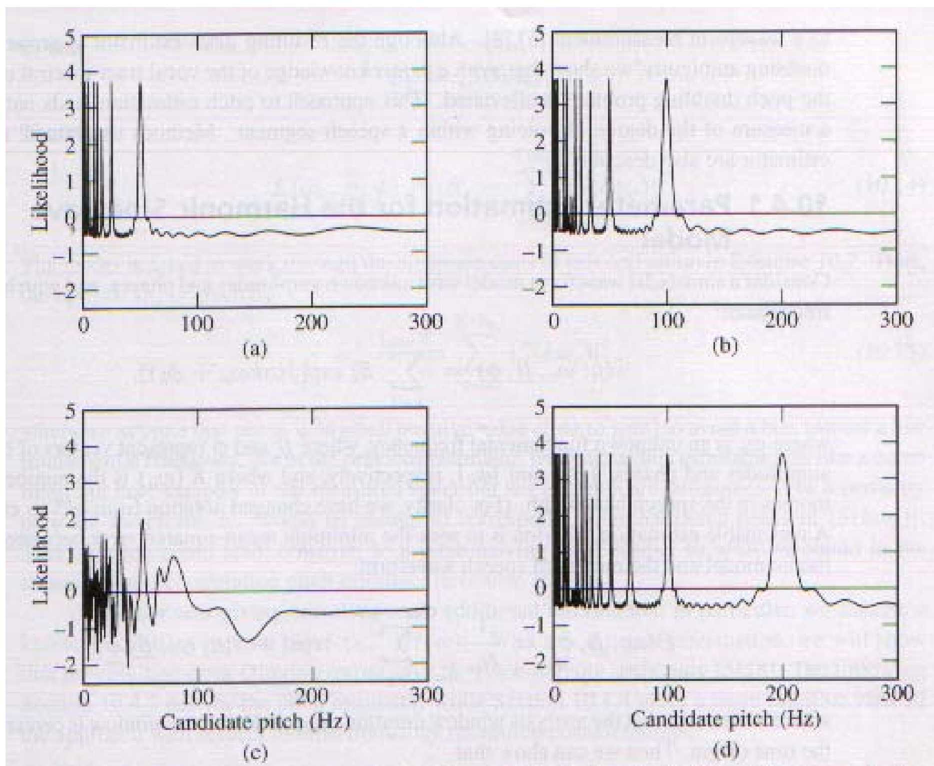
Define: $P\triangleq\frac{2\pi}{\omega_0}$ and $Q(\omega_0)\triangleq\sum_{k=1}^{K}A_k^2\cos\left(\frac{2\pi}{\omega_0}\omega_k\right)$

$Q(\omega_0)$ is the *likelihood function* of ω_0 to be the true pitch.

- Algorithm: $\hat{\omega}_0 = \arg\max Q(\omega_0)$
- If the ω_k's are multiples of a fundamental frequency $\tilde{\omega}_0$, and if $\omega_0 = \tilde{\omega}_0$, then $E[P]=0$

Pitch freq = 50Hz

Pitch freq = 100Hz



Pitch freq = 50Hz
With noise

Pitch freq = 200Hz

Figure 10.3 Pitch likelihood function $Q(\omega_o)$ for (a) 50, (b) 100, and (d) 200 Hz. The effect of noise on (a) is shown in (c).

Pitch estimation based on sinusoidal model

- Pitch doubling problem is solved.
- Pitch halving problem still exists.

Pitch Estimation Based on a Harmonic Sinewave Model

- Harmonic sinewave model:

$$s\left[n;\omega_0,\underline{\phi}\right] = \sum_{k=1}^{K(\omega_0)} \overline{A}\left(k\omega_0\right) e^{j\left(nk\omega_0+\phi_k\right)}$$

$K(\omega_0) - \textit{Number of harmonics}$ in $\textit{the speech bandwidth}.$

$\overline{A}\left(\omega\right) = \left|H\left(\omega\right)\right| - \textit{Vocal tract envelope}$

$(\textit{for unity excitation magnitude}).$

$\underline{\phi} - \textit{Phases of the harmonics}$

- Our goal is to fit the best harmonic model to the speech.

The MSE:

$$E\left(\omega_0,\underline{\phi}\right)=\frac{1}{N_w}\sum_{m=-\left(N_w-1\right)/2}^{\left(N_w-1\right)/2}\left|s\left[n\right]-\hat{s}\left[n;\omega_0,\underline{\phi}\right]\right|^2$$

Using the harmonic model representation leads to:

$$E\left(\omega_0\right)=P_s-2\underbrace{\left(\sum_{k=1}^{K(\omega_0)}\overline{A}\left(k\omega_0\right)\left|S\left(k\omega_0\right)\right|-\tfrac{1}{2}\sum_{k=1}^{K(\omega_0)}\overline{A}^2\left(k\omega_0\right)\right)}_{\rho(\omega_0)}$$

$P_s - signal\ average\ energy$

$S\left(k\omega_0\right) - STFT\ at\ frequency\ k\omega_0$

Algorithm: $\hat{\omega}_0=\underset{\omega_0}{\arg\max}\ \rho\left(\omega_0\right)$

Pitch estimation based on harmonic sinewave model

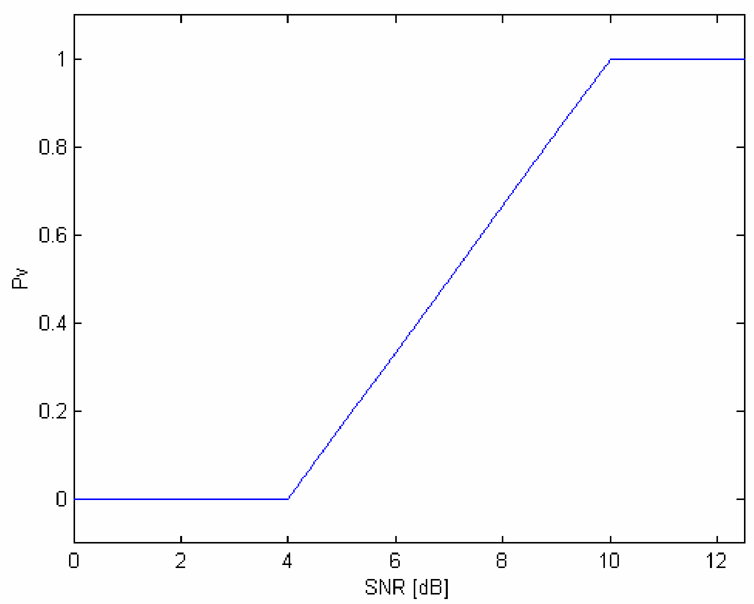
- Pitch doubling problem is solved.
- Pitch halving problem is solved.
- Enhancements:
  - Eliminate formant interaction problem.
  - Adaptive pitch resolution.

Voicing detection

$$SNR = \frac{\sum_{n=-(N-1)/2}^{(N-1)/2} \left| s[n] \right|^2}{\sum_{n=-(N-1)/2}^{(N-1)/2} \left| s[n] - \hat{s}(n;\omega_0) \right|^2}$$

$$P_v = \begin{cases} 1 & SNR > 10dB \\ \frac{1}{6}(SNR-4) & 4dB < SNR \le 10dB \\ 0 & SNR \le 4dB \end{cases}$$

P_v – *probability that speech is voiced*



Multi band pitch and voicing estimation

- MBE -Multiband Excitation Model,(Griffin & Lim, 1988).
- Speech model:

$$\hat{S}_\omega(\omega) = H_\omega(\omega)\left|E_\omega(\omega)\right|$$

$$H_\omega(\omega) - spectral\ envelope$$

$$\left|E_\omega(\omega)\right| - excitation\ spectrum$$

- Pitch period and Spectral envelope are calculated.
- The bandwidth is divided into ~20 subbands around the pitch harmonics.

Multi band pitch and voicing estimation

- The normalized MSE for subband m is defined by:

$$E_m\left(\hat{\omega}_0\right)=\frac{\int_{\gamma_m}\left|S_\omega\left(\omega\right)-\hat{S}_\omega\left(\omega;\hat{\omega}_0\right)\right|^2 d\omega}{\int_{\gamma_m}\left|S_\omega\left(\omega\right)\right|^2 d\omega}$$

$$\gamma_m=\left\{\omega:\omega_{m-1}\le\omega\le\omega_m\right\},\quad {}_{m=1,2,\ldots M}$$

$$\hat{S}_\omega\left(\omega;\omega_0\right)-Harmonic\ Model$$

$$if\ E_m\left(\hat{\omega}_0\right)<Threshold\Rightarrow Band\ is\ voiced$$

MBE example:

Original spectrum

Spectral envelope

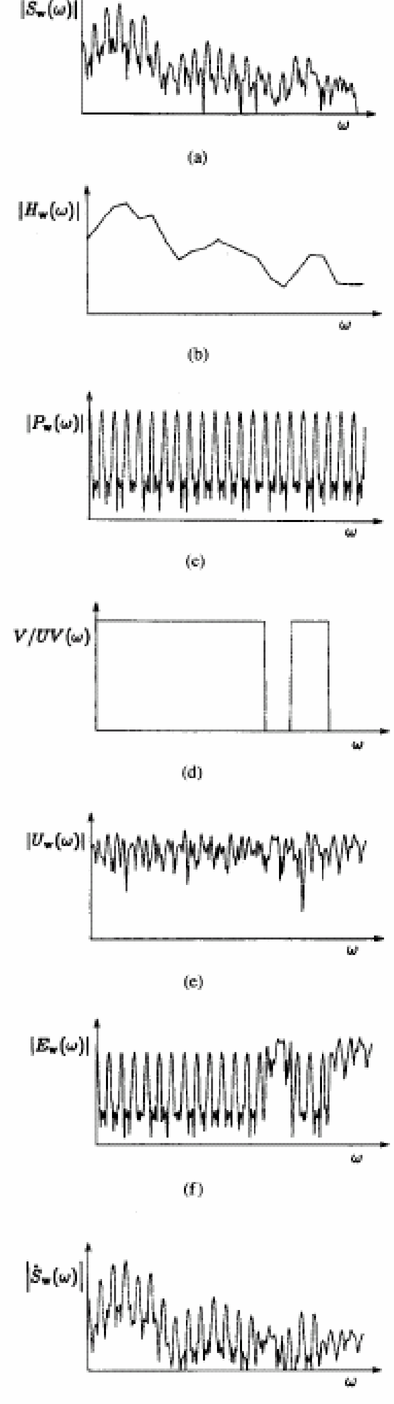
Periodic spectrum

V/UV information

Noise spectrum

Excitation spectrum

Synthetic spectrum



Advantages of the MBE model

- Quality for mixed V+UV speech.
- Robustness for additive acoustic noise.
- Lower bitrate for noiselike subbands.

Summary

- Pitch estimation can be calculated by fitting an harmonic sinewave model to the speech signal.
- The pitch estimate is robust to pitch doubling and pitch halving problems.
- The harmonic model naturally leads to a voicing detection based on the SNR.
- Implementation of voicing detection in sub bands was introduced.
- Pitch estimation under noise conditions remains an open problem.

Biliography

1. **Discrete-Time Speech Signal Processing** \ Quatieri, chapters 9,10
2. **A comparative Performance Study of Several Pitch Detection Algorithms**\ Rabiner, Cheng, IEEE tran. on acoustics, speech and sig. Proc., 1976
3. **Multiband Excitation Vocider**\Griffin, Lim, IEEE tran. on acoustics, speech and sig. Proc., 1988