Signal Processing

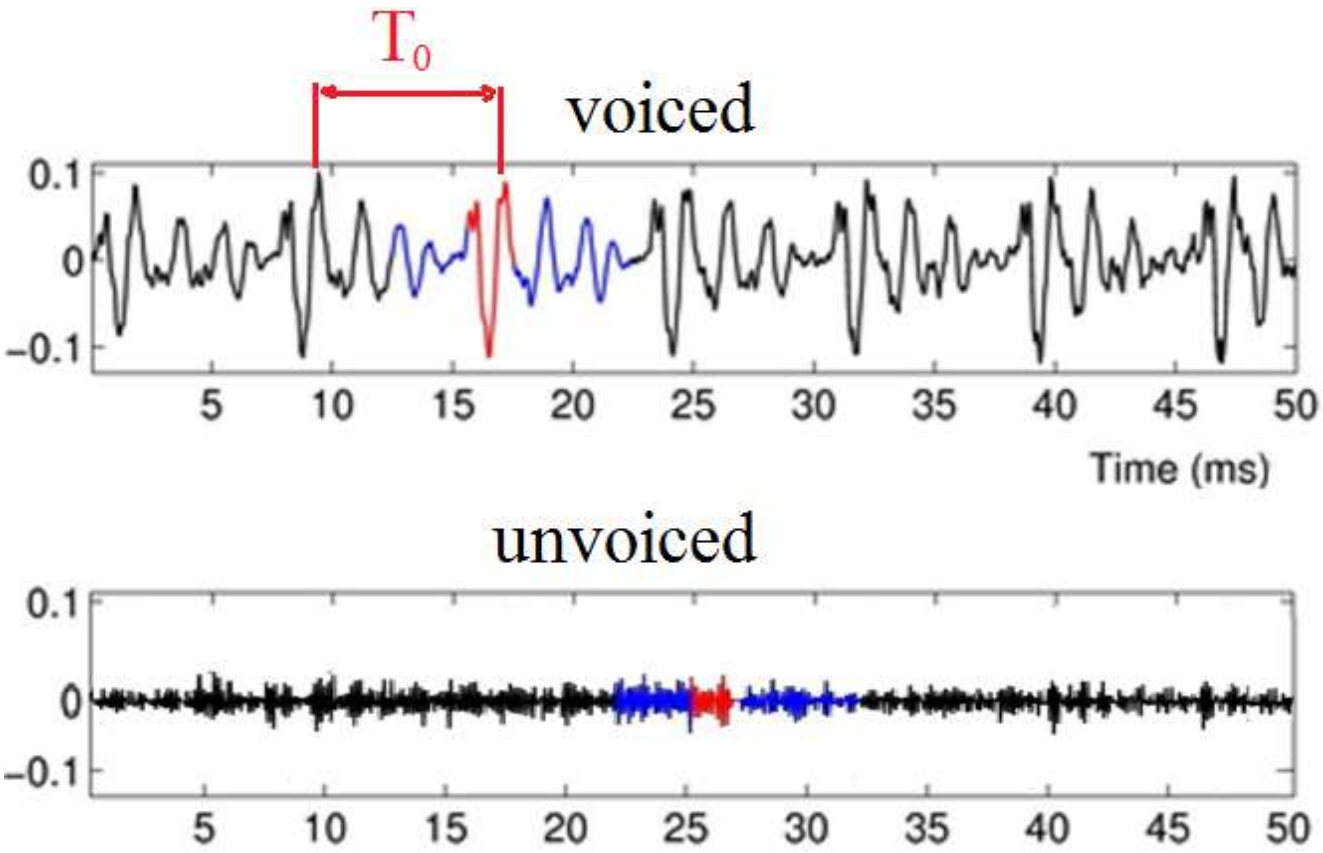
10.A Speech processing

Włodzimierz Kasprzak

2021

1. Pitch and voicing detection

- Pitch (F0) detection algorithms
- Voiced vs. unvoiced signal parts



Pitch detection algorithms (PDA)

Pitch (or the fundamental frequency F0 and the fundamental period T0) of a quasi-periodic signal can be measured in many ways.

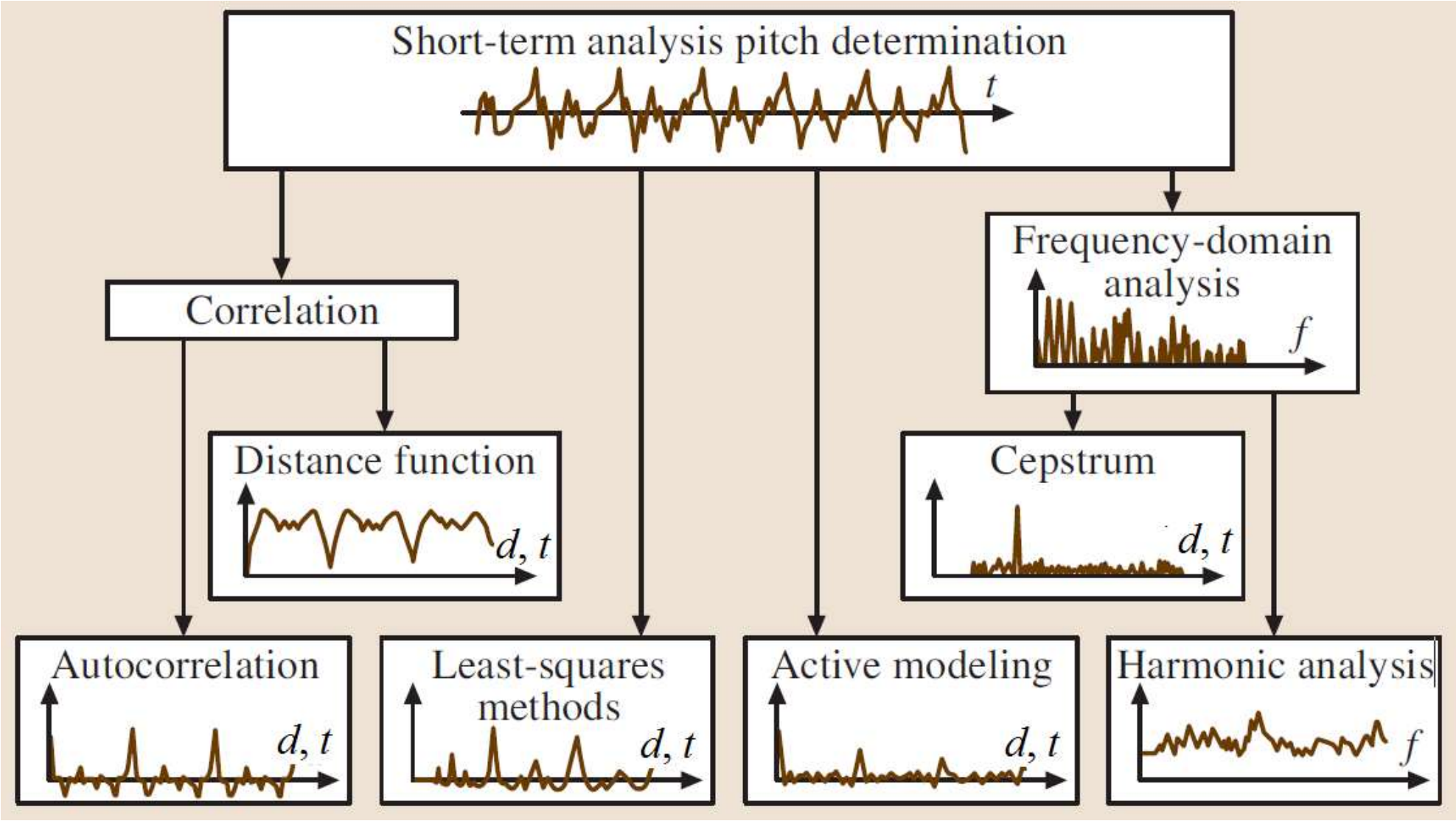
Since audio signals (e.g., speech, music) are both nonstationary and time-variant results may differ depending on method:

- A speech production-oriented PDA measures the rate of vocal fold vibration.
- Speech processing-oriented PDA: measures the periodicity of the signal (fundamental frequency or fundamental period)
- The perception-oriented PDA: measures the pitch – the frequency of a sinusoid that evokes the same intensity as the complex audio signal.

General classification of PDAs:

- Time-domain PDA;
- Frequency-domain PDA;

Pitch detection algorithms (PDA)



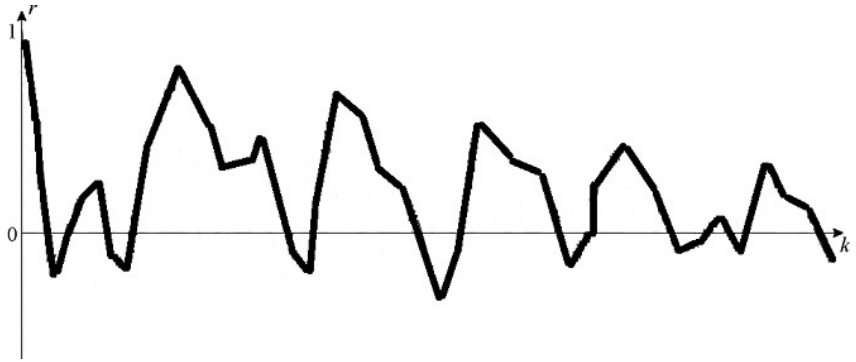
Benesty, Sondhi, Huang: Springer Handbook, Ch. 10]

Auto-correlation function of a signal frame

For a signal frame of size N, compute a **normalized auto-correlation** of signal samples with its shifted samples (by k).

Starting with m-th sample the k-shifted normalized auto-correlation is:

$$r_k^{(m)} = \frac{\sum_{n=m}^{m+N-k-1} f_n f_{n+k}}{|[f_n]||[f_{n+k}]|_n}$$



The maximum of normalized auto-correlation values always appears for $k=0$ and is equal to 1.

For a voiced speech part local maxima appear every k_0 samples:

r_k, r_{k+jk_0} ; e.g., for some k_0, $2k_0$, $3k_0$, etc., because many strong harmonic frequencies of F0 exist in this part.

Mutual correlation of frames

By **normalized mutual correlation** we mean a normalized correlation function of two consecutive signal frames.

Let the frames of signal samples are given (in vector form):

$$\boldsymbol{f}(m-k, m-1) = [f_{m-k}, \ldots, f_{m-1}]$$

$$\boldsymbol{f}(m, m+k-1) = [f_{m}, \ldots, f_{m+k-1}]$$

The normalized mutual correlation of these two vectors is:

$$\rho_m(k) = \frac{\boldsymbol{f}(m-k, m-1) \cdot \boldsymbol{f}(m, m+k-1)}{\| \boldsymbol{f}(m-k, m-1) \| \cdot \| \boldsymbol{f}(m, m+k-1) \|}$$

Start with k=2 and compute the correlation factor for different values of k. In a voiced part of the speech the maximum of normalized mutual correlation will be at k that corresponds to the base period (and base frequency F0) of the speaker.

Frequency-domain PDA

Principle - analysis of harmonic structure in the frequency domain:

- the power spectrum is compressed by a factor of two, three, etc. and then added to the original power spectrum;

→ this gives a peak at F0 resulting from the coherent additive contribution of the harmonic frequencies.

Example - the Martin's PDA:

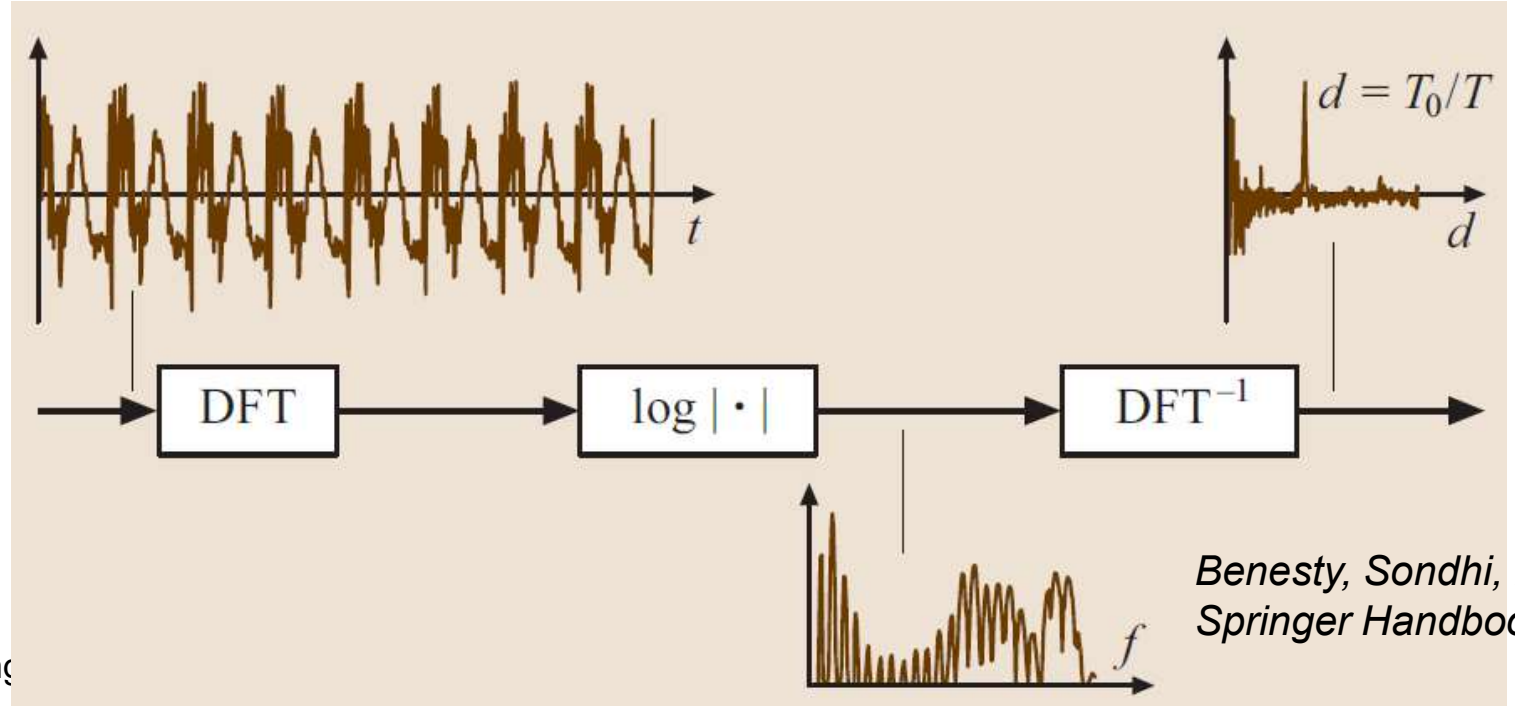
- The signal is decimated to 4kHz and transformed by a 128-point FFT.
- Only spectral local maxima with neighbours are retained – remaining points are set to zero.
- The spectrum is interpolated to increase the resolution of 1 Hz.
- The spectrum is multiplied by a spectral comb filter with parameter size p and the value p_0 is found for which a maximum of this product is reached.

Cepstrum PDA

Sensitivity against strong first formants, especially when they coincide with harmonics of F0 is a major problem in pitch detection.

It is known that the inverse Fourier transform of the power spectrum gives the autocorrelation function. Thus, a lag-domain representation shows a large peak at $d = T_0 / T_s$.

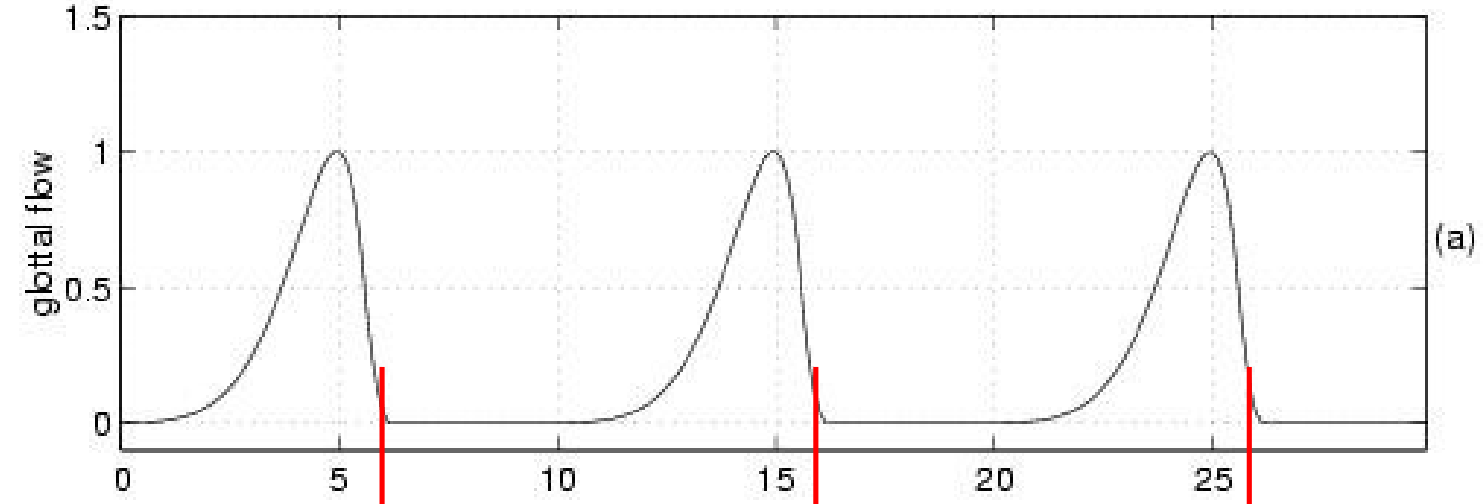
If we take the logarithm of the power spectrum (*spectral flatening*), we get the cepstrum PDA.



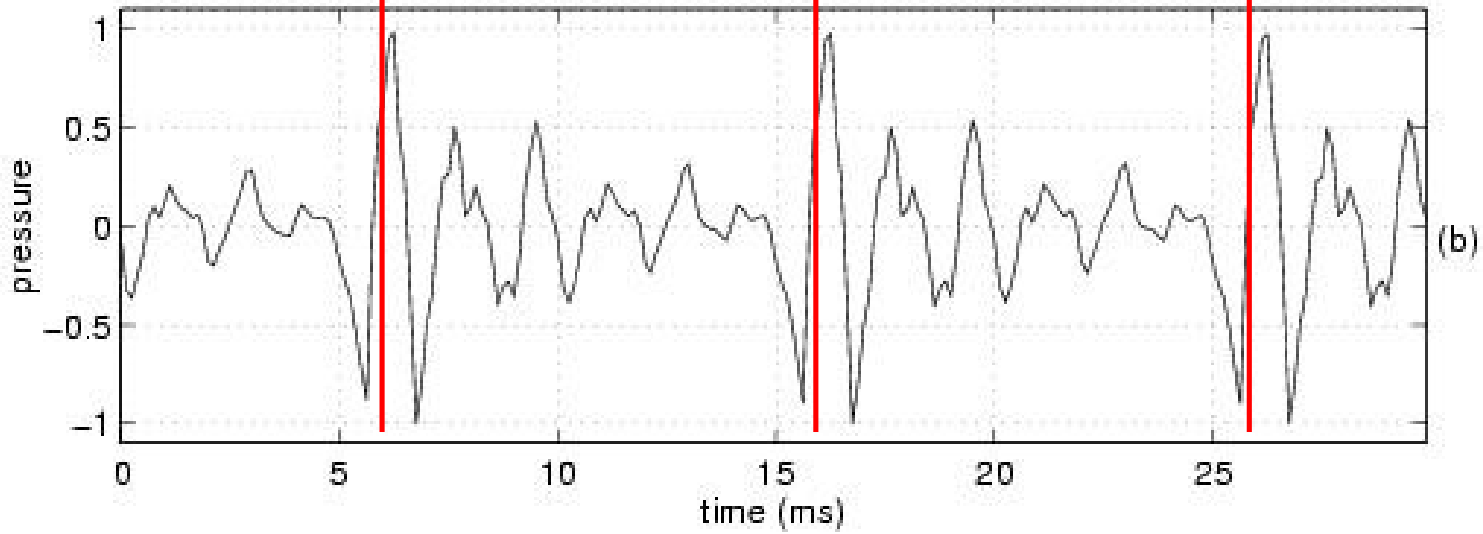
Benesty, Sondhi, Huang: Springer Handbook, Ch. 10]

Glottal Closure Instant (GCI)

Glottal volume velocity



Speech signal



Group delay CGI

Let $x[n]$ be a discrete-time signal and $X(\omega)$ its frequency-domain representation:

$$X(\omega) = \sum_n x(n)\, e^{-j\omega n} = |X(\omega)| e^{j\theta(\omega)}$$
$$= X_R(\omega) + jX_I(\omega)$$

Group delay is the derivative of the phase spectrum with respect to frequency:

$$\tau(\omega) = -\frac{d\theta(\omega)}{d\omega}$$

Computation of group delay in discrete-time case :

$$X(m) = DFT\{x(n)\}\ ; \quad \tilde{X}(m) = DFT\{n\, x(n)\}$$

$$\tau_G(m) = Re\left(\frac{\tilde{X}(m)}{X(m)}\right)$$

Computation of Group-delay

$$\tau(\omega) = -\frac{d\theta(\omega)}{d\omega} = -\mathrm{Im}\left(\frac{d(\ln X(\omega))}{d\omega}\right) = -\mathrm{Im}\left(\frac{-j\sum_n nx(n)e^{-j\omega n}}{X(\omega)}\right)$$

$$= \mathrm{Re}\left(\frac{Y(\omega)}{X(\omega)}\right)$$

$$\text{where,} \quad Y(\omega) = \sum_n nx(n)e^{-j\omega n} = Y_R(\omega) + jY_I(\omega)$$

Hence,

$$\tau(\omega) = \mathrm{Re}\left(\frac{[Y_R(\omega) + jY_I(\omega)][X_R(\omega) - jX_I(\omega)]}{[X_R(\omega) + jX_I(\omega)][X_R(\omega) - jX_I(\omega)]}\right)$$

$$= \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{X_R^2(\omega) + X_I^2(\omega)}$$

Group delay CGI

If a frame contains a single impulse at position $n_0 = k$, this will produce a fixed group delay index k for all frequency indexes m.

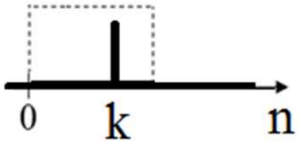
If there is additional noise in the frame, one can expect that at least the average group delay (averaged over all frequencies) equals k.

If we now compute the average group delay for each signal frame t starting with each sample of the analysed signal it will show a negative-going zero crossing (decreasing with a slope of -1) for $t=k$.

This method of CGI detection is often applied to a transformed signal, e.g., a LP residual signal, in which CGIs show up as impulse-like structures.

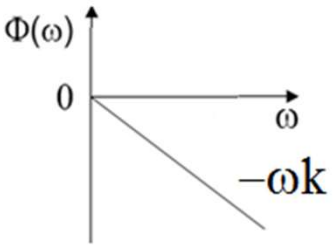
Principle of group delay CGI

a single impulse



$s[n] = \delta[n-k]$

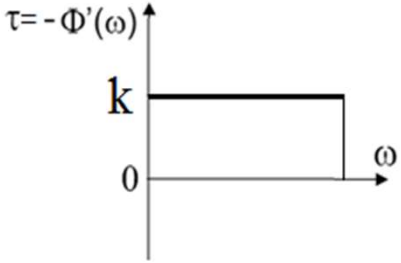
$S[m] = DFT(s[n]) = e^{-j\omega k}$



Phase(S[m]):

$\Phi(\omega) = -\omega k$

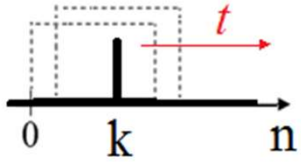
Group delay: $-\Phi'(\omega) = k$



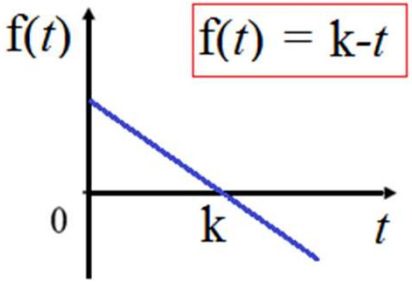
Average group delay:

$\overline{-\Phi'(\omega)} = k$

If the frame is moved to the right



relative k(*t*) decreases



$f(t) = k-t$

the average group delay decreases with a slope of (-1)

Multiple pitch determination

Multiple pitch appears:

- In music – where various tones (pitches) may coexist,
- In multi-speaker signals.

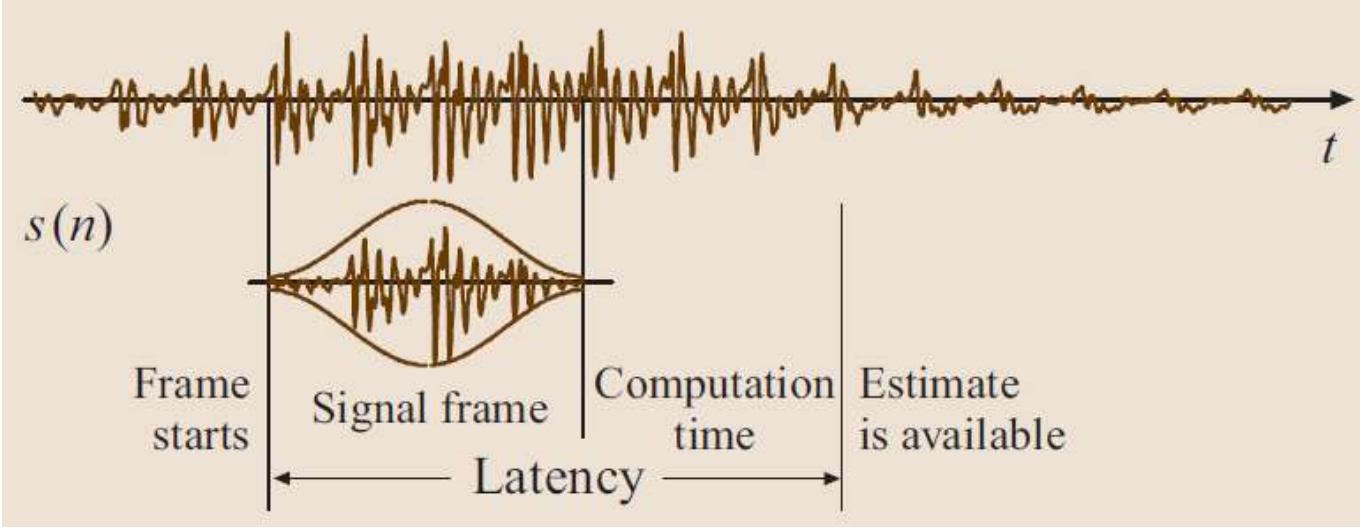
A standard procedure is:

- detect first pitch,
- signal enhancement (eg. spectral subtraction of the first source from the signal),
- detect next pitch in the enhanced signal,
- etc.

Instantaneousness vs. reliability

It is desirable to obtain processing results instantaneously. But a delay of results (**latency**) cannot be avoided. Latency depends on:

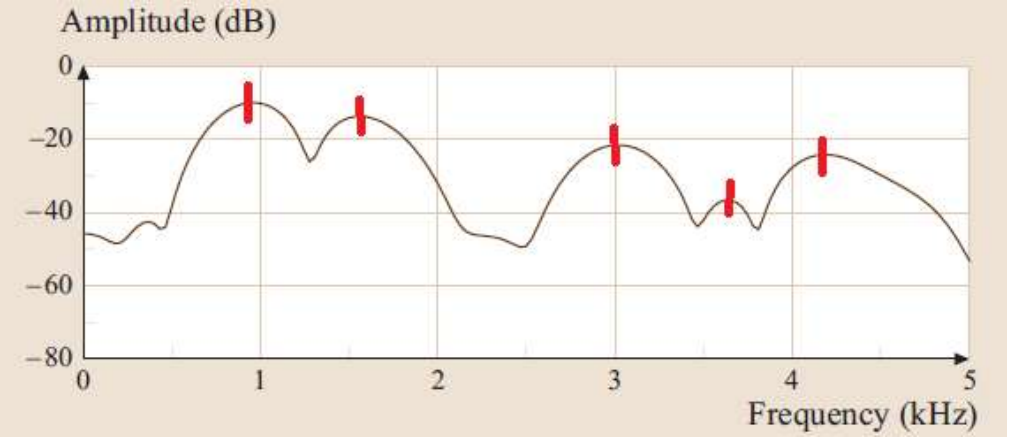
1. Time of computing the method – induced by computational complexity of the method and processing speed of the computing device;
2. Data collecting time, i.e., the amount of time needed to get the required number of signal samples. For example, a standard PDA needs at least two complete pitch periods of the signal.



Benesty, Sondhi, Huang: Springer Handbook, Ch. 10]

2. Formant estimation

Formants (energy maxima in frequency bands): **F1-F5**



Examples

F1 can vary from 300 Hz to 1000 Hz

The lower it is, the closer the tongue is to the roof of the mouth.

/i:/ has the lowest F1: ~ 300 Hz; /A/ has the highest F1 ~ 950 Hz.

F2 can vary from 850 Hz to 2500 Hz

The F2 value is proportional to the front or back position of the tongue tip. In addition, lip rounding causes a lower F2 than the case with unrounded lips.

/i:/ has an F2 of 2200 Hz, the highest F2 of any vowel; /u/ has an F2 of 850 Hz - the tongue tip is far back, and the lips are rounded.

Formant detection

Typical method of formant detection:

Searching for peaks in spectral representation, resulting from:

- STFT or
- STFT(LPC) (induces smoothed spectrum that simplifies peak search) or
- a group-delay spectrum

In LPC-based methods 2 poles (LPC parameters) per kHz of bandwith + 2-4 poles to model the spectral tilt effect. The optimal case is when the pole number matches the number of resonances in the signal.

In purely STFT methods, an initial smoothing of the spectrum is also applied by selecting a smaller number of samples than the expected pitch period and padding the remaing samples with zeros.

LPC-based spectrum

LPC parameters (explained in p. 11.5)

a_i for $i = 0, ..., m$; where $a_0 = 1$.

The number of prediction parameters in speech signal analysis is: $m \in \langle 12, 20 \rangle$.

I.e., for the sampling frequency f_s [expressed in kHz]: $m \approx f_s + [2, 4]$.

Spectral features (smoothed signal spectrum)

By transforming the parameter vector into frequency domain, we get a **smooth spectrum** of the signal frame. The required resolution in frequency domain is achieved by padding the parameter vector with zeros to get a vector with M elements:

$$\mathbf{A}_M = DFT([1, a_1, a_2, ..., a_m, 0, 0, ..., 0])$$

Group-delay spectrum

$$\tau(\omega) = -\frac{d\theta(\omega)}{d\omega} = -\mathrm{Im}\left(\frac{d(\log X(\omega))}{d\omega}\right) = -\mathrm{Im}\left(\frac{-j\sum_n nx(n)e^{-j\omega n}}{X(\omega)}\right)$$

$$= \mathrm{Re}\left(\frac{Y(\omega)}{X(\omega)}\right)$$

$$\text{where,} \quad Y(\omega) = \sum_n nx(n)e^{-j\omega n} = Y_R(\omega) + jY_I(\omega)$$

Hence,

$$\tau(\omega) = \mathrm{Re}\left(\frac{[Y_R(\omega) + jY_I(\omega)][X_R(\omega) - X_I(\omega)]}{X_R^2(\omega) + X_I^2(\omega)}\right)$$

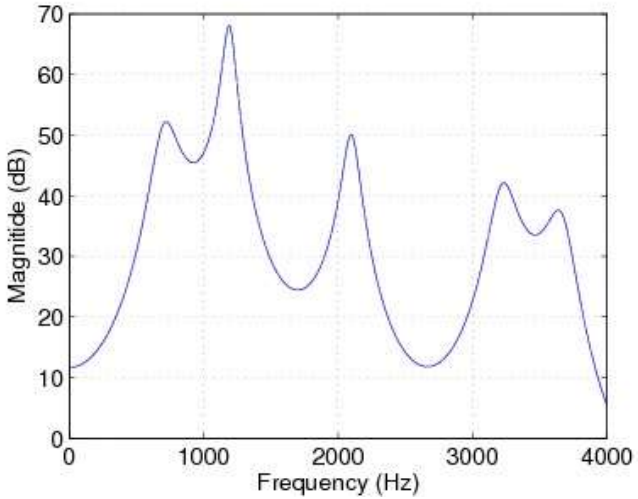
$$= \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{X_R^2(\omega) + X_I^2(\omega)}$$

Idea: use the non-normalized value of group delay (the enumerator of above fraction).

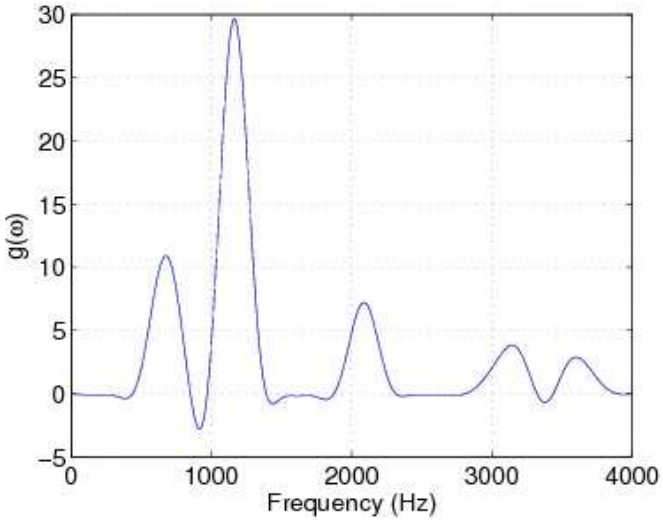
Obtain the spectrum of the group-delay's enumerator.

Group-delay spectrum

Log-magnitude spectrum (for a 5ms frame):



Group-delay spectrum:



Non-normalized – the numerator of group-delay spectrum: