BE2M31ZRE - Speech processing

Spectral characteristics of speech signal

Doc. Ing. Petr Pollák, CSc.

February 20, 2022 - 23:19

Content

- **Time-domain and frequency-domain representation of speech signals**
  - Computation of DFT for speech
    (FFT, weighting, frequency resolution)
  - Filter banks
  - Preemphasis
- **Linear Predictive Analysis - AR modelling**
  - Basic principles of LPC, LPC spectrum, AR model
  - Algorithms of computation (Levinson-Durbin, Burg)
- **Formant Analysis**
  - Formant definition and meaning
  - Methods of formant estimation

Part I

DFT-based spectral characteristics

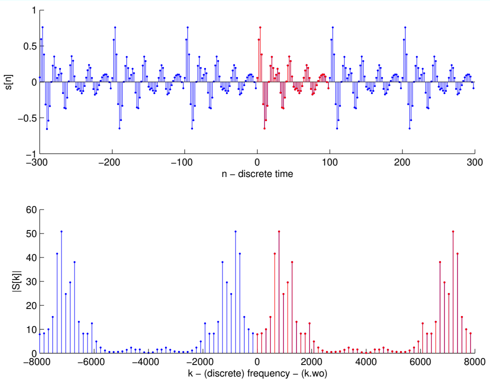
Discrete Fourier Transform (DFT) - basic properties

Forward transform - DFT

$$S[k] = \sum_{n=0}^{N-1} s[n] e^{-j\frac{2\pi}{N}kn}$$

Inverse transform - IDFT

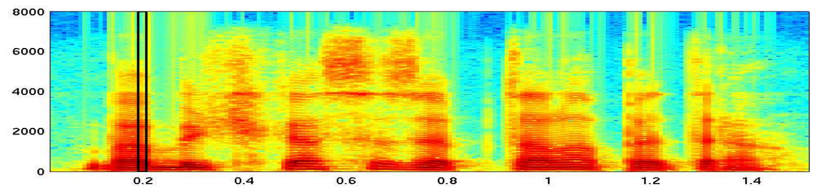
$$s[n] = \frac{1}{N} \sum_{k=0}^{N-1} S[k] e^{j\frac{2\pi}{N}kn}$$



- **Transform** : finite-length discrete signal $\rightarrow$
  $\rightarrow$ finite-length discrete spectrum
- **Frequency resolution** : N spectral samples
  $\rightarrow$ frequency range $0 \div f_s$ or $-\frac{f_s}{2} \div \frac{f_s}{2}$ resp. $\rightarrow \Delta_f = \frac{f_s}{N}$
- **FFT**: the algorithm for efficient and fast computation ($N = 2^n$!!!)

Various spectral representations of an utterance

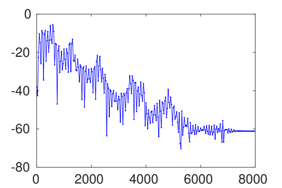
Spectrogram of an utterance



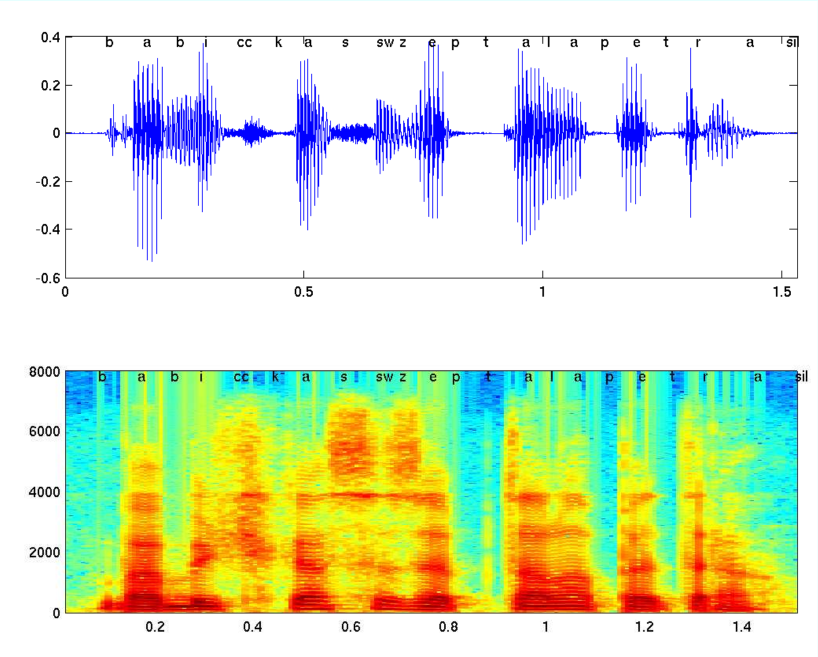
!!! Long-time (averaged) spectrum - meaning-less representation !!!

Short-time spectral representations (for selected S-T frame)

DFT spectrum:



Time- and frequency-domain representation of an utterance



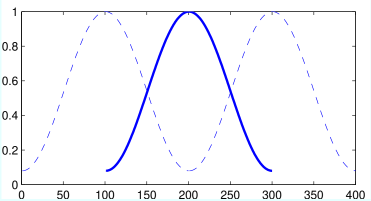
DFT based spectral analysis of speech signal - SUMMARY

Specific settings for speech analysis:

- **speech is non-stationary**
  $\Rightarrow$ processing in short-time frames is necessary (spectrogram)
- **speech is quasi-stationary**
  (i.e. stationary in short-time sense - approx 10-100 ms)
  $\Rightarrow$ 20-30 ms - length of short-time processing frame
- **DFT spectrum - standardly affected by spectral leakage**
  $\Rightarrow$ using of weighting window is necessary (Hamming)
  $\Rightarrow$ segmentation with overlapping is necessary (usually 50%)

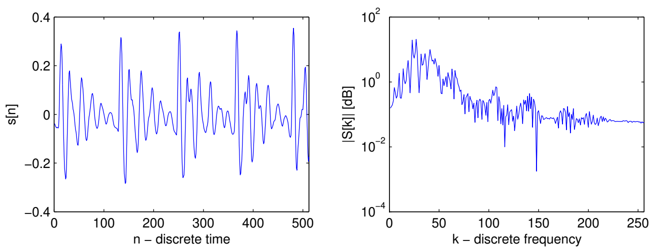
$$w[n] = 0,54 - 0,46 \cos \frac{2\pi n}{N}$$

pro $0 \leq n \leq N-1$.

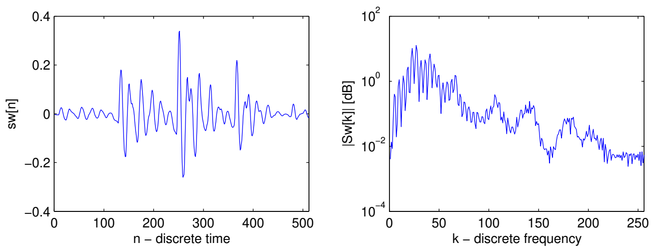


Spectral leakage in short-time spectrum of speech

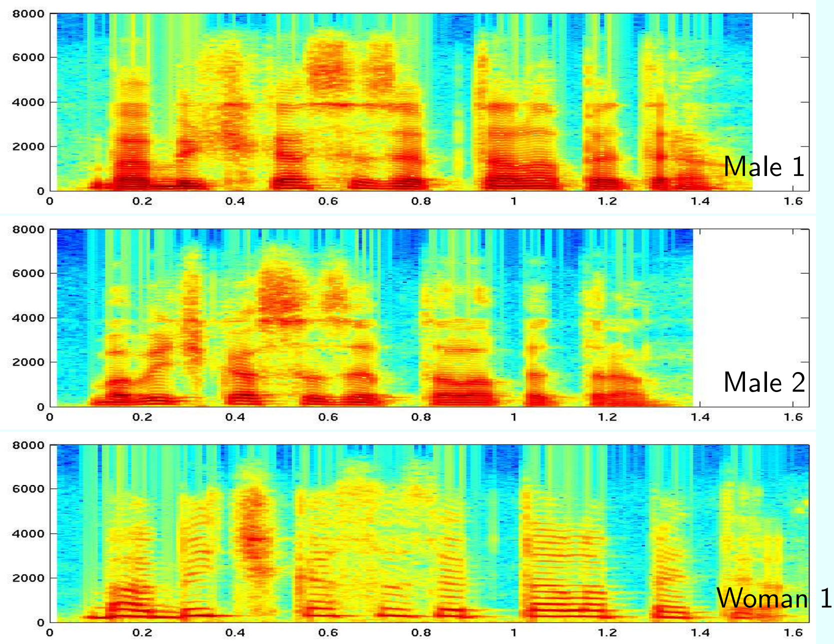
Spectrum of unweighted frame - spectral leackage
(masking of low-level details in HF band)



Spectrum of weighted frame - **spectral leackage is minimized**
(low-level details in HF band are visible)



Variability of utterance with same contents - influence on f_o



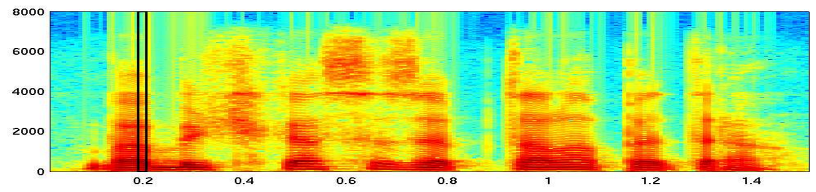
Properties of short-time DFT spectrum of speech

- particular phones can be resolved
- stochastic component is presented
- information about periodicity (f_o) is presented
- for typical values of f_s rather high number of spectral samples (redundant information)

$$\Downarrow$$

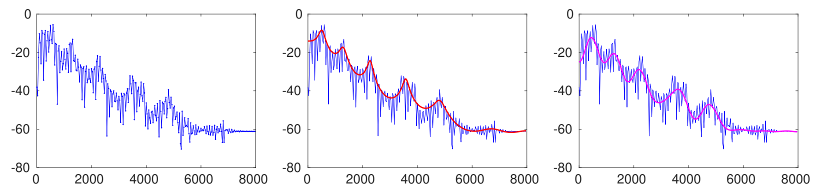
- **Smooted spectral characteristics - more suitable choice**
  - filter-banks (non-linear frequency scale)
  - LPC
  - cepstral analysis

Spectrogram of whole utterance



Short-time spectral representations (for selected S-T frame)

DFT spectrum: LPC spectrum: Cepstrum:



256 spectral samples (magnitude sp.)

$\approx$ 16 coefficients a_k (autoregressive coef.)

$\approx$ 20 coefficients c_n (real cepstrum)

Spectral analysis using filter banks

Main purpose $\rightarrow$ computation of power (energy) in given freq. bands

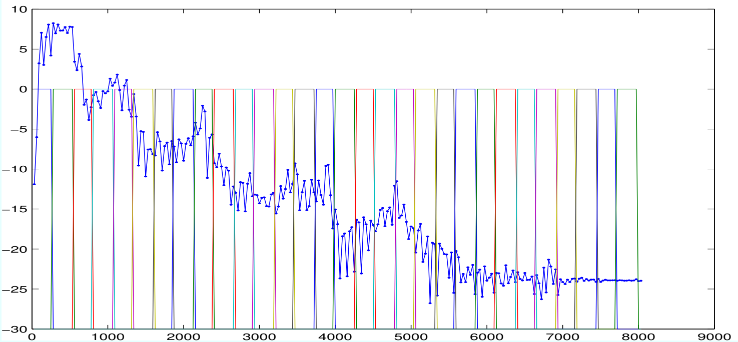
FB is typically based on DFT

$\Rightarrow$ filter are described by weights of particular DFT-bins for given frequency resolution (NDFT) and f_s

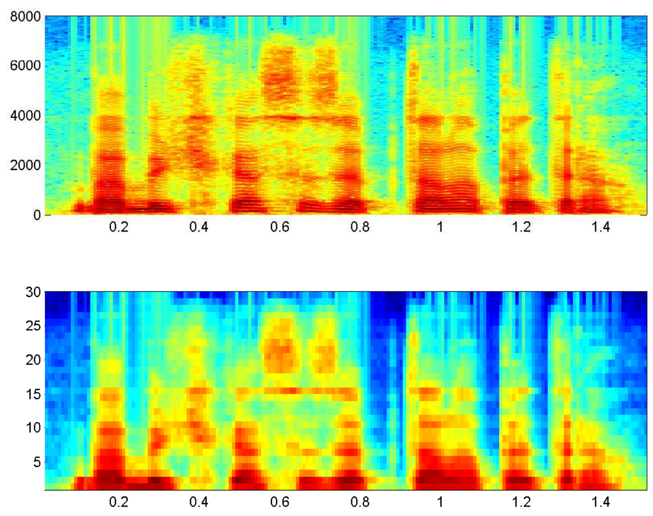
$$G_{mel}[j] = \sum_{k=0}^{N/2} |S[k]|^2 H_j[k] \quad \text{for j = 1, ..., M}$$

<u>M - number of bands</u>

- according to f_s, NDFT and type of FB



Spectral analysis using filter banks



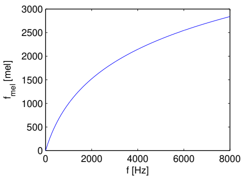
Linear scale - DISADVANTAGE - rough resolution in low-frequncy band and too detailed resolution in high-frequency band
(not related to the perception of frequency)

Filter bank with non-linear mel-frequency axis

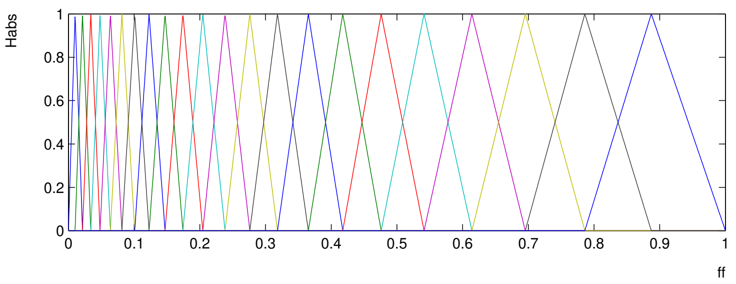
Non-linear frequency warping - *melodic scale*

$$f_{mel} = \mathrm{Mel}\,(f) = 2595\log_{10}\left(1 + \frac{f}{700}\right)$$

$$f = \mathrm{InvMel}\,(f_{mel}) = 700 \cdot \left(10^{\frac{f_{mel}}{2595}} - 1\right)$$



Triangular mel-scale filter bank (used for MFCC computation)



Realization of filter banks

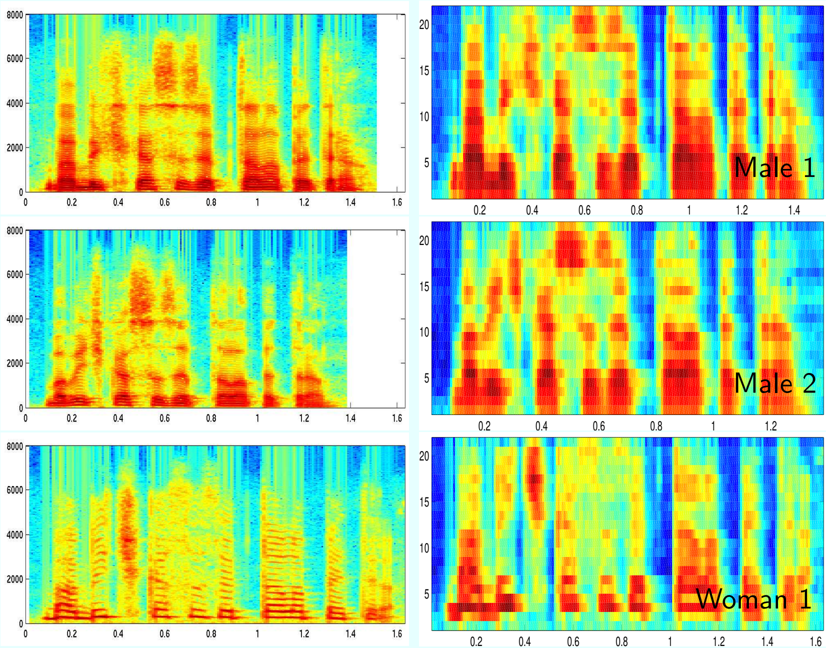
BF is again realized on the basis of DFT

$\Rightarrow$ particular filters - weights for given NDFT and f_s
$\Rightarrow$ computation principle - same for various FB
$\Rightarrow$ other FB = just other weigths

$$G_{mel}[j] = \sum_{k=0}^{N/2} |S[k]|^2 H_{mel,j}[k] \quad \text{for j = 1, ..., M}$$

<u>M - number of bands</u> ..... typical value 20-30 bands
- according to f_s and NDFT
- 22 for $f_s = 8$ kHz and frame length of 25 ms
- 30 for $f_s = 16$ kHz and frame length of 25 ms

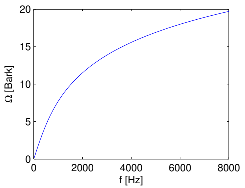
Variability of the utterance within mel-based spectrogram



Filter bank with Bark frequency axis

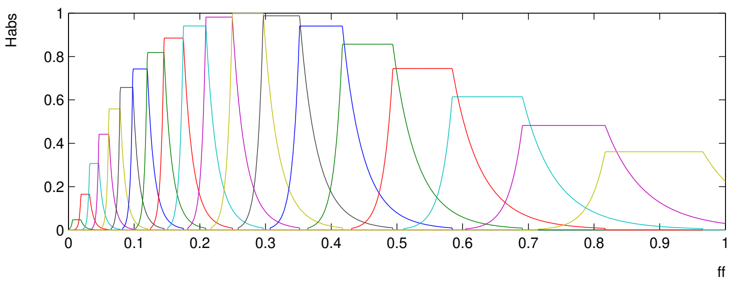
Bark scale - defined on the basis of critical bands

$$\Omega = \text{Bark}(f) = 6 \ln \left(\frac{f}{600} + \sqrt{\left(\frac{f}{600}\right)^2 + 1} \right)$$



$$f = \text{InvBark}(\Omega) = 600 \cdot \sinh \frac{\Omega}{6}$$

Trapezoidal Bark-scale filter bank (used for PLPC computation)
(contains equal-loudness curves and application of intensity law)



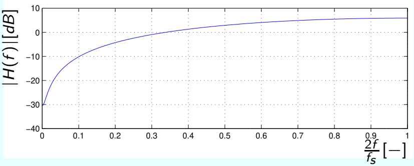
FB is realized again on the basis of DFT (for given NDFT a f_s)

Preemphasis - compensation of HF-spectrum attenuation

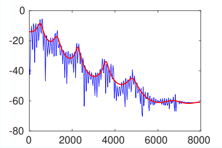
Downslope of magnitude spectrum - high frequencies - lower energy

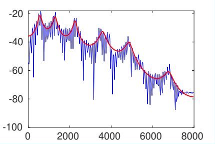
Preemphasis filter (1st-ord FIR):

$$s'[n] = s[n] - m \cdot s[n-1]$$
$$(m = 0.97)$$

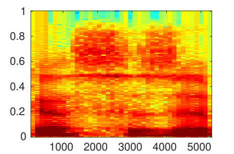


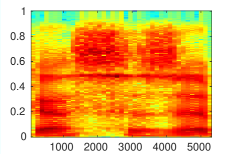
Impact of preemphasis in short-time spectrum (DFT and LPC)





Impact of preemphasis in spectrogram

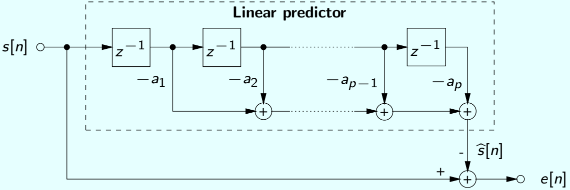




Part II

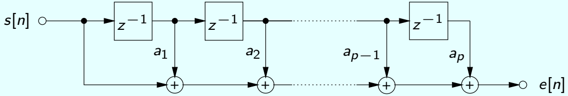
LPC, AR modelling

Linear prediction : $\widehat{s}[n] = -\sum_{k=1}^{p} a_k s[n-k]$.



Error signal (measure of predictor quality)

$$e[n] = s[n] - \widehat{s}[n] = s[n] + \sum_{k=1}^{p} a_k s[n-k] = \sum_{k=0}^{p} a_k s[n-k] \; .$$



Principles of LPC analysis

IDEA: more precise prediction $\rightarrow$ lower level of error signal

Criterion - power of error signal

$$J = E\left\{ e^2[n] \right\}$$

Looking for coefficients a_k $\equiv$ Minimizing of prediction error
$\equiv$ looking for minimum of J, i.e.

$$\frac{\partial J}{\partial a_k} = 0 \ , \quad \text{for} \quad k \ = 1, 2, ..., p \quad \Rightarrow \quad p \text{ linear equations}$$

Solutions and computational procedures
(for varying definitions of J):

- **autocorrelation method** - the most frequent approach (Yule-Walker)
- Levinson-Durbin alg. (fast computation of Yule-Walker eqs)
- Burg algorithm - originates from lattice structure of FIR filter

Autocorrelation method, Yule-Walker equations

$$\begin{bmatrix} R[0] & R[1] & R[2] & \dots & R[p-1] \\ R[1] & R[0] & R[1] & & R[p-2] \\ R[2] & R[1] & R[0] & \ddots & R[p-3] \\ \vdots & & \ddots & \ddots & \vdots \\ R[p-1] & R[p-2] & R[p-3] & \dots & R[0] \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} R[1] \\ R[2] \\ \vdots \\ \vdots \\ R[p] \end{bmatrix}$$

$R[k]$ .... autocorrelation coefficients of analyzed signal

<u>RESULT:</u>

a_k .... autoregressive coefficients (AR model)

$P_p = R[0] + \sum_{k=1}^{p} a_k R[k]$ .... power of error signal

AR model of signal

Decorrelating (analyzing) filter : $A(z) = \sum_{k=0}^{p} a_k z^{-k}$

$$s[n] \longrightarrow \boxed{A(z)} \longrightarrow e[n]$$

Synthesis with real error signal (ideal case)

$$e[n] \longrightarrow \boxed{\frac{1}{A(z)}} \longrightarrow s[n]$$

Synthesis with artifficial signal with unit power (AR model)
- G is related to the power of prediction error ($G = \sqrt{P_p}$)

$$u[n] \longrightarrow \boxed{\frac{G}{A(z)}} \longrightarrow \tilde{s}[n]$$

$$H(z) = \frac{G}{A(z)} = \frac{G}{1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_p z^{-p}}$$

Spectral representations of AR model

General description of AR model (AR synthesis) in Z-domain

$$\tilde{S}(z) = H(z) \cdot U(z)$$

Description of AR model in frequency domain

$$S_{\tilde{s}}(e^{j\Theta}) = |H(e^{j\Theta})|^2 \cdot S_u(e^{j\Theta})$$

Properties and consequences: - $S_u(e^{j\Theta})$ is flat
$\rightarrow$ shape of $S_{\tilde{s}}(e^{j\Theta})$ is completely described by AR model

$$\Downarrow$$

LPC spectrum (if $S_u(e^{j\Theta}) = 1$) $\quad S_{\tilde{S}}(e^{j\Theta}) = |H(e^{j\Theta})|^2$

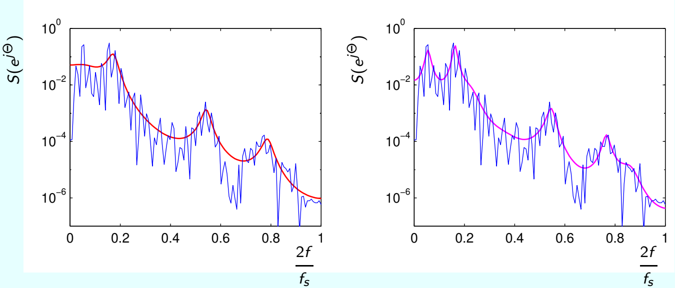
$$S_{\tilde{S}}(e^{j\Theta}) = \frac{G^2}{|A(e^{j\Theta})|^2} = \frac{G^2}{|1 + a_1 e^{-j\Theta} + a_2 e^{-j2\Theta} + ... + a_p e^{-jp\Theta}|^2}$$

$$\Downarrow$$

coefficients a_k .... compressed spectral reprezentation

Comparison of LPC and DFT spectra

$$S_{\tilde{S}}(e^{j\Theta}) = |H(e^{j\Theta})|^2 \quad \approx \quad \frac{|S[k]|^2}{N}$$



- AR model = "all-pole" filter, peak modelling (resonators of vocal tract)
- general peak = a couple of complex conjugated poles
- real pole models a peak at 0 or $f_s/2$
- higher order of AR model = more peaks in LPC spectrum
  $\rightarrow$ typical values: $p = 10$ for $f_s = 8$ kHz, $p = 16$ for $f_s = 16$ kHz

Computation of LPC in MATLAB

1. **Computation of AR model parameters**

   Function **lpc**

   $$[\mathrm{a},\ \mathrm{Ep}] = \mathrm{lpc}\ (\ \mathrm{s},\ \mathrm{p}\ )\ ;$$

   - a . . . autoregressive coefficients (includinng $a_0 = 1$)
   - Ep . . . power of prediction error
   - s . . . analyzed signal
   - p . . . order of AR model

2. **Computation of LPC spectrum**

   Function **freqz**

   $$\mathrm{H} = \mathrm{freqz}\ (\ \mathrm{sqrt(Ep)},\ \mathrm{a},\ \mathrm{N}\ )\ ;$$

   - H . . . complex LPC spectrum
   - N . . . number of points of LPC spectrum

Levinson-Durbin algorithm

Fast and **recurent** computation of coefficients a_k defined by autocorrelation method (fast solution of Yule-Walker equations)

Inicialization:

$$P_0 = R[0] \quad a_1^{(1)} = k_1 = -\frac{R[1]}{R[0]}$$

$$P_1 = P_0 \cdot (1 - k_1^2)$$

Steps for $m = 2, 3, \ldots, p$:

$$a_m^{(m)} = k_m = -\frac{R[m] + \sum_{j=1}^{m-1} a_j^{(m-1)} R[m-j]}{P_{m-1}}$$

$$a_j^{(m)} = a_j^{(m-1)} + k_m a_{m-j}^{(m-1)}, \quad j = 1, 2, \ldots, m-1$$

$$P_m = P_{m-1} \cdot (1 - k_m^2)$$

Result:

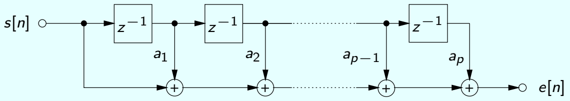
$$a_i = a_i^{(p)}, \quad i = 1, 2, \ldots, p$$

k_k .... reflection coefficients (lattice structure of the filter)

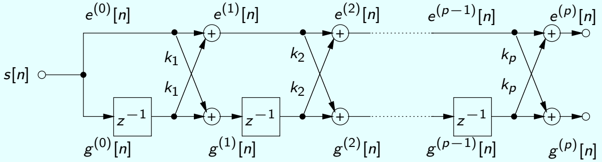
.... PARCOR coefficients (partial correlation coeff.)

AR model - standard and lattice structure

Trasnversal structure of analyzing FIR filter:



Lattice structure of analyzing FIR filter:



k_k ..... reflection coefficients, relationship k_k vs. a_k - Levinson recursion

Inicialization:

$$a_1^{(1)} = k_1$$

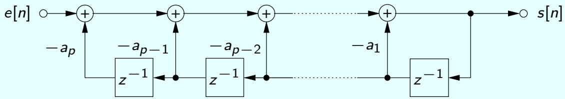
Computation for $m = 2, 3, \ldots, p$**:**

$$a_m^{(m)} = k_m$$

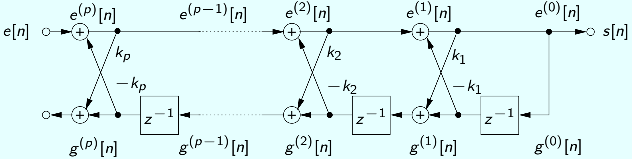
$$a_j^{(m)} = a_j^{(m-1)} + k_m a_{m-j}^{(m-1)}, \quad j = 1, 2, \ldots, m-1$$

AR model with lattice structure

Trasnversal structure of synthezing all-pole IIR filter:



Lattice structure of synthezing all-pole IIR filter:



Properties of reflection coefficients k_k :

- stable synthezing filter for $-1 < k_k < 1$
- more robust than a_k for low variability of signal ($\rightarrow$ suitable features)
- suitable for implementation (less problems due to quantization)
- possible interpolation
- **direct computation of reflection coeff.** possible $\rightarrow$ Burg algorithm

Burg algorithm

Criterion to be minimized (for each section of lattice structure):

$$J_m = \frac{1}{2} \sum_{n=0}^{N-1} \left[\left(e^{(m)}[n] \right)^2 + \left(g^{(m)}[n] \right)^2 \right] \qquad \text{for } m = 1,\, 2,\, ...,\, p \,.$$

Inicialization: $e^{(0)}[n] = g^{(0)}[n] = s[n]$

Computation for $m = 1, 2,\, 3,\, \ldots,\, p$**:**

$$k_m = -\frac{2 \cdot \sum_{n=m}^{N-1} \left(e^{(m-1)}[n] \cdot g^{(m-1)}[n-1] \right)}{\sum_{n=m}^{N-1} \left(e^{(m-1)}[n] \right)^2 + \sum_{n=m}^{N-1} \left(g^{(m-1)}[n-1] \right)^2}$$

Always fulfilled $|k_m| < 1$ $\longrightarrow$ always stable solution

$$e^{(m)}[n] = e^{(m-1)}[n] + k_m \cdot g^{(m-1)}[n-1], \quad n = 0,\, 1,\, \ldots,\, N-m$$

$$g^{(m)}[n] = g^{(m-1)}[n-1] + k_m \cdot e^{(m-1)}[n], \quad n = 0,\, 1,\, \ldots,\, N-m$$

Further computations: - autoregr. coeff a_k - Lev. rec., see L.-D. alg.
- power of prediction error P_k - see L.-D. alg.

Part III

Formants and their Estimation

Formants - definition

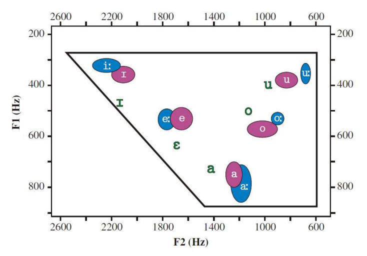
- **Formant (formant frequency)**
  $\rightarrow$ central frequency of vocal tract rezonator

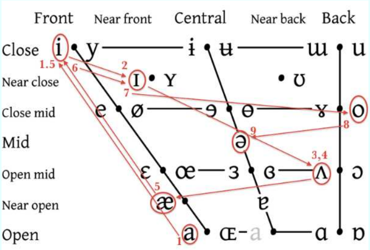
- significant peaks in **SMOOTHED short-time spectrum**
- significant formants are F1 - F4, i.e. in the band upto 4 kHz
- F5 - negligible (also higher estimation error)
- !! Do not confuse formant vs. pitch f_0 !!
  (f_0 is not identified in smoothed spectrum)

- **Applications:**
  - elementar speech analysis
  - formant speech synthesis
  - transformations of voice characteristics (Lombard effect)

Formants of vowels (formant triangle)

	I	E	A	O	U
F1	300 - 500	480 - 700	700 - 1100	500 - 700	300 - 500
F2	2000 - 2800	1560 - 2100	1100 - 1500	850 - 1200	600 - 1000
F3	2600 - 3500	2500 - 3000	2500 - 3000	2500 - 3000	2400 - 2900



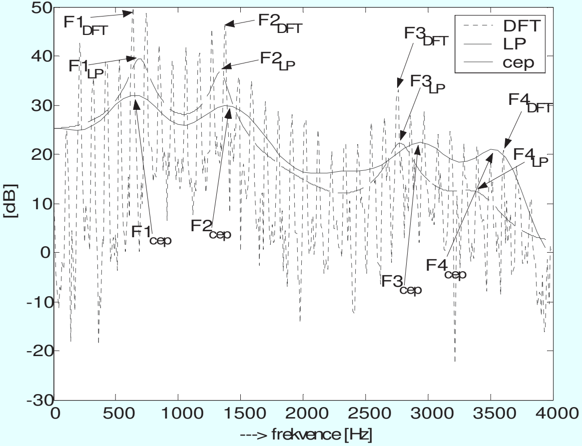


Formants - estimation techniques

- **from smoothed DFT spectrum**
  - short window, zero-padding, looking for maxima
  - not too precise
- **Using LPC**
  - LPC analysis - smoothed spectrum
  - peaks in LPC spectrum - rezonators of vocal tract
  - the most frequently used technique
- **Using cepstral analysis**
  - estimation of smoothed spectrum using cepstral liftering
  - looking for the maxima

Formants - short-time spectrum

Phone 'a' - formants in smoothed and non-smoothed spectrum



Formants - LPC-based estimation

- peaks in LPC spectrum - rezonators = formants
- peaks are determined by **poles p_i of transfer function** $H(z)$

$$F_i = \frac{\arg\ p_i}{2\pi} \cdot f_s$$

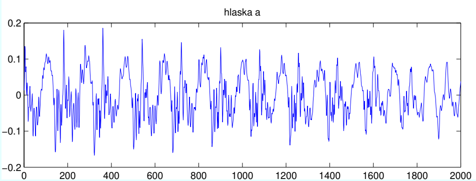
$$B_i = -\frac{\ln |p_i|}{\pi} \cdot f_s$$

- F_i - formant frequency (central freuency of rezonator)
- B_i - band-width of formant (rezonator)

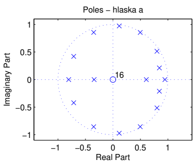
- Problems:
  - generally lower robustness of LPC analysis (data dependency)
  - sensitivity to choice of AR model order (for noisy conditions)
  - removing of redundant poles (negligible peaks)
  - sorting of computed poles (tracking of particular formants)

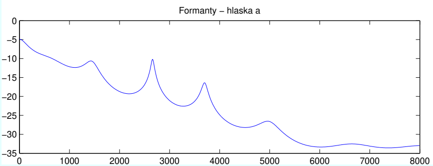
LPC-based formant estimation - example

Waveform of signal

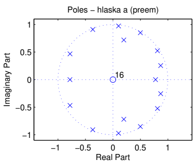


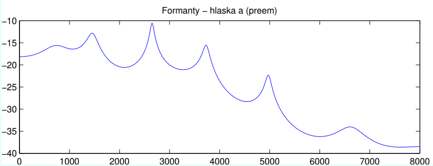
poles & LPC spectrum with formants





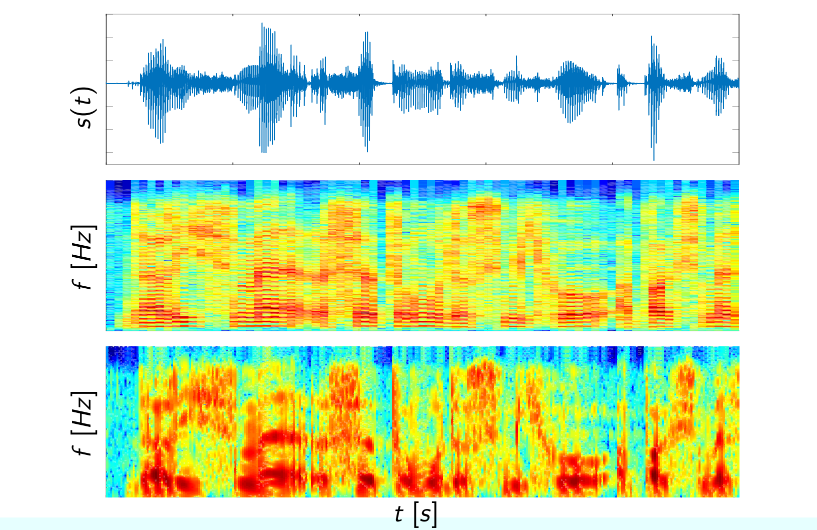
poles & LPC spectrum with formants (preemphasis)



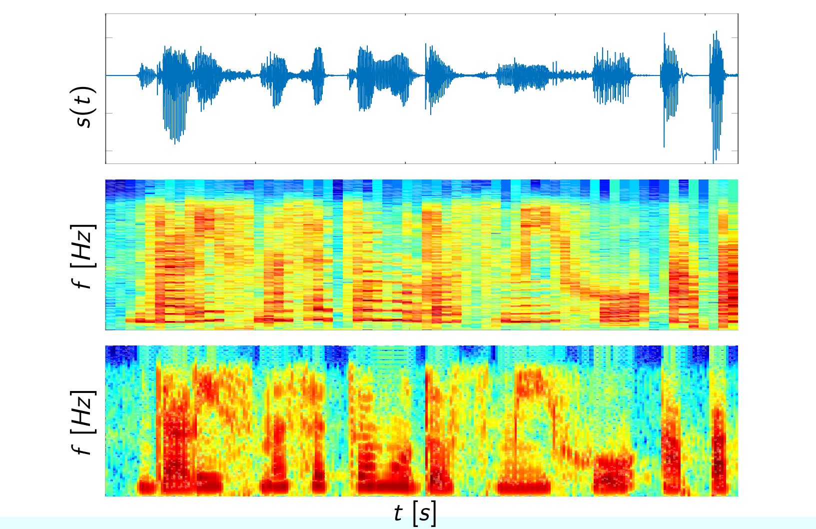


Formants in DFT spectrogram

Male voice - longer vs. shorter analyzing short-time frame
(harmonic components vs. smoothed spectrum)



Female voice - longer vs. shorter analyzing short-time frame
(harmonic components vs. smoothed spectrum)



Thank you for your attention