Určování základního tónu řeči

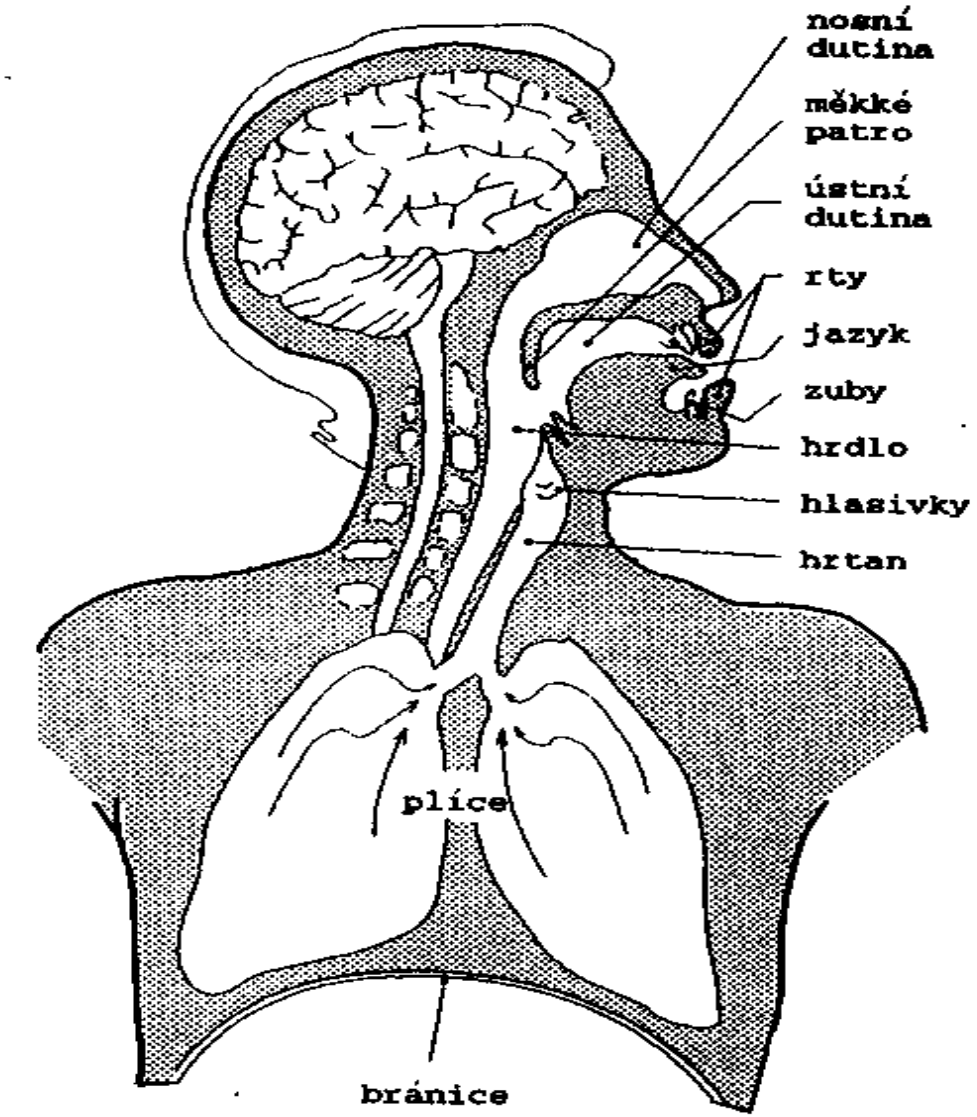
Jan Černocký ÚPGM FIT VUT Brno, cernocky@fit.vutbr.cz

FIT VUT Brno

Plán

- Charakteristiky základního tónu
- Problémy určování.
- Autokorealční metoda, AMDF, NFFC.
- Omezení vlivu formantů.
- Dlouhodobý prediktor
- Cepstrum
- Vylepšení odhadu základního tónu.

Opakování – tvorba řeči a její model



Úvod

- Frekvence základního tónu je základním kmitočtem, na kterém kmitají hlasivky: F_0, anglický název *pitch*.
- Periodu základního tónu *(pitch period)* spočítáme jako převrácenou hodnotu frekvence: $T_0 = \frac{1}{F_0}$.
- Jako *lag* označujeme periodu základního tónu vyjádřenou ve vzorcích: $L = T_0 F_s$, kde F_s je vzorkovací frekvence.

Využití základního tónu

- **syntezátory řeči** – generování melodie.
- **kódování**
  - v jednoduchém kódování označovaném jako LPC se zmenšení bitového toku dosáhne tak, že se samostatně přenáší parametry artikulačního ústrojí (např. koeficienty predikčního filtru a_i nebo odvozené), energie, příznak znělý/neznělý a F_0.
  - v modernějších kodérech (např. v RPE-LTP nebo ACELP pro mobilní telefony GSM) se využívá **dlouhodobého prediktoru LTP** (long time predictor). Jedná se o filtr s "dlouhou" impulsní odezvou, která však obsahuje jen jeden nebo několik nenulových prvků. $\Rightarrow$ další "bělení" signálu.

Charakteristiky základního tónu

- F_0 může nabývat hodnot od 50 Hz (muži) až do 400 Hz (děti), při F_s=8000 Hz tyto frekvence odpovídají lagům L=160 až 20 vzorků. Je patrné, že při malých hodnotách F_0 se blížíme délkám běžně používaných oken (20 ms, což odpovídá 160 vzorkům).
- kolísání u jednoho mluvčího může být až v poměru 2:1.
- pro různé hlásky mívá základní tón typické průběhy, malé změny po prvním kmitu ($\Delta F_0 < 10$ Hz) charakterizují mluvčího, ale obtížně se zjišťují. V radiotechnice se těmto malým posuvům říká "jitter".
- F_0 je ovlivněn *vším* – větnou melodií, náladou, únavou, atd. Velikosti změn F_0 jsou větší (větší "modulování" hlasu) u profesionálních mluvčích, obyčejní lidé mluví monotónněji.

Problémy určování základního tónu

- ani znělé hlásky nejsou zcela periodické. Čistě periodický může být pouze velmi čistý zpěv. Při generování řeči s F_0=konst. je výsledná řeč monotónní.
- nevyskytuje se čistě znělé nebo neznělé buzení. Většinou je buzení smíšené (šum na vyšších frekvencích).
- při nízké energii signálu je určení znělosti a základního tónu obtížné.
- vysoký F_0 může být ovlivněn nízkým formantem F_1 (ženy, děti).
- při přenosu řeči v telefonním pásmu (300–3400 Hz) nemáme k dispozici základní harmonickou základního tónu F_0, pouze násobky (vyšší harmonické). Filtrace za účelem získání F_0 by tedy k ničemu nevedla...

Metody pro určování základního tónu

- autokorelační + NCCF, kterou budeme aplikovat na původní signál, dále na tzv. klipovaný signál a na chybu lineární predikce.
- využití prediktoru chyby lineární predikce.
- cepstrální metoda.

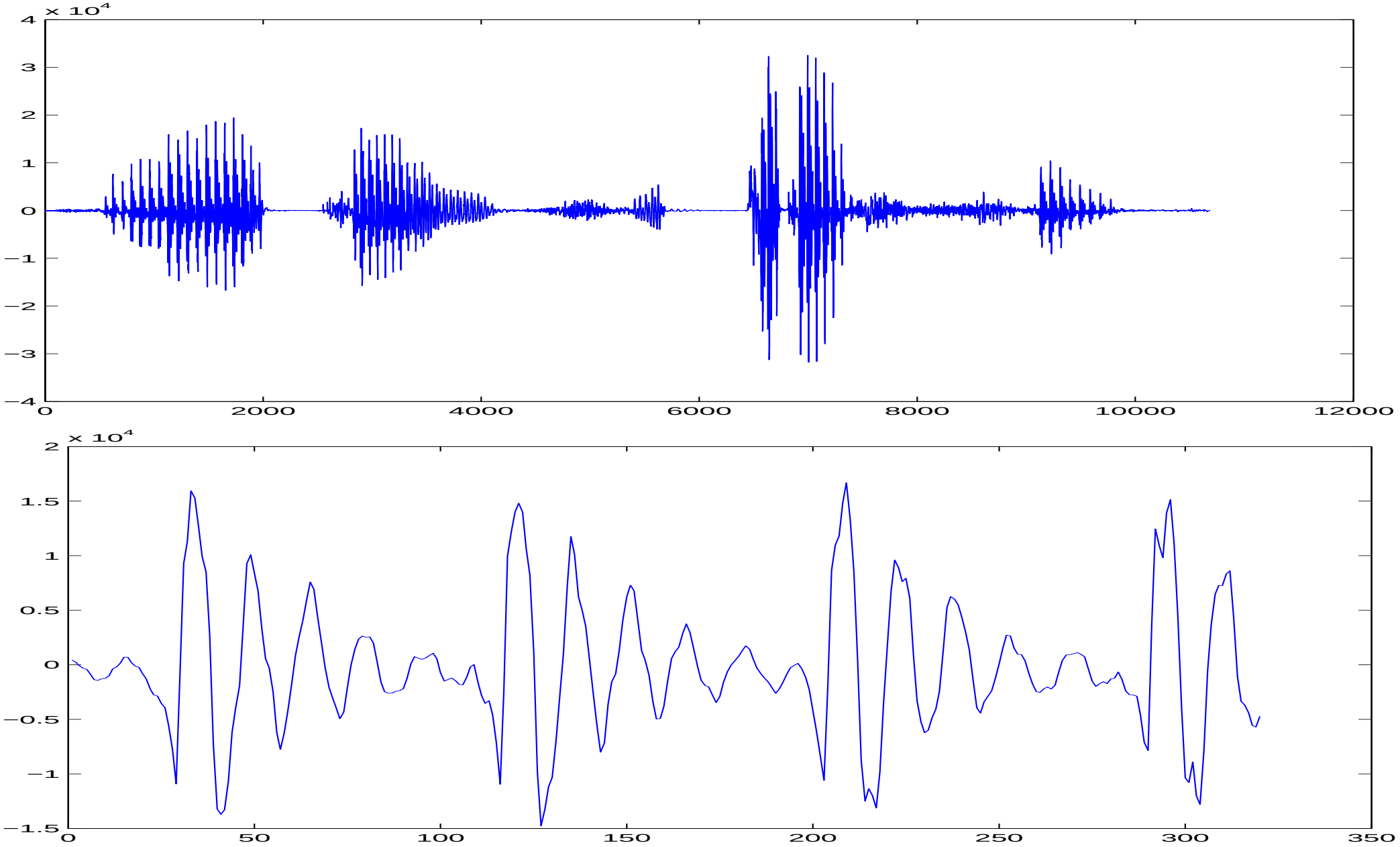
Autokorelační funkce – ACF (Autocorrelation function)

$$R(m) = \sum_{n=0}^{N-1-m} s(n)s(n+m) \tag{1}$$

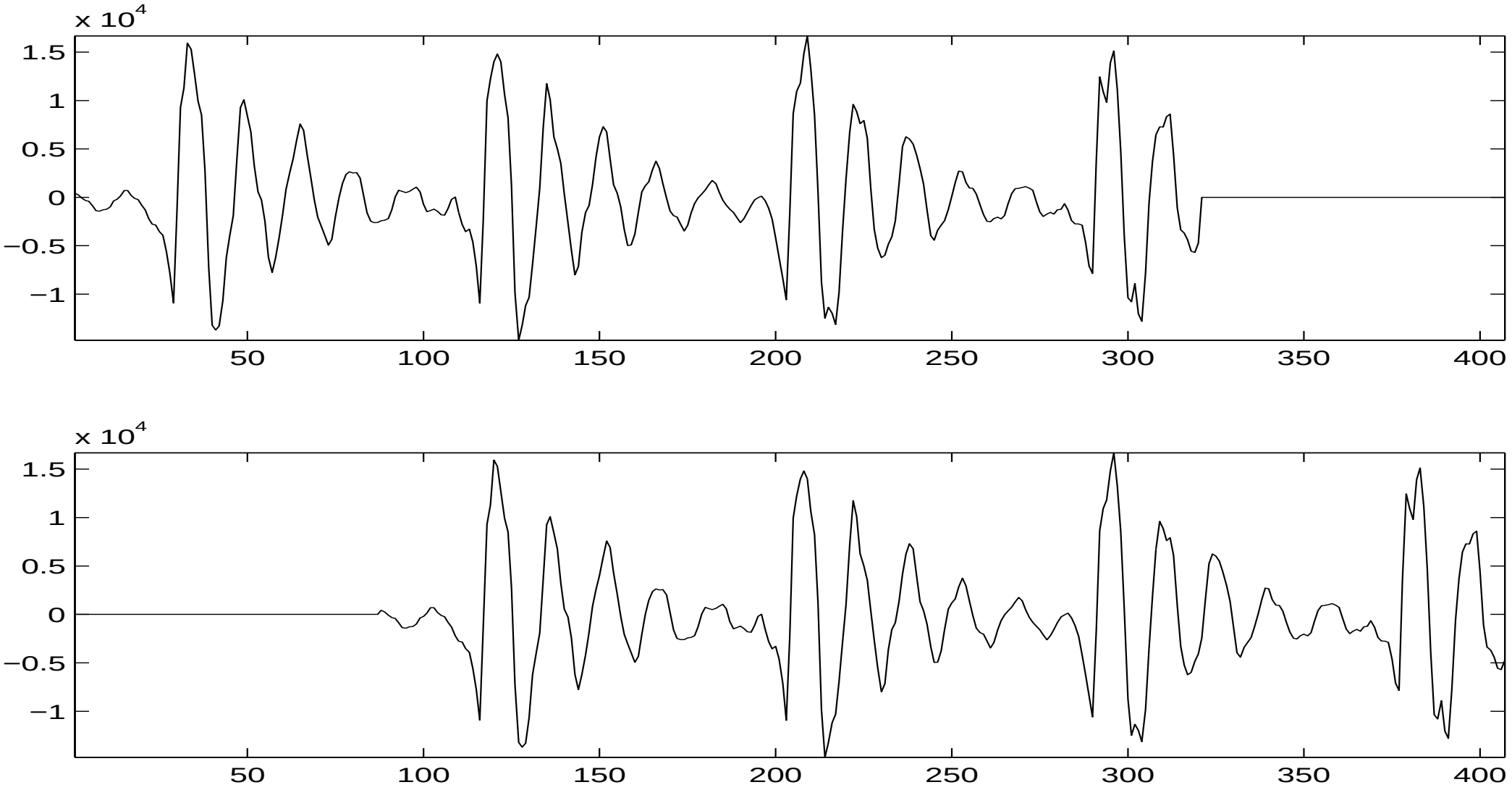
S využitím symetrie autokorelačních koeficientů:

$$R(m) = \sum_{n=m}^{N-1} s(n)s(n-m) \tag{2}$$

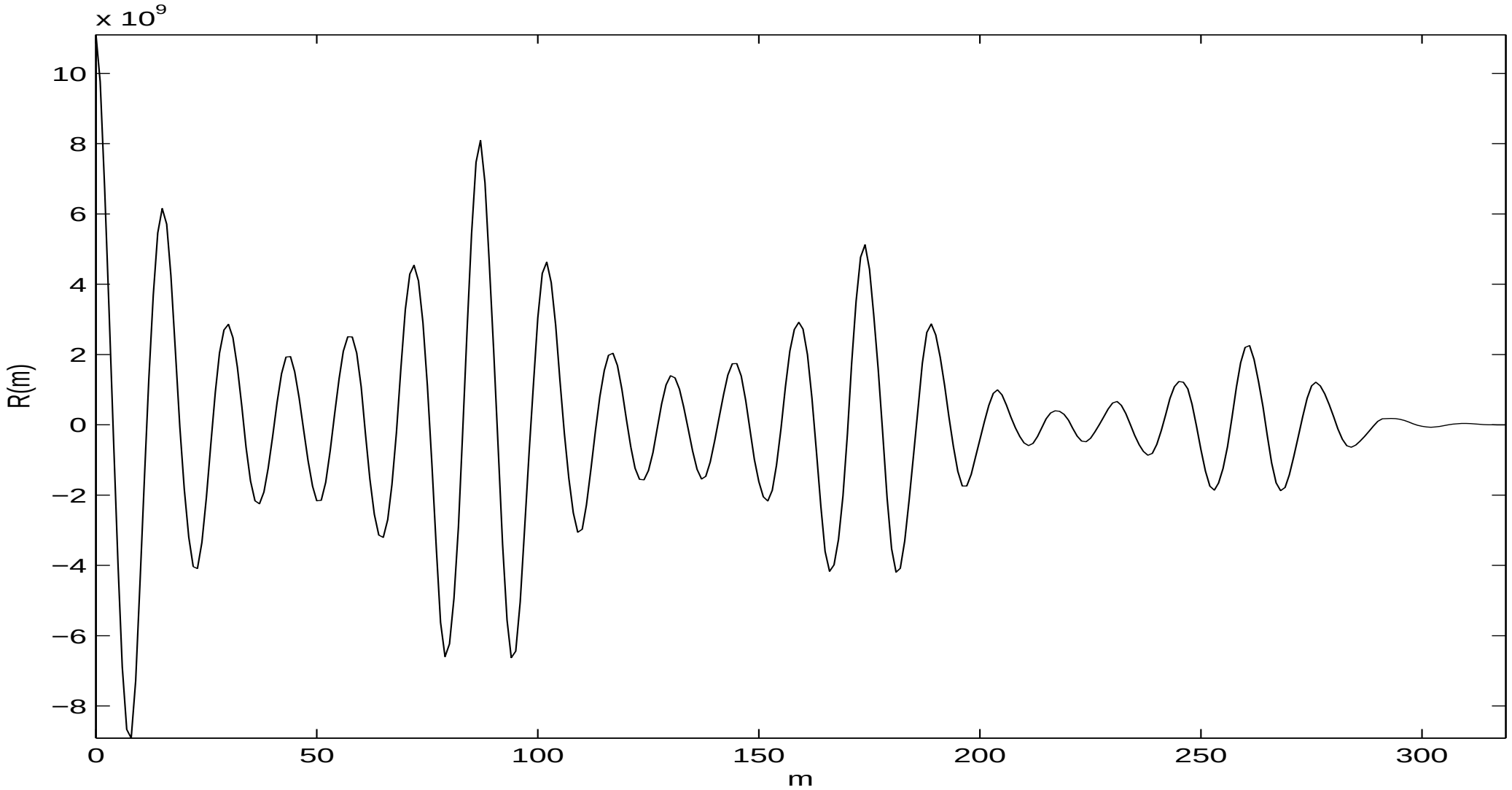
Celý signál a jeden rámec



Ilustrace posunu



Vypočtená autokorelační funkce



Výpočet lagu a určení znělosti

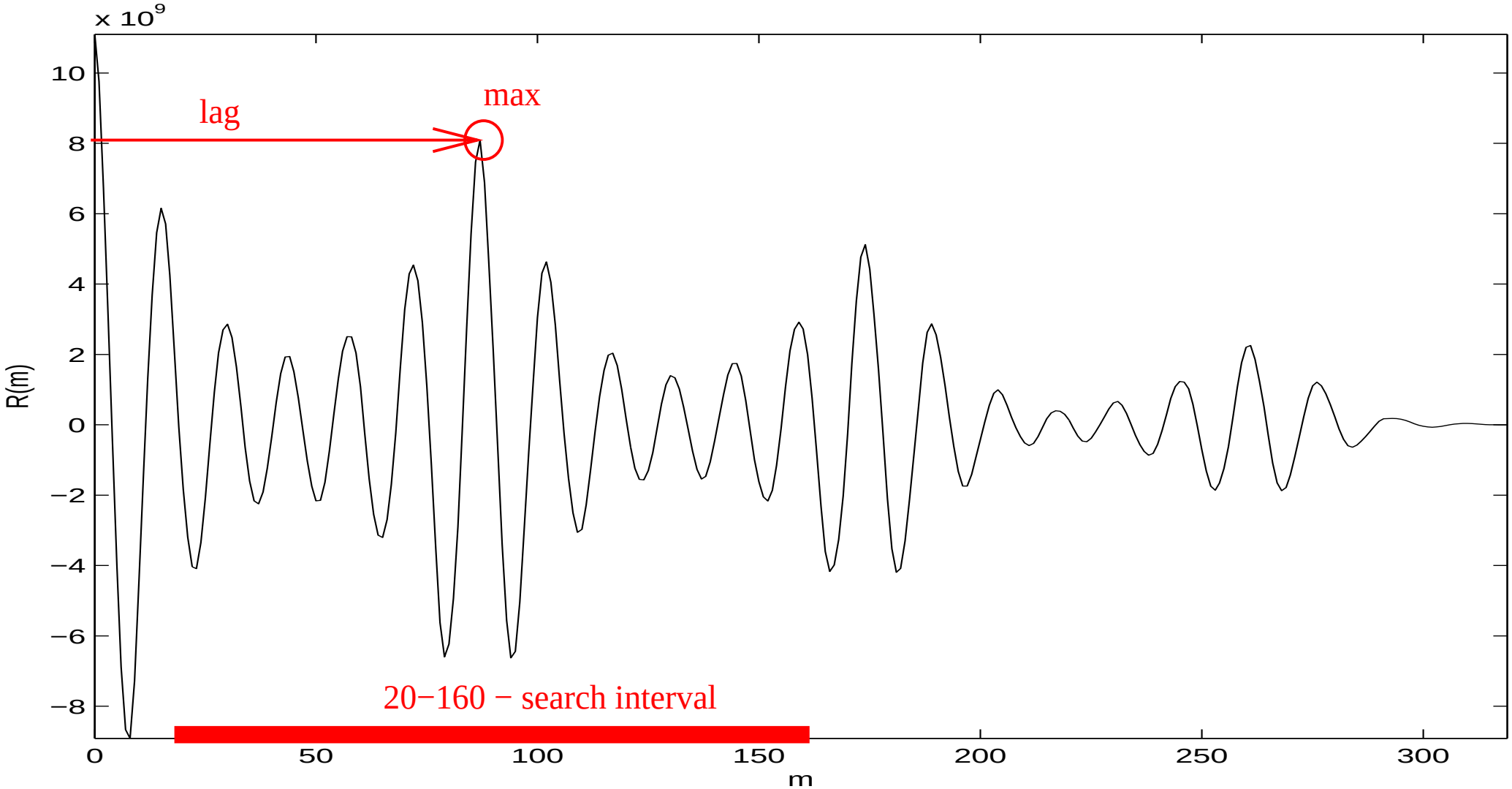
Lag se z ACF, že hledáme maximum autokorelační funkce:

$$R(m) = \sum_{n=0}^{N-1-m} c[s(n)]c[s(n+m)] \tag{3}$$

Znělost rámce můžeme odhadnout porovnáním nalezeného maxima s nultým (maximálním) autokorelačním koeficientem. Konstanta α se musí zvolit experimentálně.

$$\begin{aligned} R_{max} < \alpha R(0) \quad &\Rightarrow \quad \text{neznělý} \\ R_{max} \geq \alpha R(0) \quad &\Rightarrow \quad \text{znělý} \end{aligned} \tag{4}$$

Hledání maxima ACF $\Rightarrow$ lag (pro uvedený obrázek L=87):

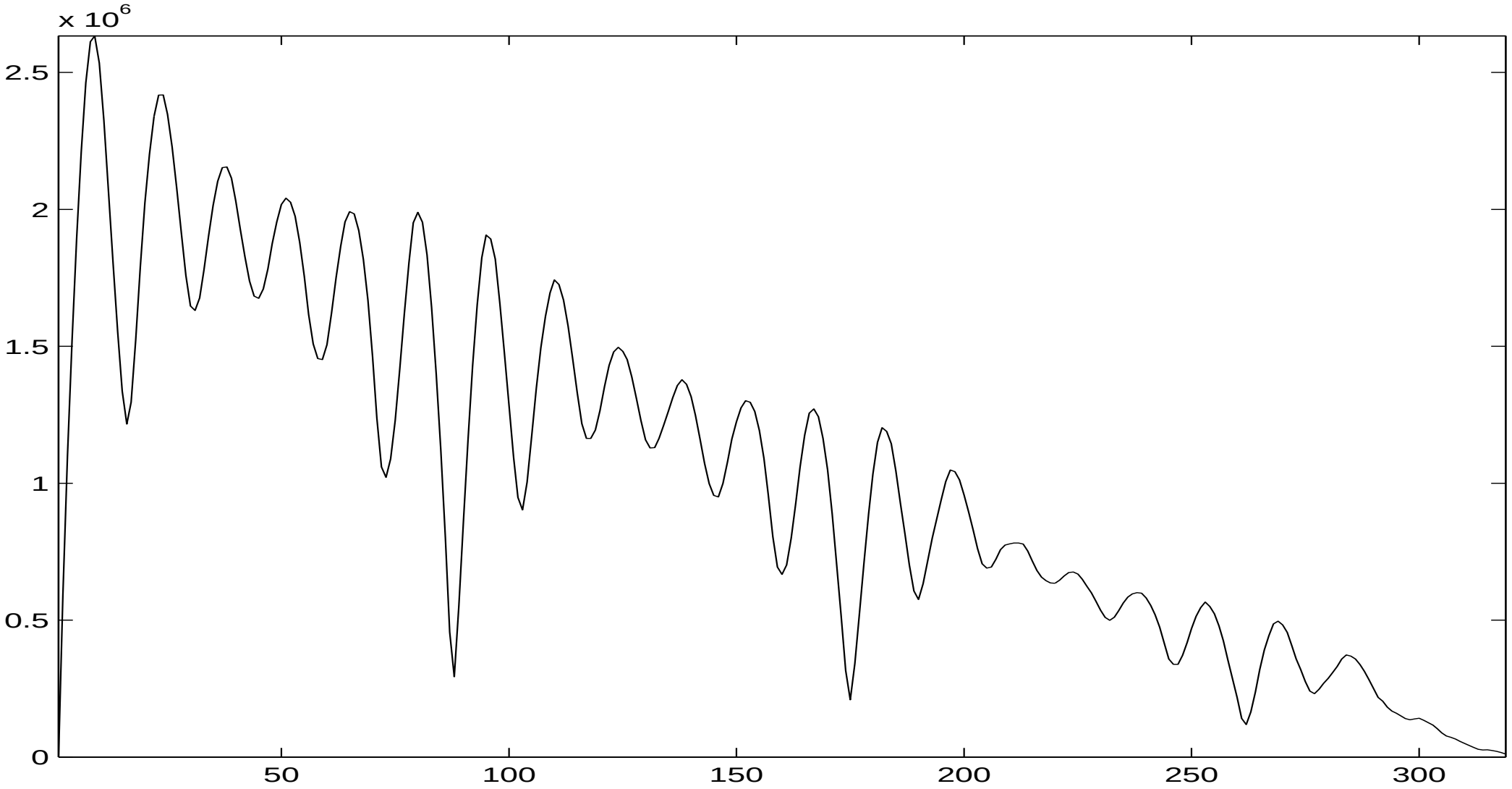


AMDF

V dávných dobách, kdy bylo násobení náročnější na čas procesoru, se autokorelační funkce nahrazovala funkcí AMDF (Average Magnitude Difference Function):

$$R_D(m) = \sum_{n=0}^{N-1-m} |s(n) - s(n+m)|, \tag{5}$$

kde bylo naopak nutné hledat pro určení lagu *minimum*.

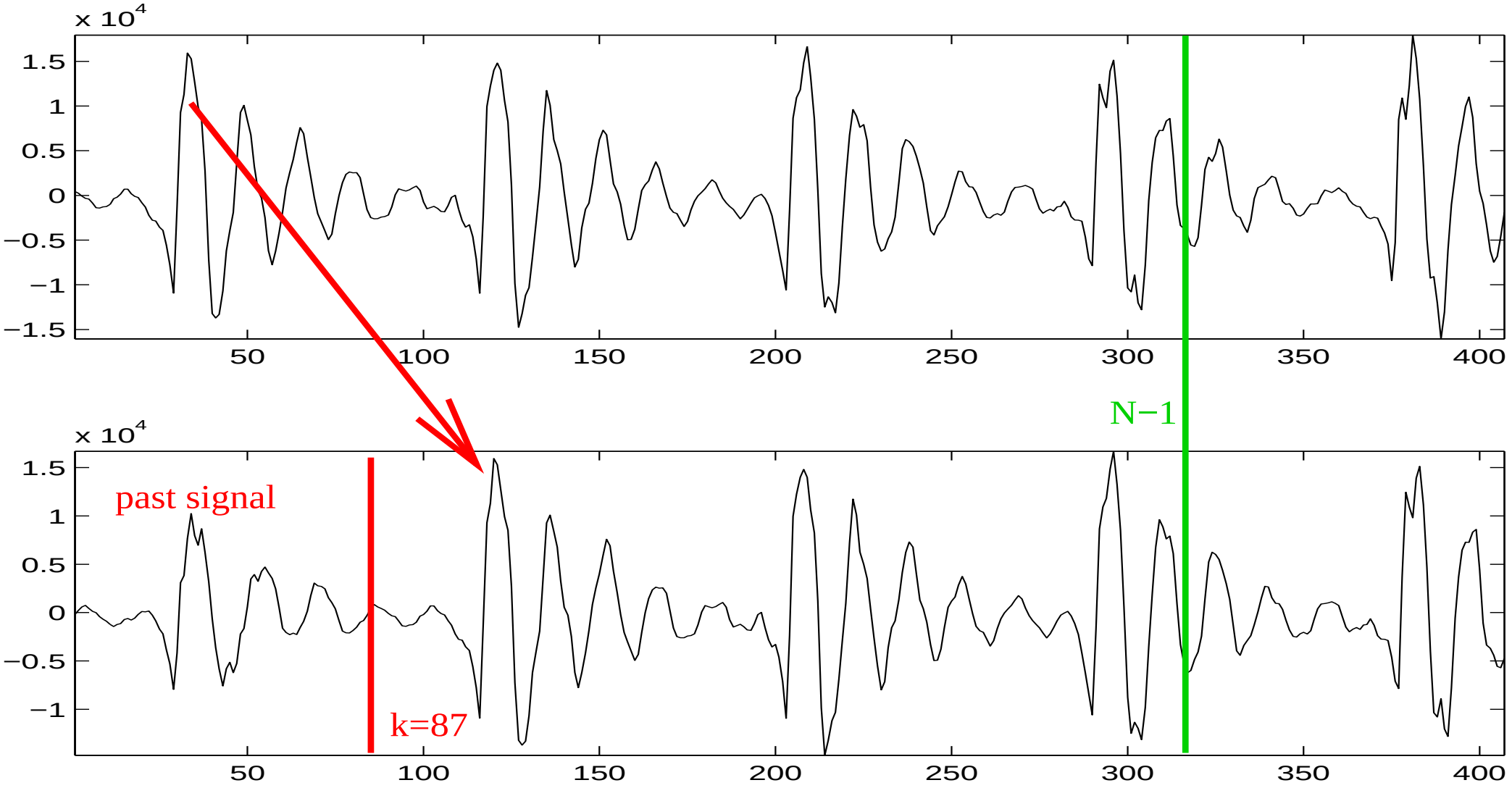

x 10^6
2.5
2
1.5
1
0.5
0
50
100
150
200
250
300

Cross-correlation function

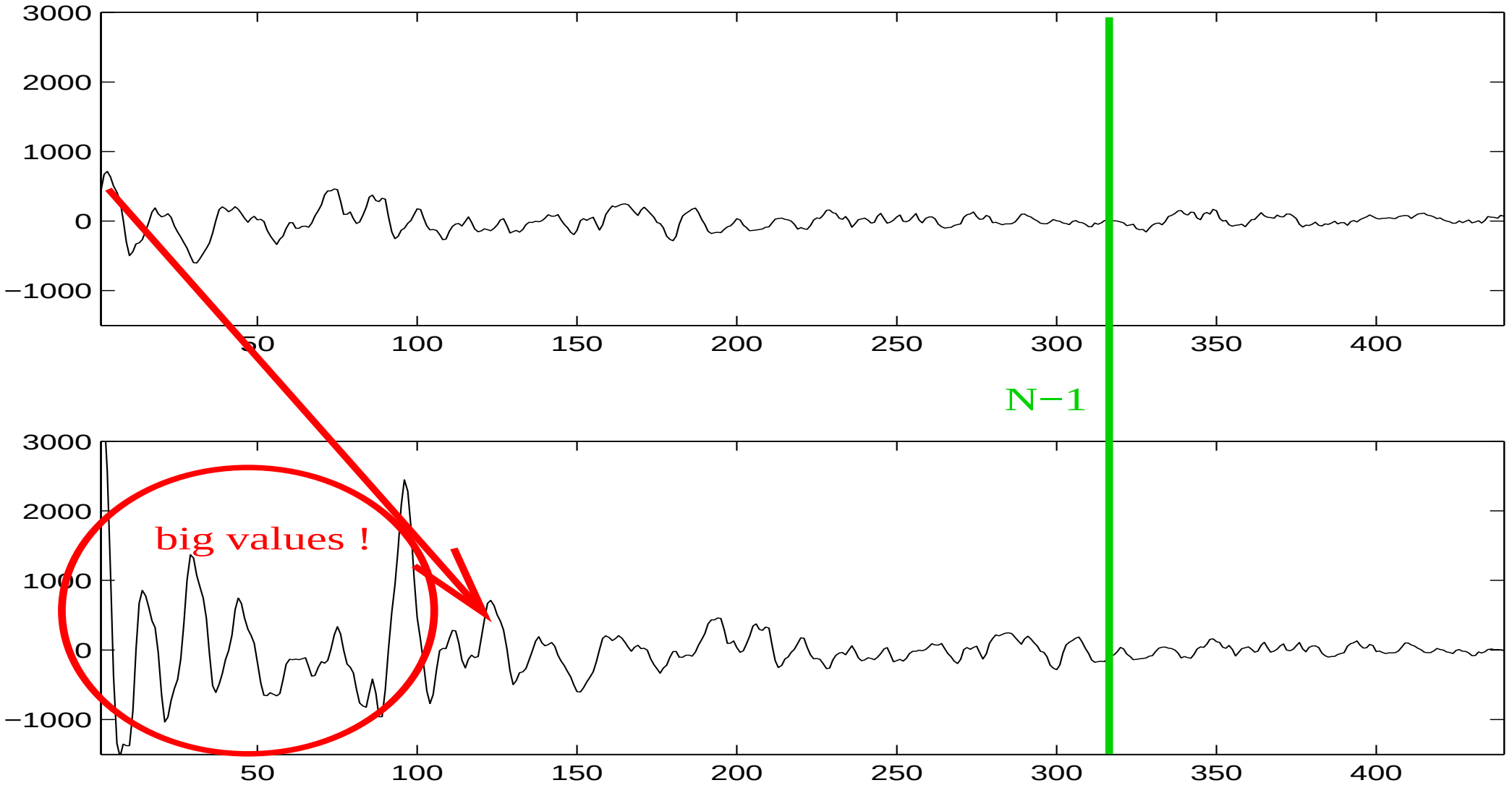
Navýhoda standardní ACF je postupné "zkracování" oblasti, ze které autokorelační koeficienty počítáme. Můžeme si dovolit použít celý signál (při reálném zpracování si ho musíme zapamatovat) $\Rightarrow$ **CCF**. Začátek rámce označíme zr:

$$CCF(m) = \sum_{n=zr}^{zr+N-1} s(n)s(n-m) \tag{6}$$

Posunutí pro výpočet NCCF:



Posunutí pro výpočet NCCF - problém, protože posunutý signál má mnohem větší energii !



Normalized cross-correlation function

Rozdílnost energií originálního a posunutého rámce můžeme řešit pomocí normalizace:

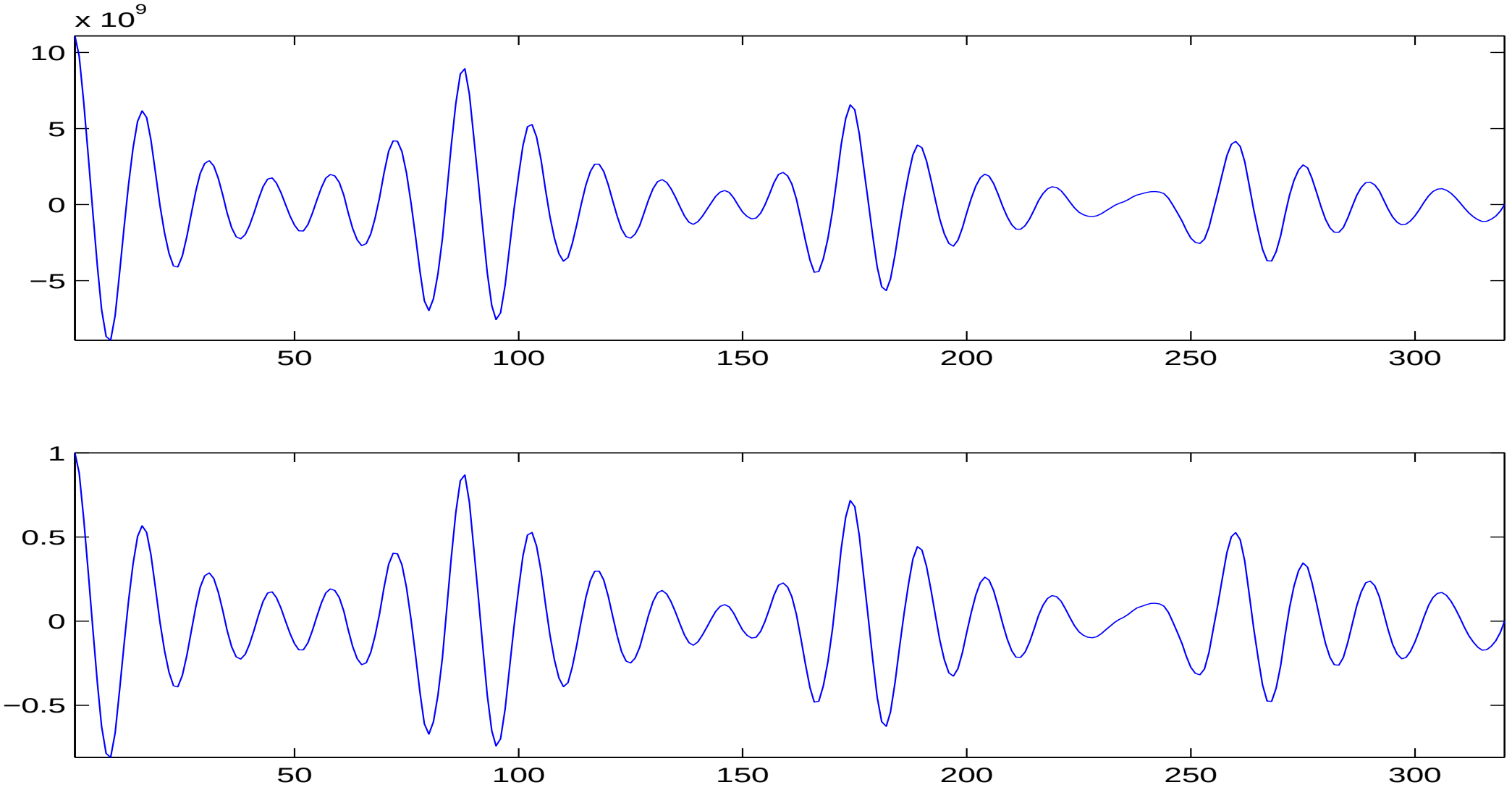
NCCF

$$CCF(m) = \frac{\sum_{n=zr}^{zr+N-1} s(n)s(n-m)}{\sqrt{E_1 E_2}} \tag{7}$$

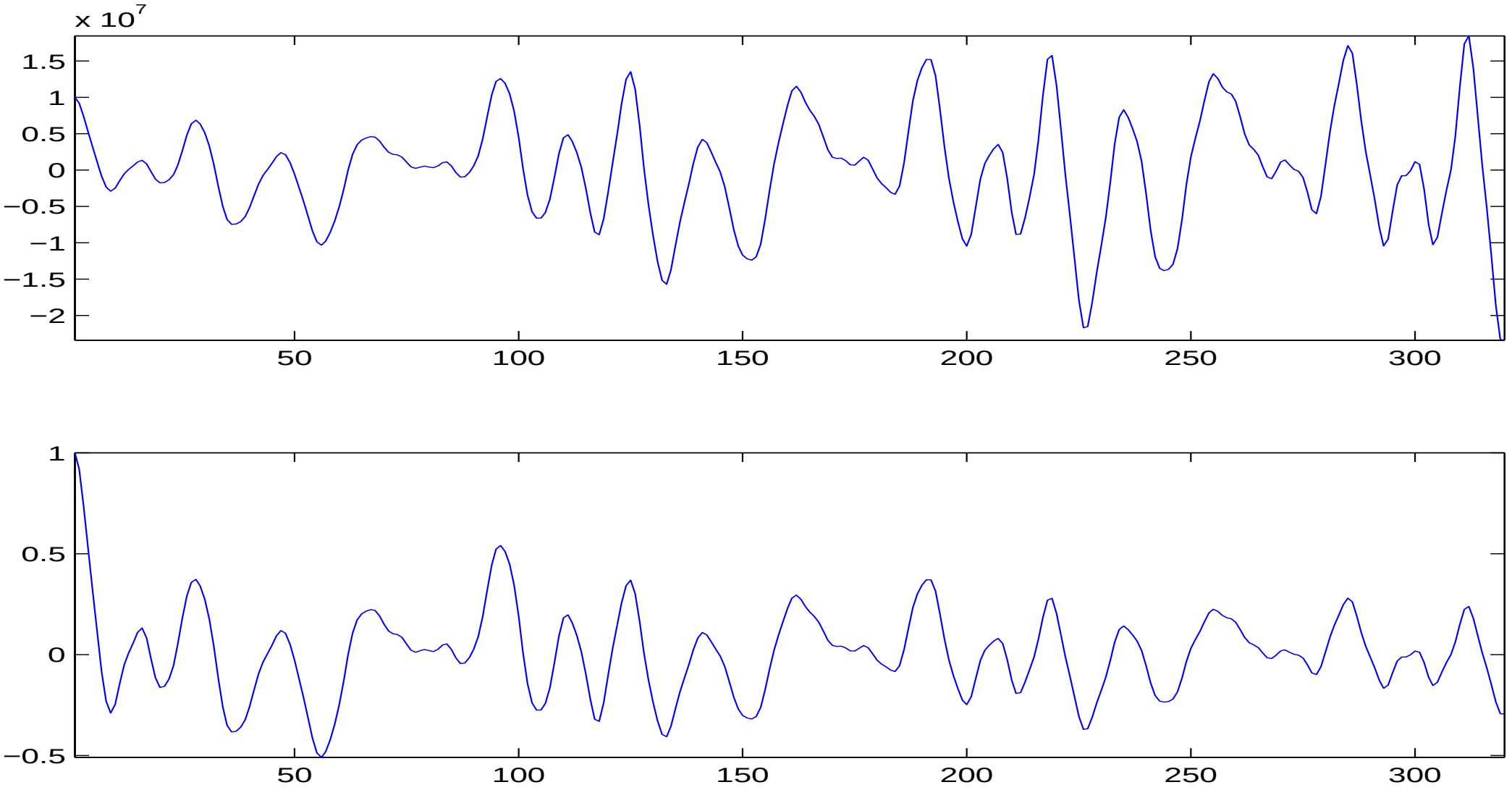
E_1 a E_2 jsou energie originálního a posunutého rámce:

$$E_1 = \sum_{n=zr}^{zr+N-1} s^2(n) \qquad E_2 = \sum_{n=zr}^{zr+N-1} s^2(n-m) \tag{8}$$

CCF a NCCF pro “dobrý příklad”



CCF a NCCF pro "špatný příklad"



Nevýhoda: metody nedostatečně potlačují vliv formantů (to se projeví dalšími maximy v ACF nebo v AMDF).

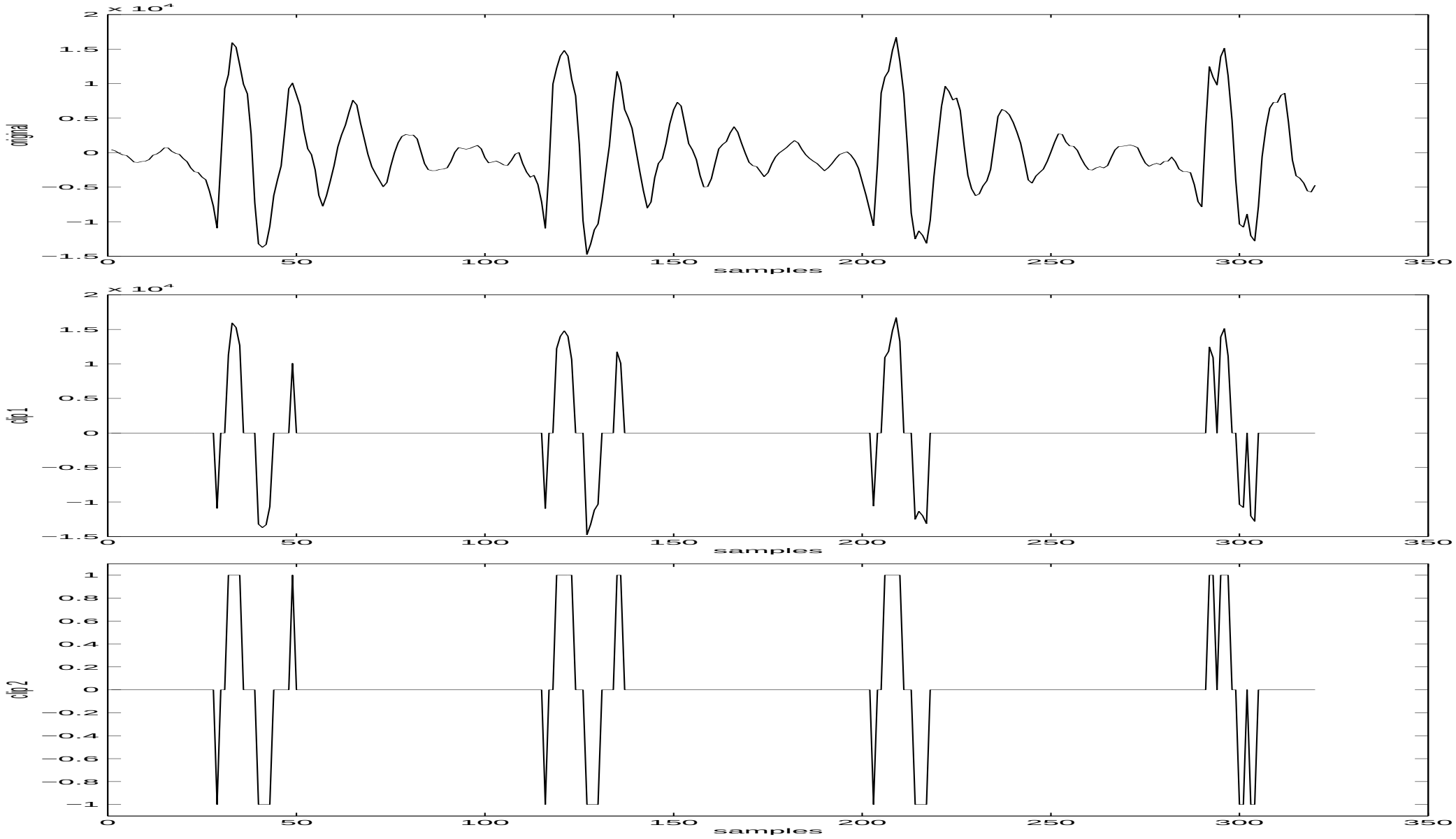
Centrální klipování – Center Clipping

předzpracovává signál pro ACF, se zajímáme pouze o *špičky signálu*. Definujeme tzv. klipovací úroveň c_L. V první variantě této metody ze signálu "vynecháváme interval" $< -c_L, +c_L >$. Ve druhé variantě nahrazujeme hodnotou 1 signál tam, kde je překročena úroveň c_L a hodnotou -1 tam, kde signál nedosáhne úrovně $-c_L$:

$$c_1[s(n)] = \begin{cases} s(n) - c_L & \text{pro} \quad s(n) > c_L \\ 0 & \text{pro} \quad -c_L \leq s(n) \leq c_L \\ s(n) + c_L & \text{pro} \quad s(n) < -c_L \end{cases} \tag{9}$$

$$c_2[s(n)] = \begin{cases} +1 & \text{pro} \quad s(n) > c_L \\ 0 & \text{pro} \quad -c_L \leq s(n) \leq c_L \\ -1 & \text{pro} \quad s(n) < -c_L \end{cases} \tag{10}$$

Obrázky ilustrují klipování na rámci řečového signálu pro klipovací úroveň 9562:



Určení klipovací úrovně

Vzhledem ke kolísání signálu $s(n)$ nemůže být konstantní a je nutné ji určovat pro každý rámec, na kterém odhadujeme základní tón. Jednoduchou metodou je určení klipovací úrovně z maximální absolutní hodnoty vzorků v rámci:

$$c_L = k \max_{n=0\ldots N-1} |x(n)|, \tag{11}$$

kde konstanta k se volí od 0.6 do 0.8. Sofistikovanější metoda využívá rozdělení rámce na několik mikro-rámců, např. $x_1(n)$, $x_2(n)$, $x_3(n)$ o třetinové délce. Klipovací úroveň je pak určena pomocí "nejslabšího" maxima z těchto mikro-rámců jako:

$$c_L = k \min \{\max |x_1(n)|, \max |x_2(n)|, \max |x_3(n)|\} \tag{12}$$

Problém: klipování šumu v pauzách, kde může být následně detekován základní tón a znělost. Metodu je vhodné doplnit určením úrovně ticha s_L (silence level). Pokud maximum signálu $< s_L$, pak se znělost a lag neurčují.

Využití chyby lineární predikce

Jedná se o předzpracování nejen pro metodu ACF, ale i pro jiné algoritmy určení základního tónu. Opakování: chybu lineární predikce získáme jako rozdíl skutečného a předpovězeného signálu:

$$e(n) = s(n) - \hat{s}(n) \tag{13}$$

$$E(Z) = S(z)[1 - (1 - A(z))] = S(z)A(z) \tag{14}$$

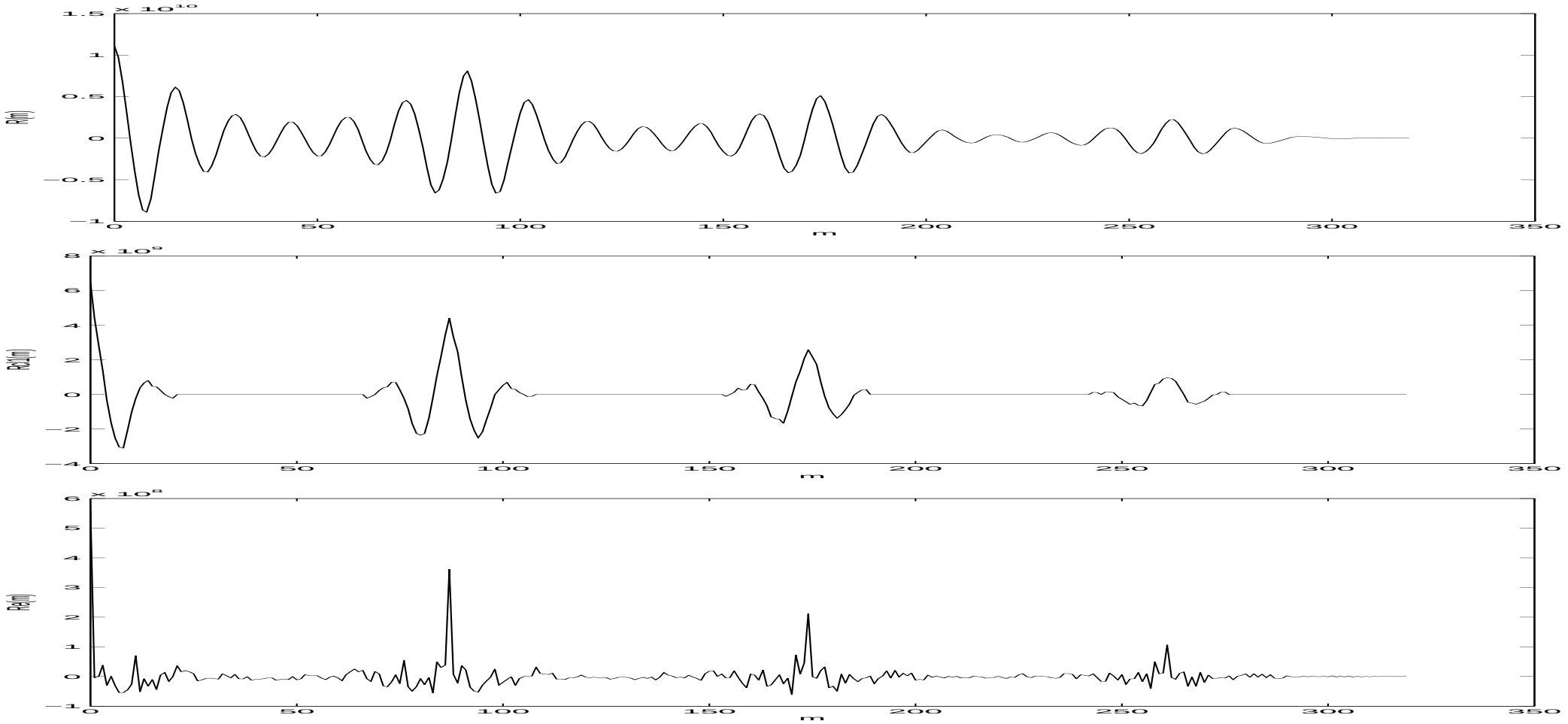
$$e(n) = s(n) + \sum_{i=1}^{P} a_i s(n-i) \tag{15}$$

$$\tag{16}$$

Signál $e(n)$ již neobsahuje informaci o formantech, proto je k určování základního tónu vhodnější než základní signál. Určení lagu z chybového signálu můžeme provést pomocí ACF apod.

Srovnání autokorelačních funkcí

Následující obrázek presentuje autokorelační funkce vypočítané ze základního signálu, z klipovaného signálu a z chyby lineární predikce.



Dlouhodobý prediktor chyby predikce pro určení základního tónu

Snažíme se předpovědět n-tý vzorek signálu ne z P předcházejících vzorků (jako u LPC), ale ze dvou vzorků vzdálených o předpokládaný lag. Pokud určíme posun s minimální energií chyby predikce, lag jsme nalezli. Predikovanou chybu predikce zapíšeme:

$$\hat{e}(n) = -\beta_1 e(n-m+1) - \beta_2 e(n-m) \tag{17}$$

Pak je chyba prediktoru chyby predikce dána:

$$ee(n) = e(n) - \hat{e}(n) = e(n) + \beta_1 e(n-m+1) + \beta_2 e(n-m) \tag{18}$$

Chceme minimalisovat energii tohoto signálu:

$$\min E = \min \sum_{n=0}^{N-1} ee^2(n) \tag{19}$$

Postupujeme podobně jako při výpočtu LPC koeficientů, jako řešení dostáváme pro

koeficienty β_1 a β_2:

$$\begin{aligned}\beta_1 &= [r_e(1)r_e(m) - r_e(m-1)]/[1 - r_e^2(1)] \\ \beta_2 &= [r_e(1)r_e(m-1) - r_e(m)]/[1 - r_e^2(1)],\end{aligned} \tag{20}$$

kde $r_e(m)$ jsou normované autokorelační koeficienty chybového signálu $e(n)$. Po dosazení těchto koeficientů do vzorce pro energii 19 můžeme tuto energii zapsat v závislosti na posunutí m jako:

$$E(m) = 1 - K(m)/[1 - r_e^2(1)] \tag{21}$$

$$\text{kde} \quad K(m) = r_e^2(m-1) + r_e^2(m) - 2r_e(1)r_e(m-1)r_e(m) \tag{22}$$

Lag nyní můžeme najít buď tak, že vyhledáme minimální energii nebo tak, že najdeme maximální hodnotu funkce $K(m)$ (uvědomme si, že jmenovatel $1 - r_e^2(1)$ na m nezávisí).

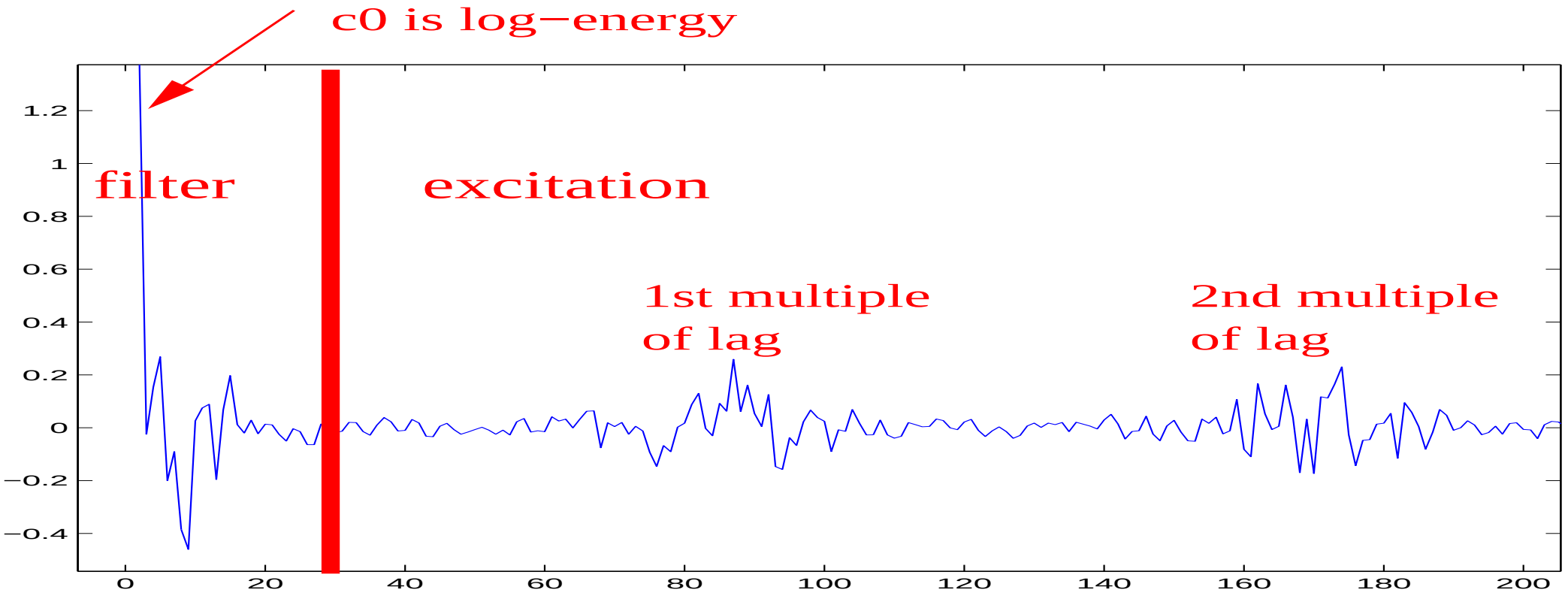
$$L = \arg \min_{m \in [L_{min}, L_{max}]} E(m) = \arg \max_{m \in [L_{min}, L_{max}]} K(m) \tag{23}$$

Cepstrální analýza pro určení základního tónu

Cepstrální koeficienty můžeme získat pomocí tohoto vztahu:

$$c(m) = \mathcal{F}^{-1}\left[ln|\mathcal{F}s(n)|^2\right] \tag{24}$$

V cepstrálních koeficientech se daří oddělit část koeficientů zodpovědnou za hlasový trakt (nízké indexy) od části zodpovědné za buzení, a tedy i za základné tón (vyšší indexy). Lag je nutné opět nalézt hledáním maxima $c(m)$ v rozsahu povolenách lagů.


c0 is log−energy
filter
excitation
1st multiple
of lag
2nd multiple
of lag
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
0
20
40
60
80
100
120
140
160
180
200

Zlepšení spolehlivosti určení základního tónu

Místo skutečného lagu je často detekován poloviční či několikanásobný lag. Předpokládejme např., že v pěti po sobě jdoucích rámcích byly detekovány tyto lagy: 50, 50, 100, 50, 50. V prostředním rámci se evidentně jedná o chybu: detekci dvojnásobného lagu. Chyby tohoto typu se můžeme snažit opravit několika způsoby.

Nelineární filtrace mediánovým filtrem

$$L(i) = \text{med}\left[L(i-k), L(i-k+1), \ldots, L(i), \ldots, L(i+k)\right] \tag{25}$$

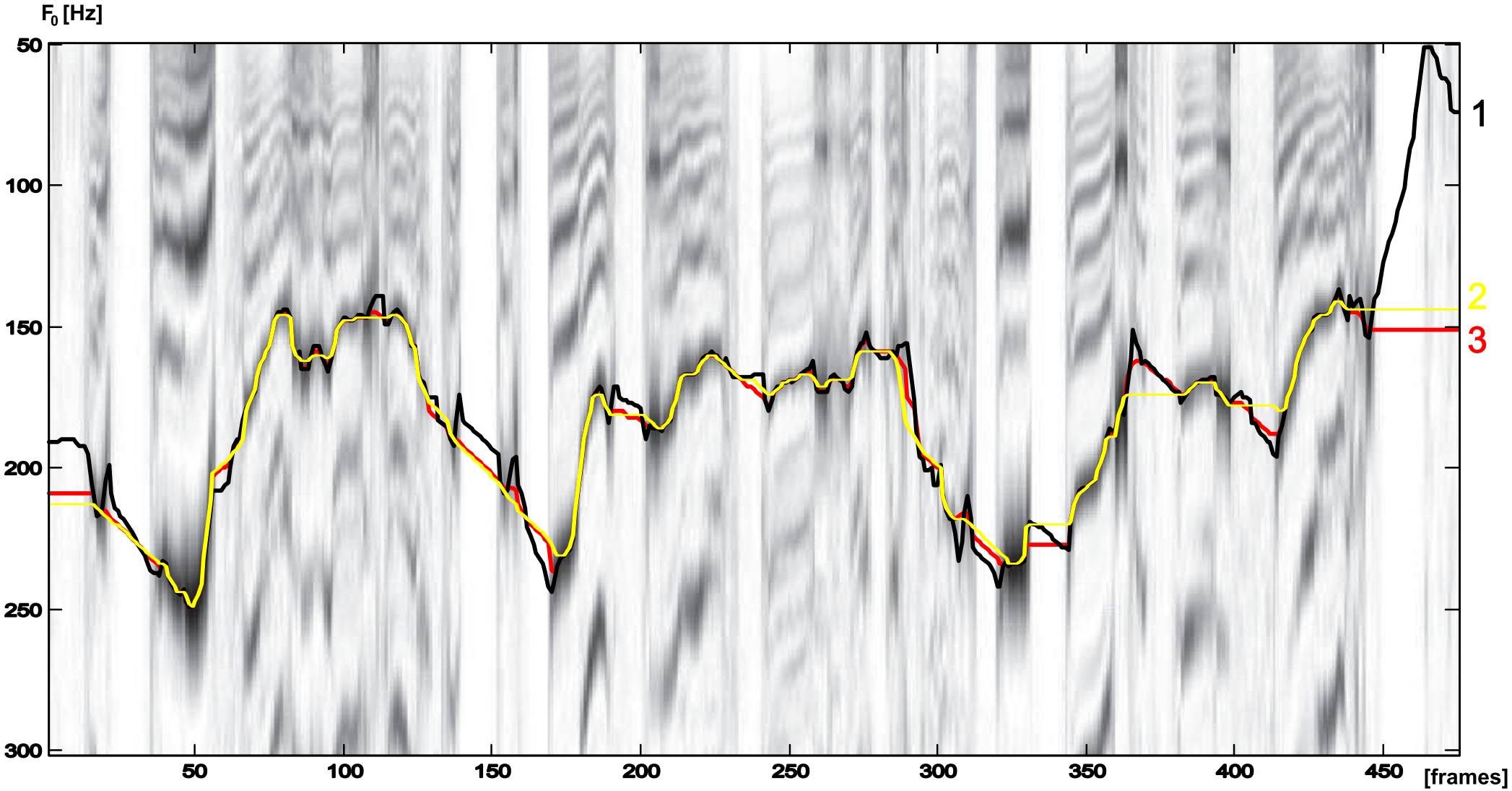
Medián seřadí hodnoty podle velikosti a vybere hodnotu, která se nachází uprostřed. Lagy z našeho příkladu tedy budou opraveny na 50, 50, 50, 50, 50.

Metoda optimálních cest

V předcházejících metodách jsme lag určovali tak, že jsme určili pouze *jedno* maximum, případně minimum na jeden rámec. Hledání maxima či minima můžeme ovšem rozšířit na několik rámců vedle sebe: nebudeme hledat hodnotu, ale "cestičku", která minimalizuje (či maximalizuje) dané kritérium. Příspěvkem ke kritériu může být např. hodnota $\frac{R(m)}{R(0)}$ nebo energie chyby predikce pro daný lag. Dále je potřeba definovat hypotézy o tvaru cesty (cesta se nemůže z jednoho rámce na druhý výrzně změnit...).

Algoritmus má pak tyto kroky:

1. určení možných cest — např tak, že rozdíp v hodnotě lagu mezi sousedními rámci nesmí být větší než konstanta ΔL.
2. určení celkového kriteria pro danou cestu.
3. výběr optimální cesty.


F0 [Hz]
50
100
150
200
250
300
50
100
150
200
250
300
350
400
450
[frames]
1
2
3

Desetinné vzorkování

Pro zvýšen přesnosti určení F_0 je vhodné signál nadvzorkovat a následně filtrovat. Dosáhneme tak zvýšení vzorkovací frekvence. Tuto operaci není nutné provádět "fyzicky"; dá se promítnout přímo do výpočtu autokorelačních koficientů. Nadvzorkování často zamezíme falešné detekci dvojnásobku skutečného lagu.

Příklad interpolovaného signálu a interpolačního filtru: