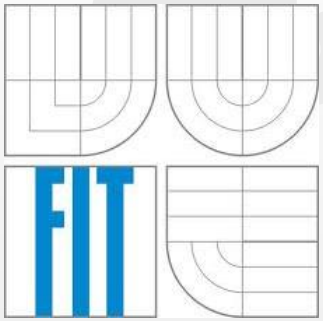
ZRE - Kódování řeči II.

Vokodéry, CELP, ACELP

Vladimír Malenovský, ÚPGM FIT VUT Brno



Historie

- 1939 – Homer Dudley's VODER
  https://www.youtube.com/watch?v=mSdFu1xdoZk

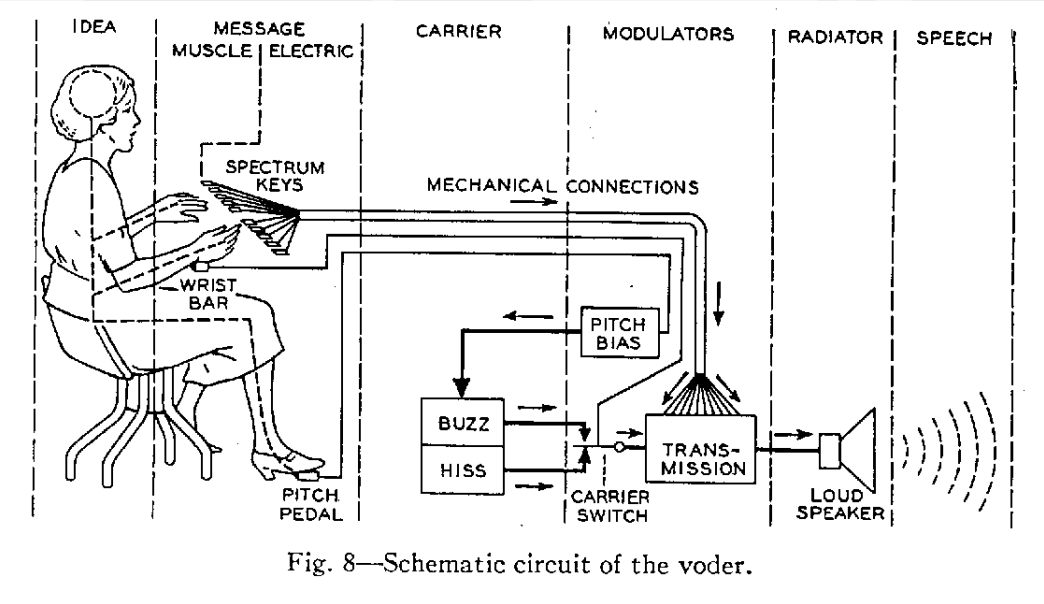


Fig. 8—Schematic circuit of the voder.

- 1960 – Siemens synthesizer (počátek elektronické hudby)
- 1978 – nástup řečových vokodérů založených na LP
- 1985 M. Schroeder a B. Atal - první článek o CELPu na koferenci ICASSP

1961 - vokodér pro přenos zašifrované hlasové zprávy přes radio a telefon

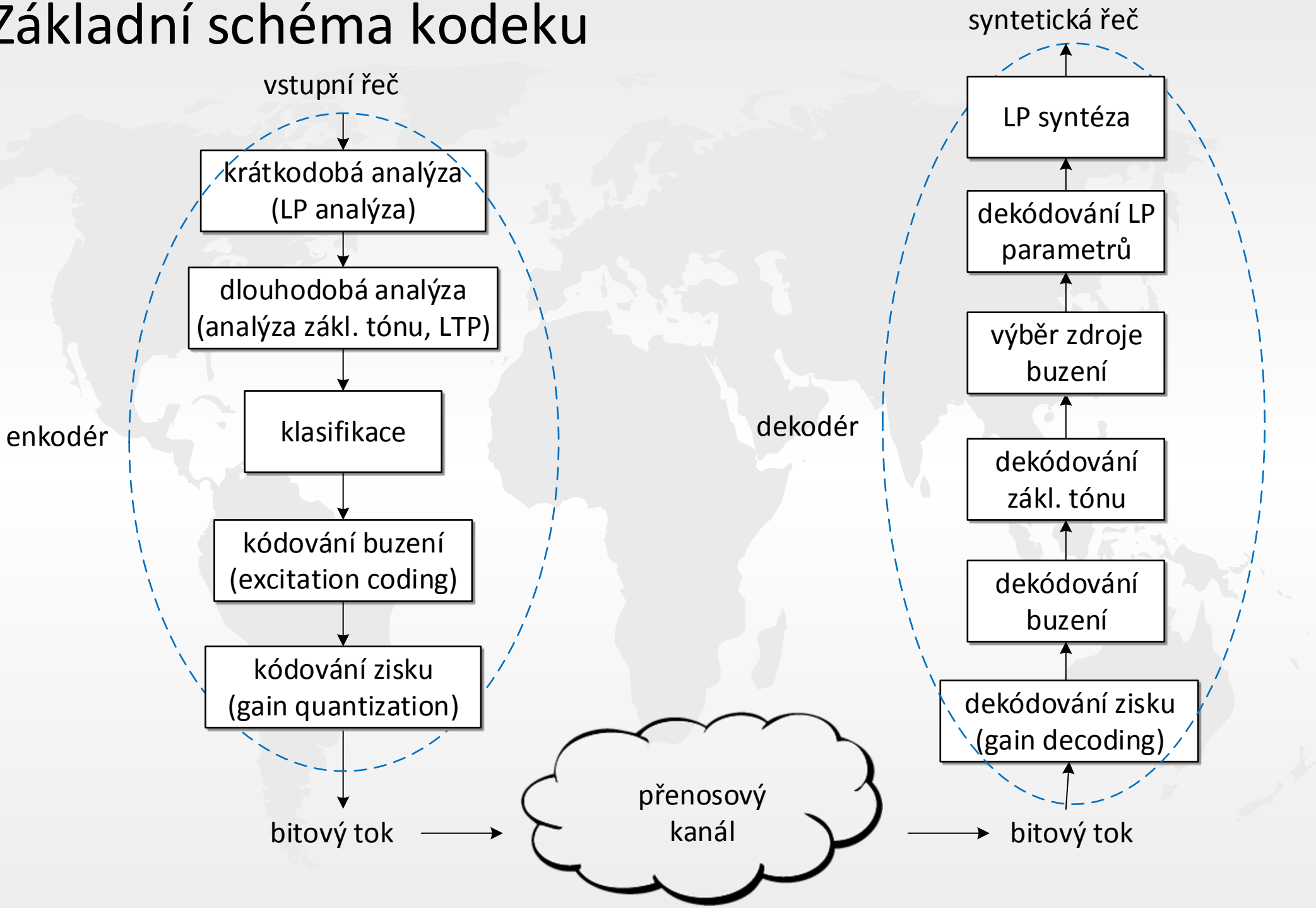

1972 - vokodér firmy Kraftwerk

Historie

- novinky oproti waveform kodekům
  - adaptivní (LTP) a fixní knihovna
  - koncept analýzy syntézou (closed-loop approach)
  - perceptuální váhování
  - vektorová kvantizace
- existuje celá řada variant CELPu použitých v různých kodecích:
  1989 – GSM FR – RPE LTP
  1992 – G.728 – LD CELP
  1995 – GSM HR – VCELP
  1996 – GSM EFR – ACELP
  1996 – G.723.1 – MPC-MLQ/ACELP
  1996 – G.729 – CS ACELP
  1999 – G.722 (AMR-NB) – ACELP
  2002 – G.722.2 (AMR-WB) – ACELP
  2006 – EVRC – RCELP
  2008 – G.718 – ACELP
  2008 – MPEG-4 – ACELP
  2012 – USAC – ACELP
  2014 – EVS – ACELP
- 1989 - ACELP je nejvýznamější varianta (vylepšení) CELPu, která je použita v téměř 98% všech mobilních zařízení na světě

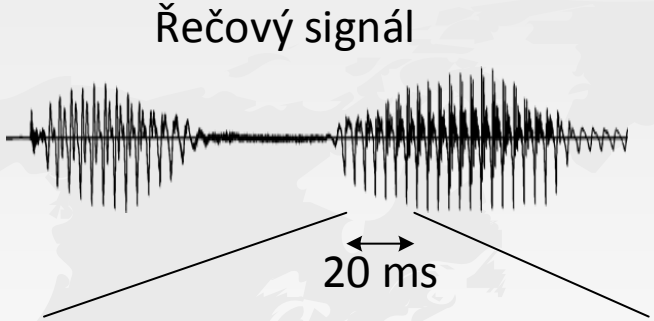
Vokodér je zařízení pro analýzu a syntézu řečového signálu na základě jeho charakteristik

Základní schéma kodeku

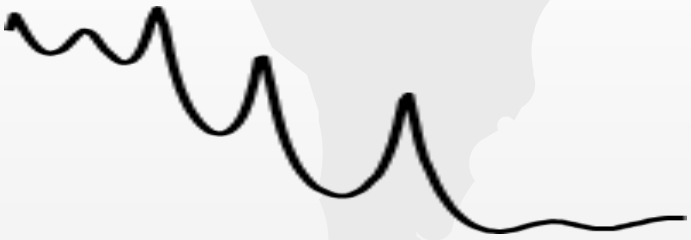


LP analýza

LP analýza



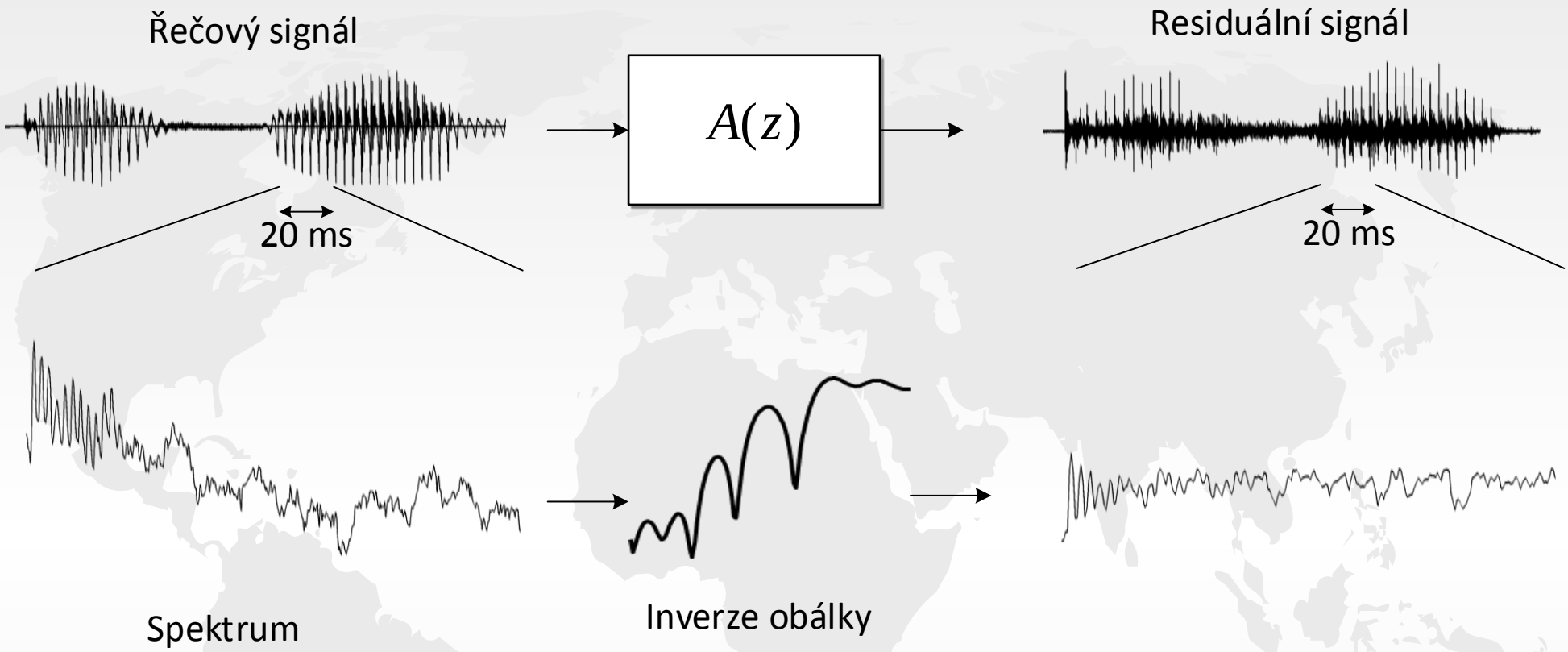
Spektrum



Obálka 1/A(z)

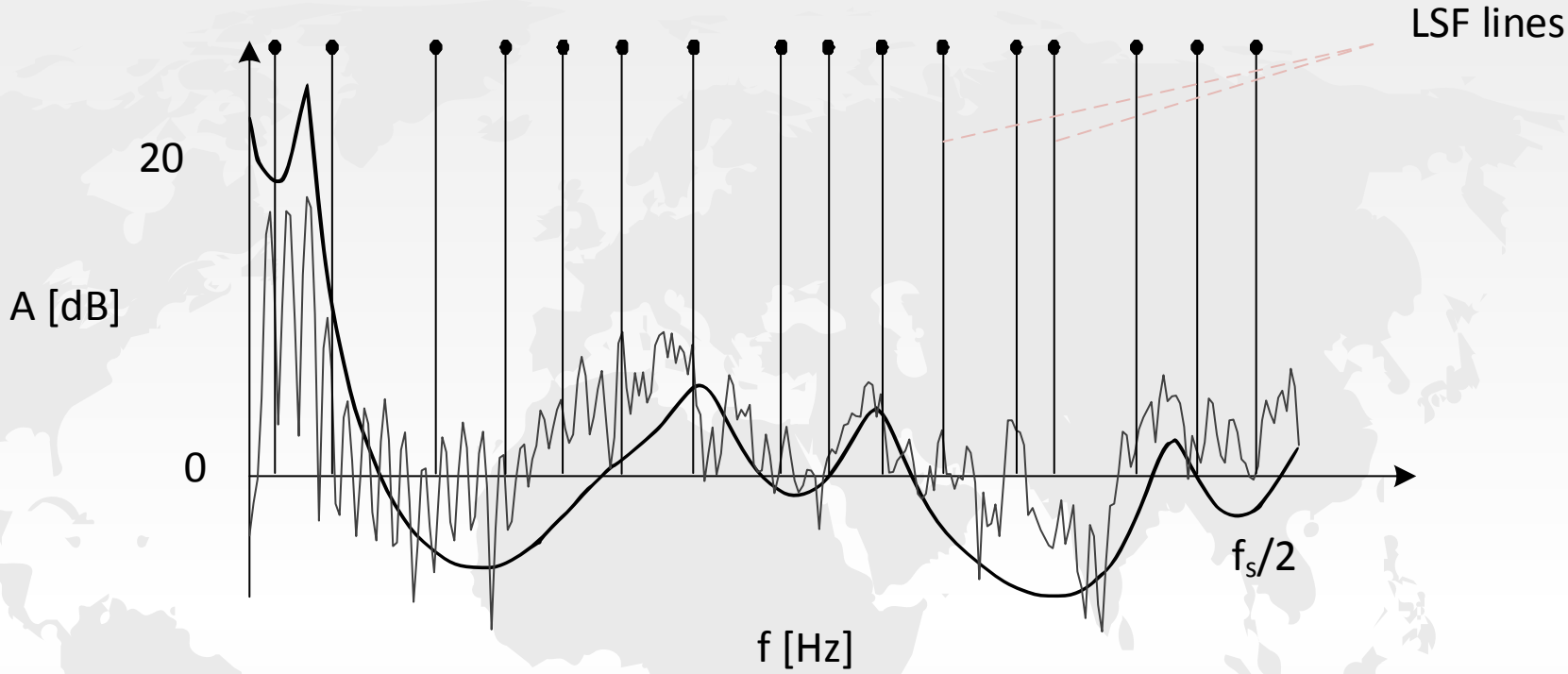
- Drtivá většina moderních řečových kodeků používaných na světě využívá LP analýzu
- LP analýza se provádí na krátkých úsecích vstupního signálu (5-30ms), kde lze signál považovat za stacionární
- chyba predikce:
  $e(n) = x(n) - [a_1x(n-1)+a_2x(n-2)+...+a_Px(n-P)]$
- A(z) lze chápat jako filtr (LP analysis filter), 1/A(z) je tzv. obálka (LP synthesis filter)
- Poloha pólů filtru 1/A(z) odpovídá formantům ve spektru
- Řád filtru P=15-20 postačuje bohatě pro řeč či šum, ne však např. pro hudbu
- pole-zero applet na http://web.mit.edu/6.302/www/pz/
- nebo na http://www.falstad.com/dfilter/index.html

LP filtrace



- Filtrace vstupního signálu filterm A(z) odstraňuje z řečového signálu jeho vlastní „obálku“ a tím zbavuje spektrum vlastních formantů (whitening, flattening)
- Reziduální (excitační) signál je to co po této operaci zbyde a to se dále kóduje

LP->LSF converze



- LP parametry získané na základě LP analýzy se převádí na LSP/LSF parametry pomocí tzv. Chebyshevových polynomů

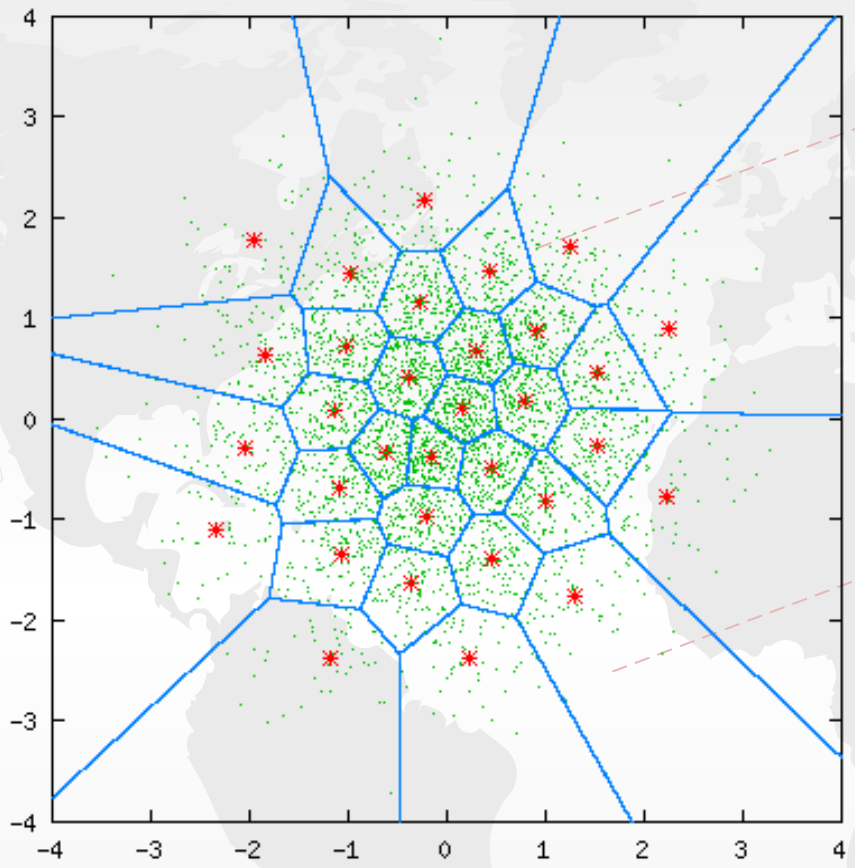
$$A(z) = 1 - \sum_{i=1}^{P} a_i z^{-i} = 0.5\left[P(z) + Q(z)\right]$$

$$P(z) = A(z) + z^{-(P+1)} A(z^{-1})$$

$$Q(z) = A(z) - z^{-(P+1)} A(z^{-1})$$

- Kořeny P(z) a Q(z) leží na jednotkové kružnici v komplex. rovině v konj. párech => koeficienty p_k a q_k jsou reálné
- Kořeny P(z) a Q(z) alternují
- P(z) a Q(z) lze vyjádřit pomocí Cheb. plynomů a iterativně zjistit jejich kořeny
- Nejlepší vysvětlení na http://www.ece.mcgill.ca/~pkabal/papers/1986/Kabal1986.pdf

LSF kvantizace

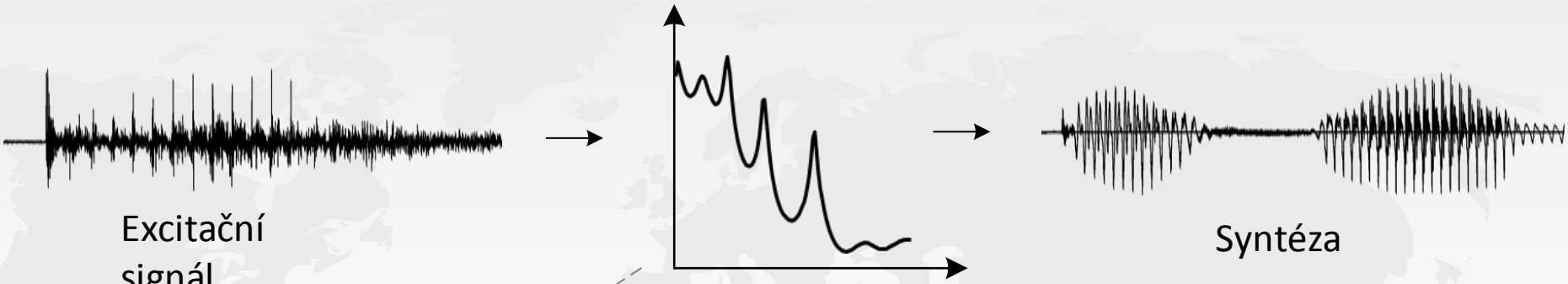


kódové slovo (codeword)

Voronoi region

- Udělá se kolekce LSF vektorů na trénovací databázi a ty jsou nějak rozmístěny v 15-rozměrném prostoru. Do něj se umístí Q reprezentativních kódových vektorů (codevector, centroid)
- Iterativním algoritmem zvaným k-means se pozice kódových vektorů zpřesňují
- Animace pohybu centroid ve 2D je např na http://www.data-compression.com/vqanim.shtml
- Je zapotřebí zhruba 20-30 bitů na jeden 16-ti dimenzionální LSF vektor

LP syntéza

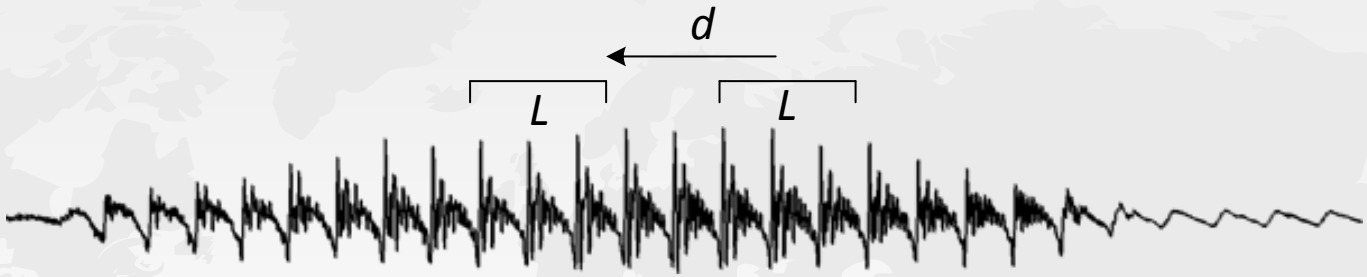


$$\frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{P} a_i z^{-i}}$$

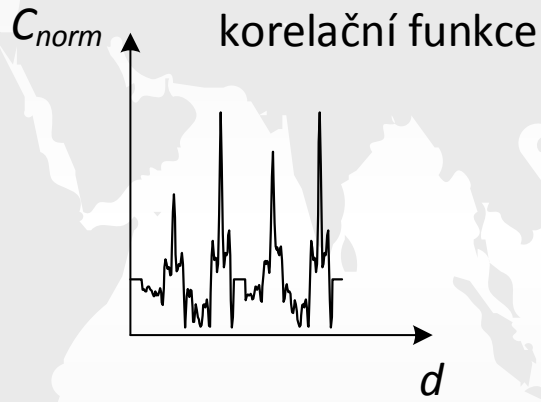
- Filtr LP syntézy je IIR filtr, takže má vlastní paměť, což je několik posledních vzorků z minulé syntézy
- Filtr LP syntézy má pouze póly, t.j. umí modelovat pouze spektrální „špičky“, nikoliv „zářezy“
- Vzhledem k charakteru filtru může dojít k jeho nestabilitě a „explozi“ syntézy

Analýza základního tónu

Analýza zákl. tónu



$$C_{norm}(d) = \frac{\sum_{n=0}^{L} s(n)s(n-d)}{\sqrt{\sum_{n=0}^{L} s^2(n) \sum_{n=0}^{L} s^2(n-d)}},$$

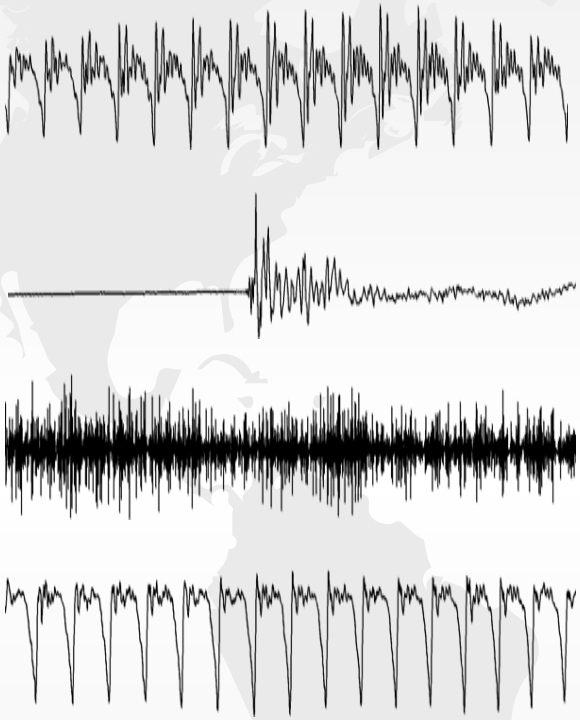


- korelační funkci počítáme pro d = [20;230] vzorků, což odpovídá F = [47;800] Hz
- vybereme první „velké“ maximum, to prohlásíme za OL pitch
- snažíme se vyhnout násobkům OL pitch
- zapamatujeme si hodnotu $C(d_{max})$, to prohlásíme za OL voicing
- kuriozita: extremely low voice https://www.youtube.com/watch?v=AaPtiFO-NLc

Klasifikace

Třídy (kategorie) řečových signálů

řečový signál



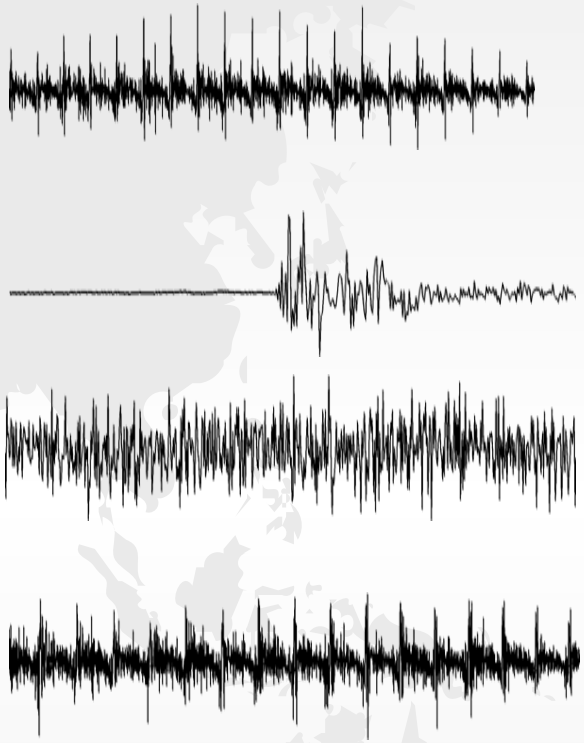
VOKÁLY (znělé úseky)
a, e, i, o, u

PLOZÍVY
t, d, k, g

FRIKATIVY (neznělé úseky)
s, z, š, h, ch, v, f

NEPÁROVÉ KONSONANTY
m, n, j, l, r

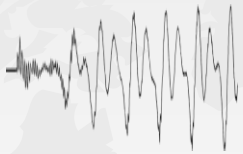
excitace



- můžete sami vyzkoušet, že při vyslovení frikativ se vám nechvějí hlasivky (glottis)
- kodek provádí klasifikaci signálů na ZNĚLÉ, NEZNĚLÉ a OSTATNÍ
- podle toho pak vybírá vhodný model pro kódování excitačního signálu
- základními modely jsou: VOICED, UNVOICED a GENERIC


CELP

Koncept analýzy syntézou (analysis-by-synthesis)



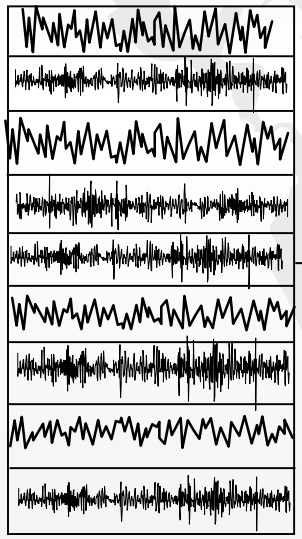
$s(n)$

1 vstupní rámec řeči
(např. 320 vzorků pro 16kHz
signál odpovídá 20ms)

Chyba (error)

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

k



G

$\frac{1}{A(z)}$

$\hat{s}(n)$

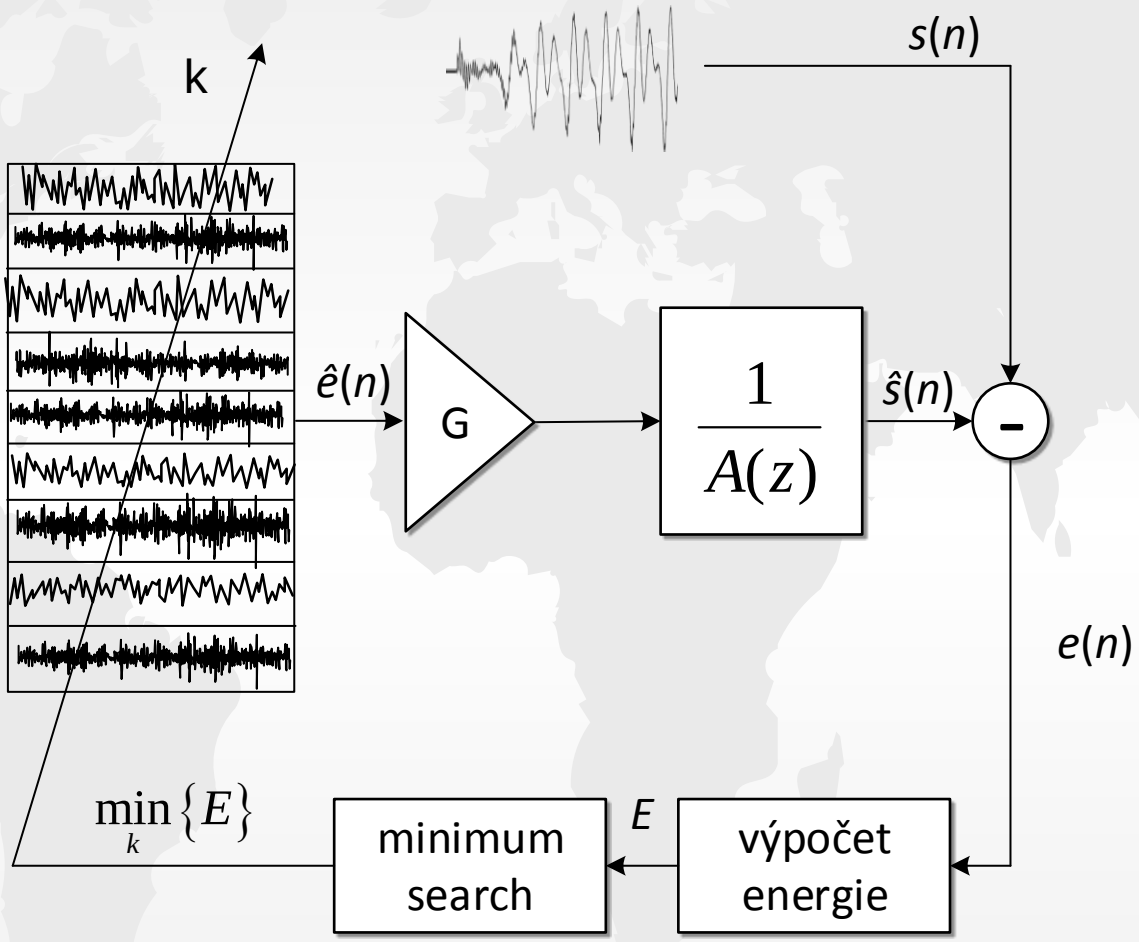
Zesílení
(gain)

Knihovna (codebook)
excitačních signálů

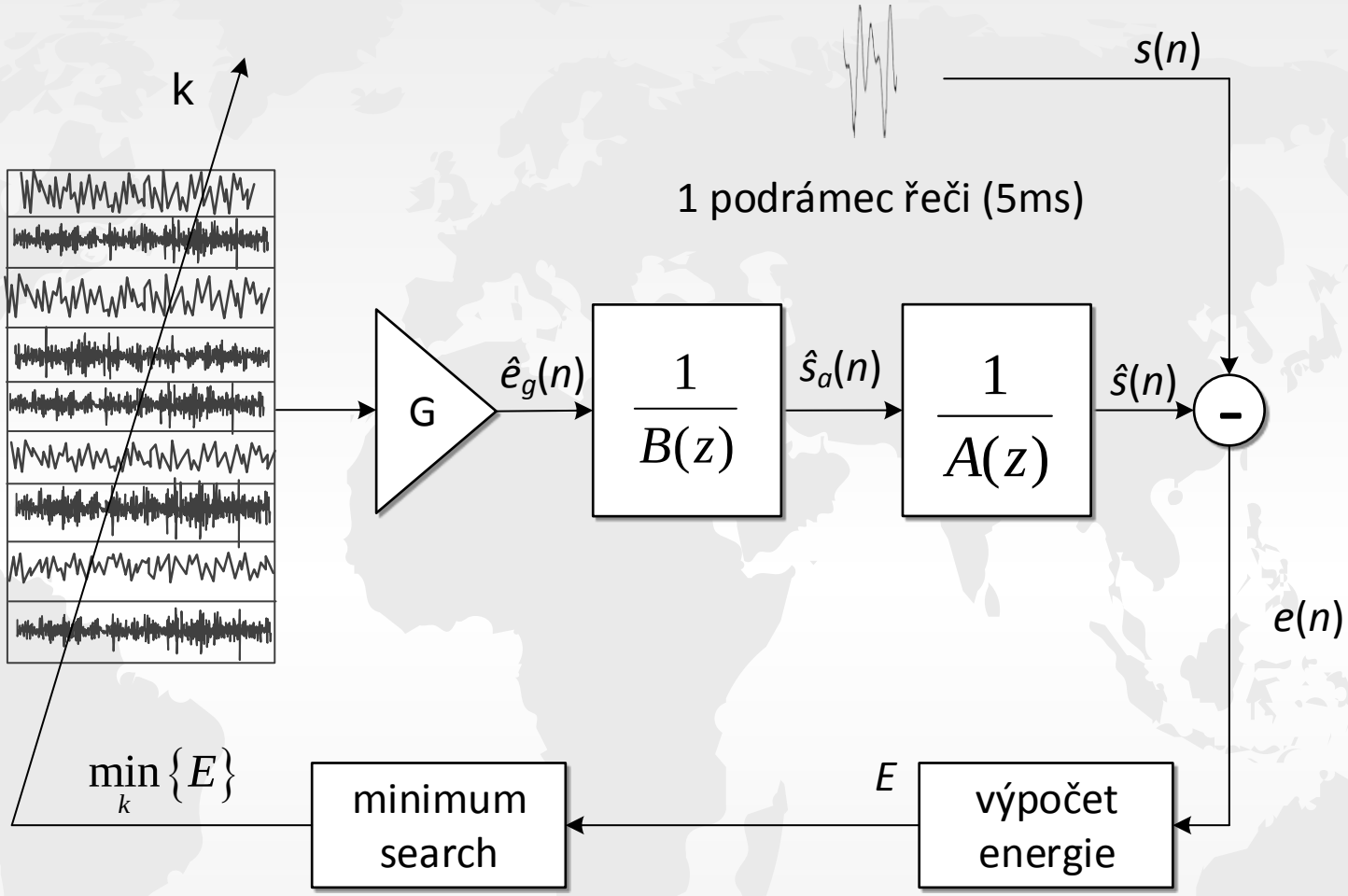
- Snažíme se minimalizovat chybu
- Nejčastěji

$$\min_k \left\{ \sum_n \left[s(n) - \hat{s}(n) \right]^2 \right\}$$

Koncept analýzy syntézou (jinak nakresleno)

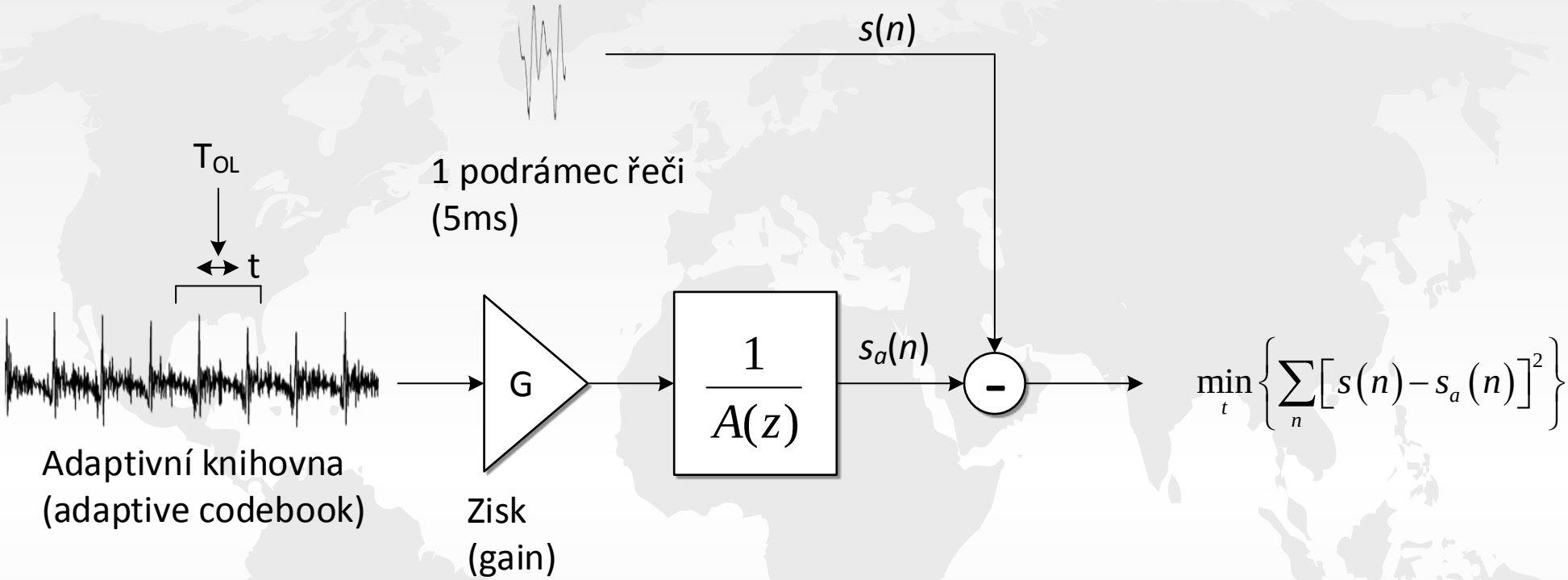


Dlouhodobý prediktor (LTP) a zavedení podrámců



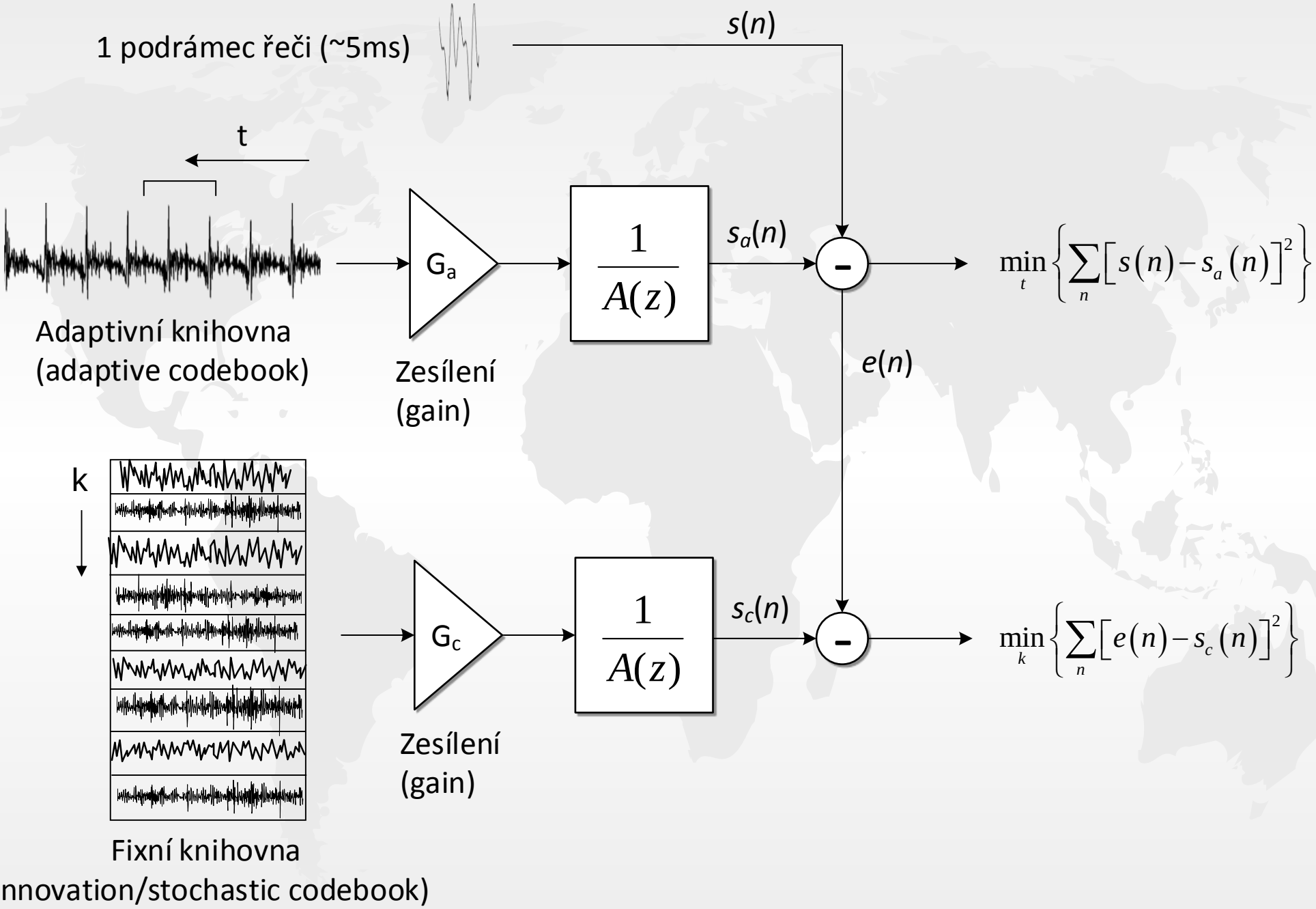
- dlouhodobý prediktor (LTP) je filtr tvaru $1/(1+bz^{-T})$
- Takže platí: $\hat{s}_a(n) = \hat{e}_g(n) - b\hat{s}_a(n-T)$ (syntéza)
- Nebo taky: $\hat{e}_g(n) = \hat{s}_a(n) + b\hat{s}_a(n-T)$ (analýza)
- říká se mu sice „dlouhodobý“ prediktor, protože prediktuje z vzorků vzdálených až 20ms, ale zákl. tón řeči T je třeba estimovat v kratších úsecích (podrámcích), typicky 5ms

Zavedení adaptivní knihovny



- dlouhodobý prediktor lze nahradit tzv. adaptivní knihovnou, která je tvořena úseky minulého excitačního signílu
- hledá se pak nejlepší úsek minulé excitace, který by mohl reprezentovat svým tvarem současný podrámec řeči
- korelace a minimalizace kvadrátu chyby okolo T_{OL} (open-loop pitch)
- zisk (gain) najdeme opět minimalizací chyby
- prohledávání adaptivní knihovny se dělá na residuálním signálu, ne na vstupním signálu!!

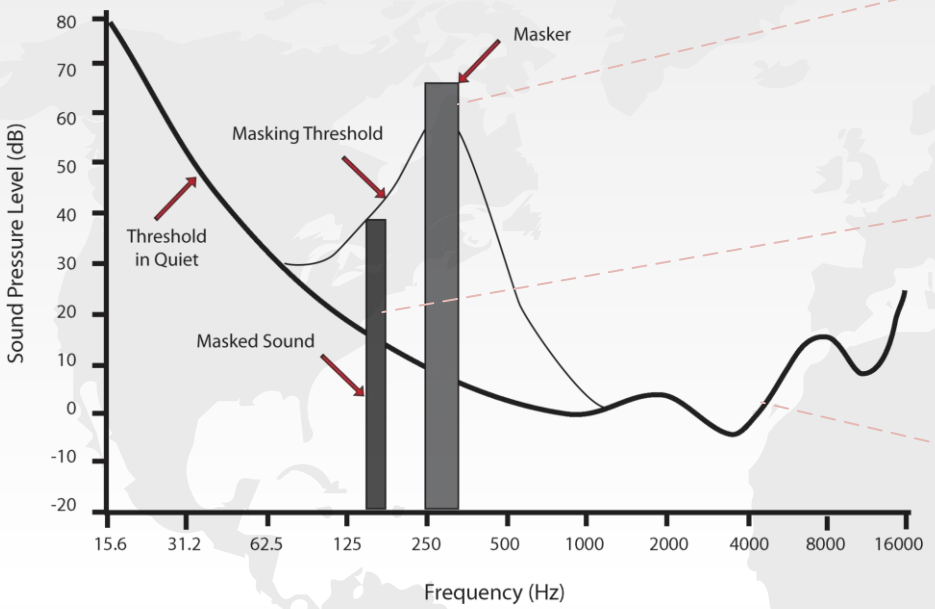
Zavedení fixní knihovny



$$\min_t \left\{ \sum_n \left[s(n) - s_a(n) \right]^2 \right\}$$

$$\min_k \left\{ \sum_n \left[e(n) - s_c(n) \right]^2 \right\}$$

Maskování tónů



maskovací šum o centralní frekvenci 300Hz a šírce pásma 100Hz a úrovni 65 dB

maskovaný signál o centralní frekvenci 200Hz a šírce pásma 50Hz a úrovni 38 dB

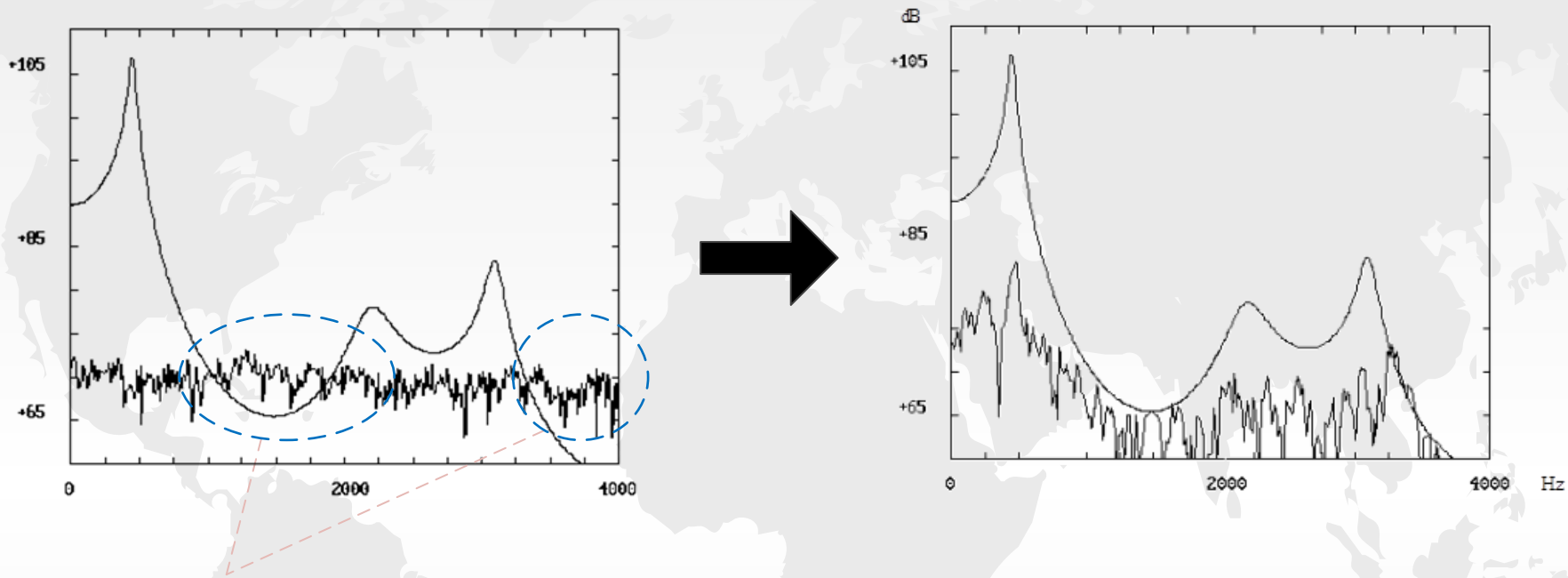
práh slyšitelnosti

DŮSLEDKY:

- tóny v blízkosti silného tónu jsou maskovány
- spektrální komponenty s úrovní pod prahem slyšitelnosti není třeba kódovat
- lze tolerovat vyšší úroveň kvantizačního šumu v blízkosti silných tónů, např. formantů

- demo na https://www.youtube.com/watch?v=k6DVywW5NR4

Maskování kvantizačního šumu



tohle bude slyšet

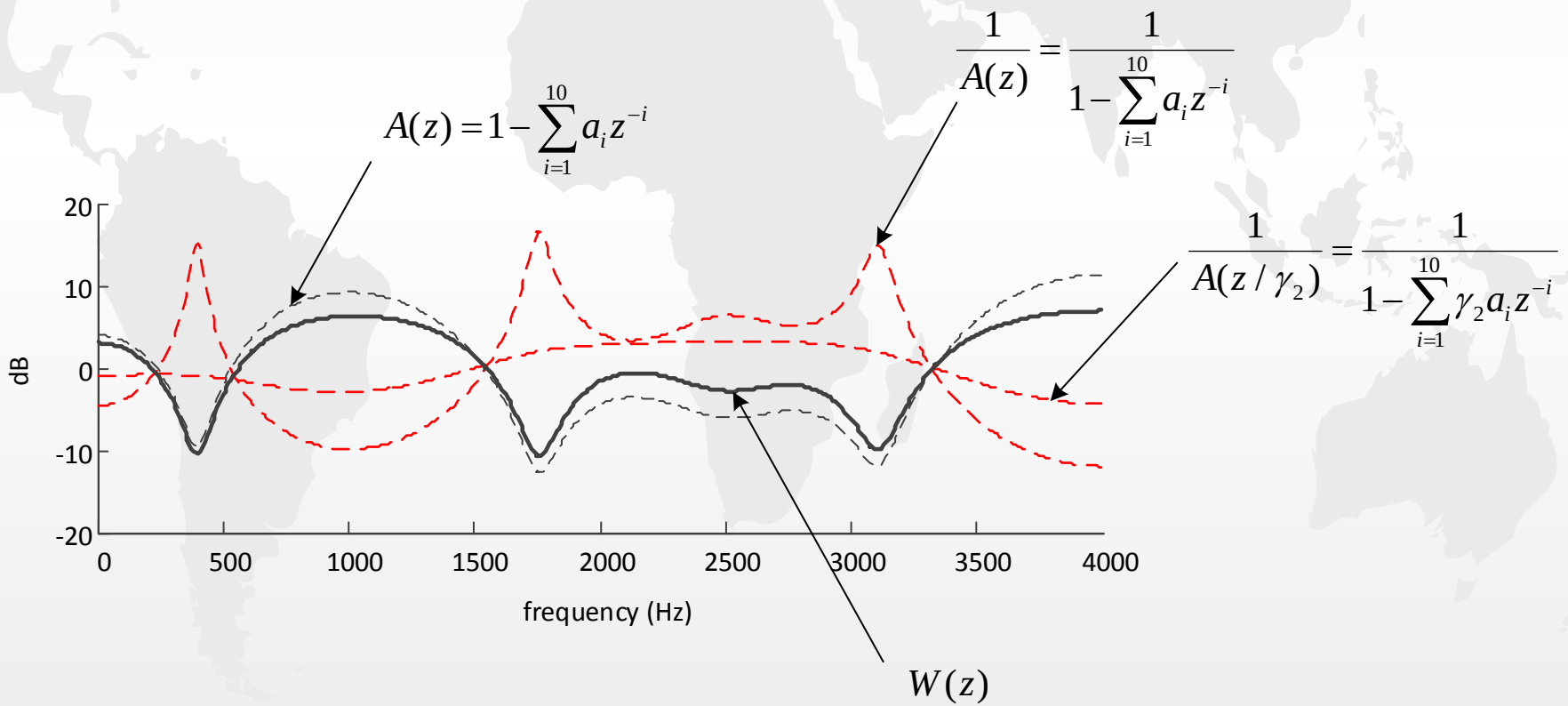
chtěli bychom, aby kvantizační šum nějak vypadal takhle

- Když je lidské ucho necitlivé na některé změny ve spektru, proč toho nevyužít a neaplikovat na chybu predikce (tzv. kvantizační šum)?
- kvantizační šum = chyba mezi orig. a syntetickou řečí
- můžeme si dovolit ho zvýšit v oblasti formantů a potlačit v „údolí“

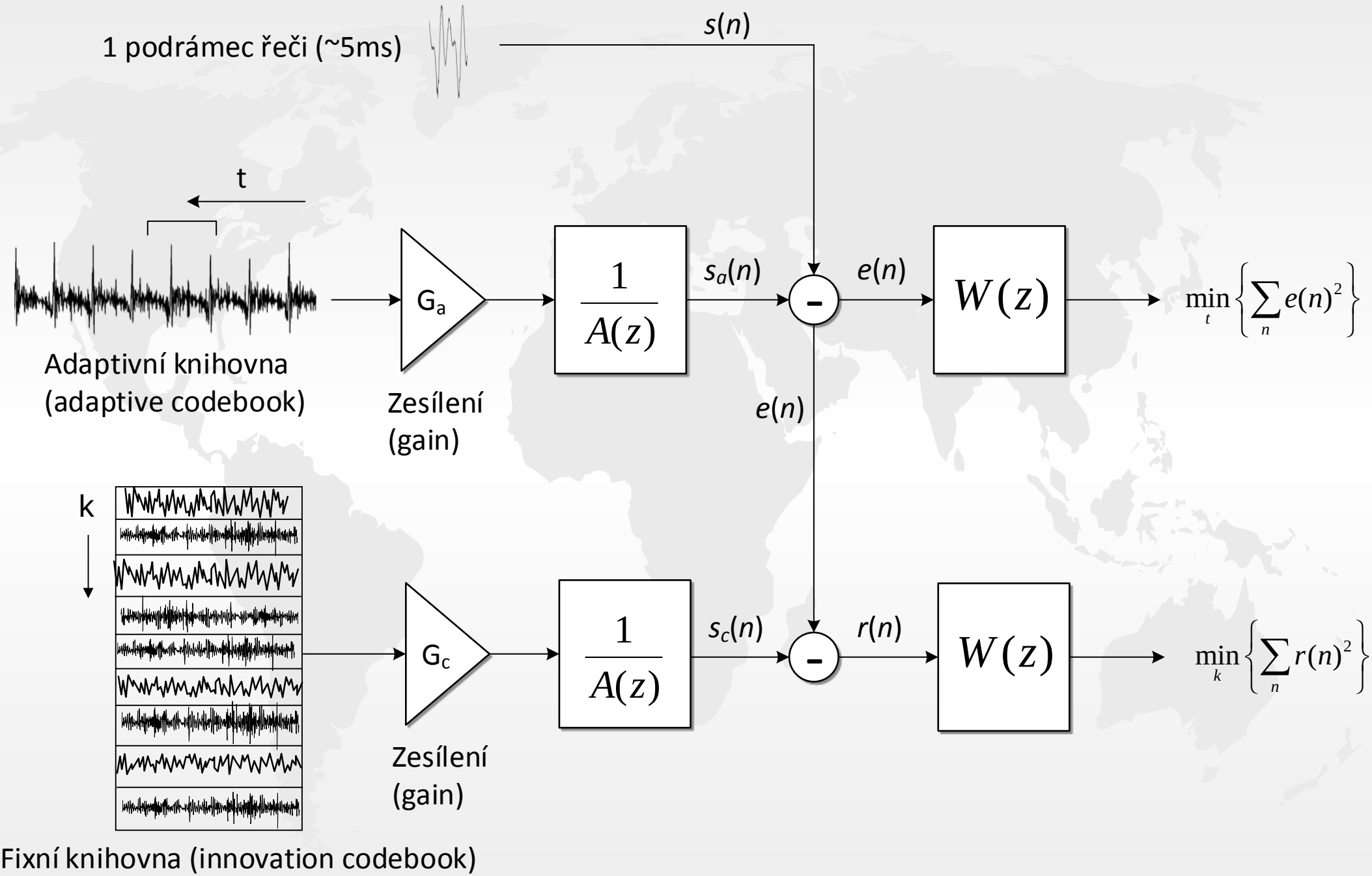
Perceptuální váhování

$$W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)} = \frac{1 - \sum_{i=1}^{10} \gamma_1^{\,i} a_i z^{-i}}{1 - \sum_{i=1}^{10} \gamma_2^{\,i} a_i z^{-i}}$$

- Musíme najít takový tvar perceptuálního filtru, který zvýrazní v kantizačním šumu ta frekvenční pásma, na která je naše ucho citlivé a potlačí ta, na které je necitlivé
- celkový zisk filtru musí být 1, jinak by hrozilo umělé zesilování signálu
- Nejlépe je využít samotného filtru A(z) jako základ
- γ_1= 0.94, γ_2= 0.6 (adaptivní)



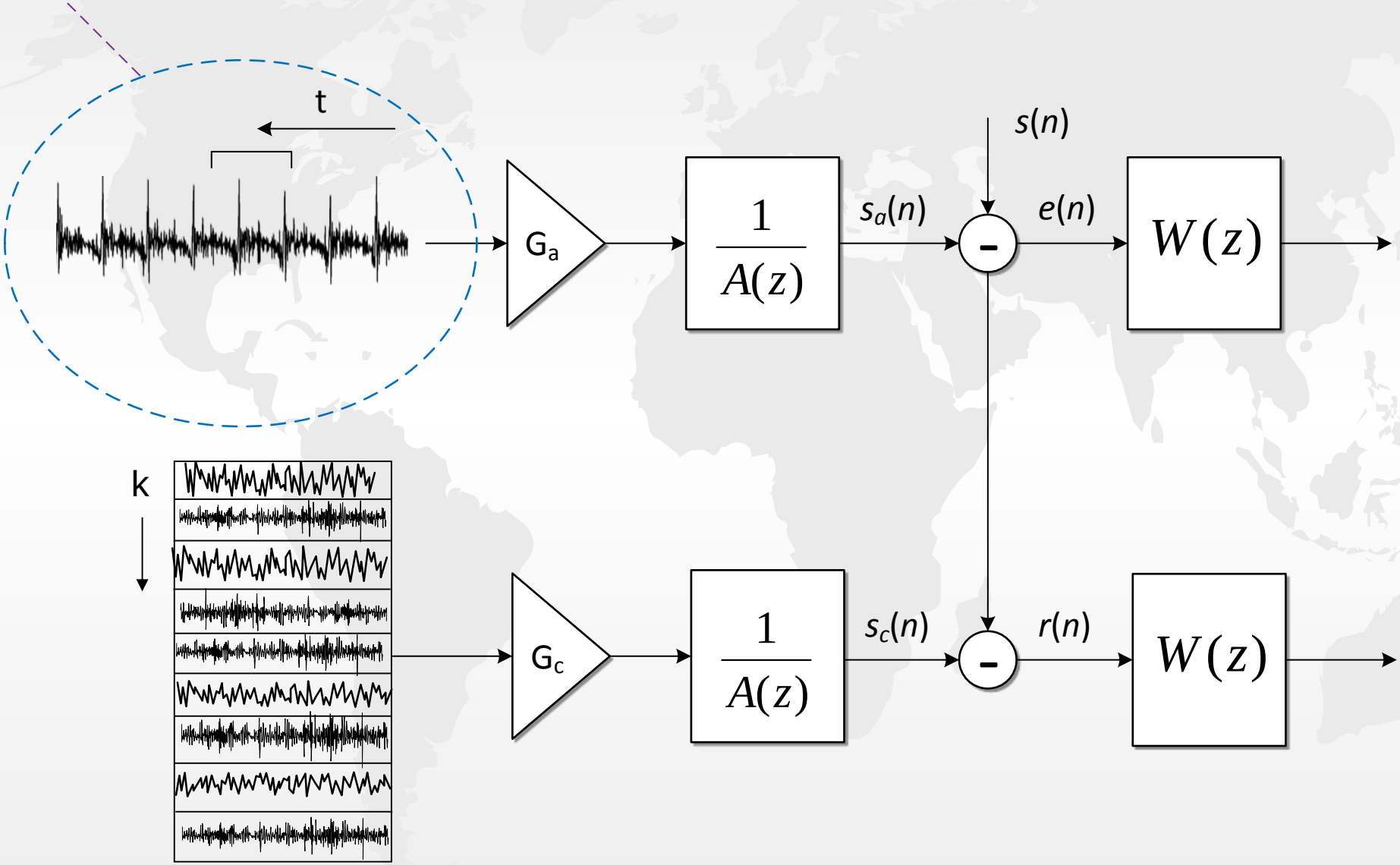
Zavedení perceptuálního filtru



Výpočetní náročnost

Adaptivní knihovna: (7-9 bitů, t.j. 128 – 512 vektorů)

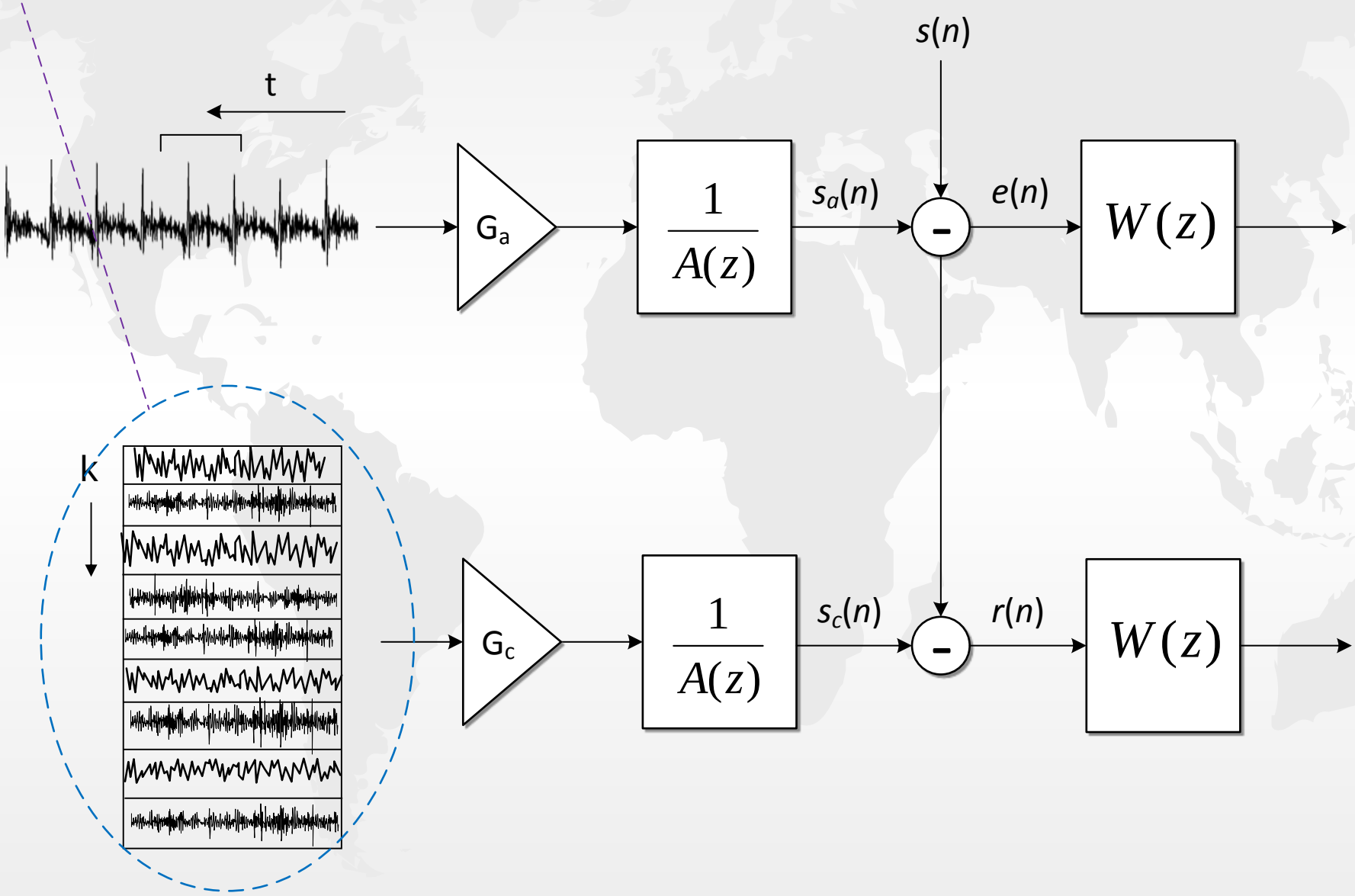
(estimace periodicity předem (OL pitch analysis), prohledávání knihovny kolem této hodnoty)



Výpočetní náročnost

Fixní knihovna: (10-88 bitů, t.j. 1024 – 3.10^{26} vektorů)

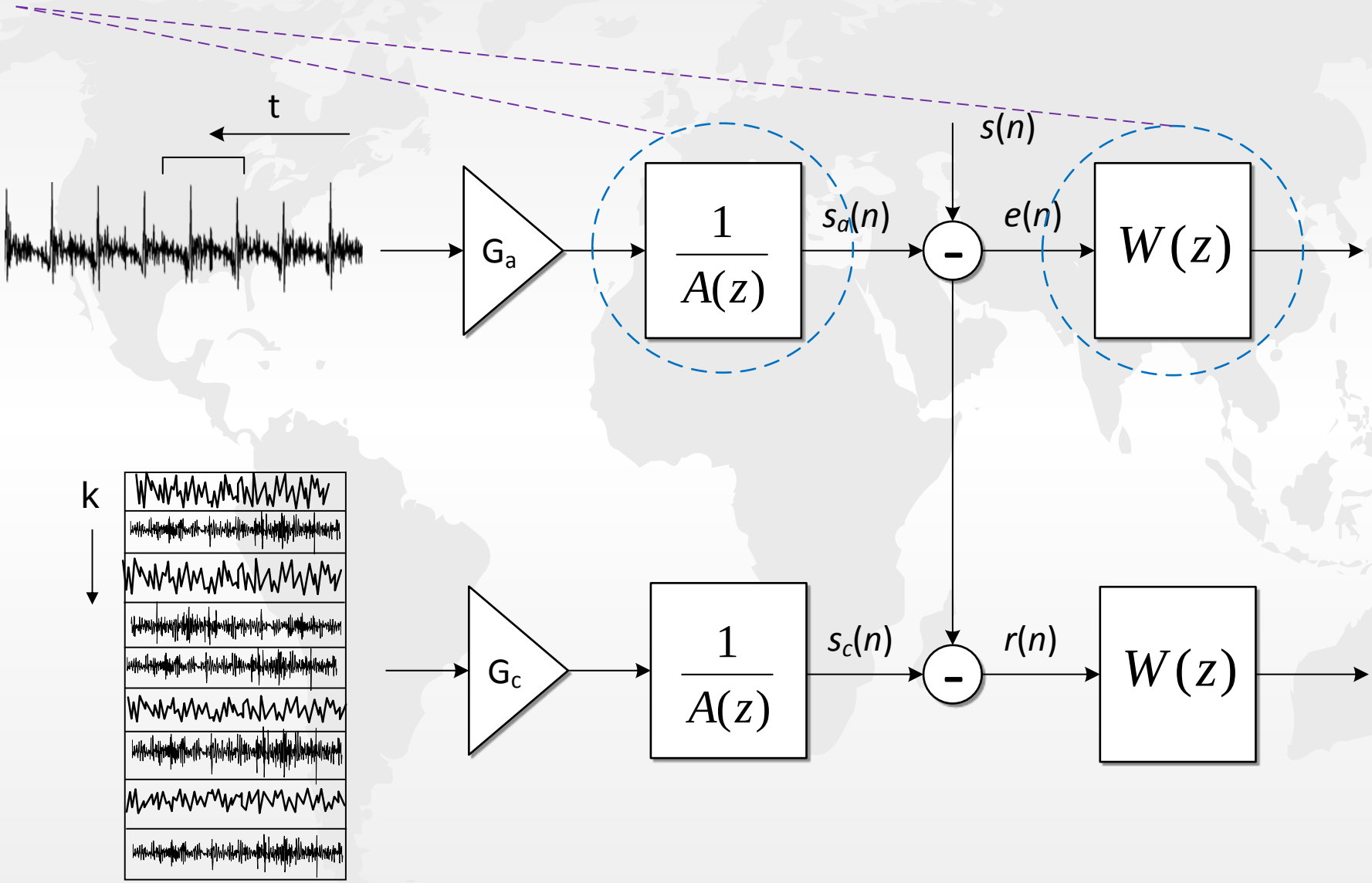
(vnucení jednoduchých struktur – pouze několik pulzů na stopu, omezený počet pozic pulzů, znaménka)



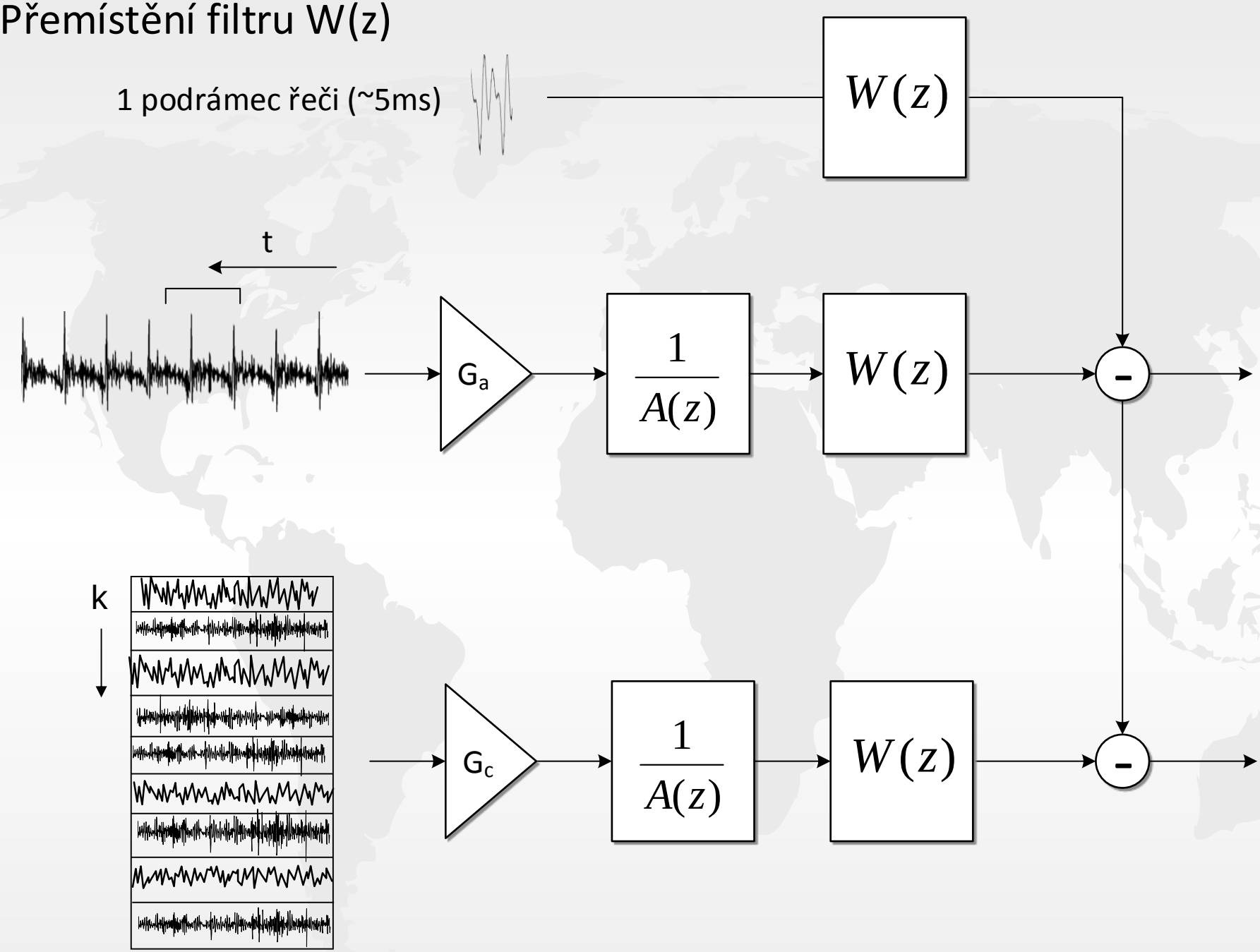
Výpočetní náročnost

Filtrace: (pro každý codevector nutno provést konvoluci s filtry 16.řádu)

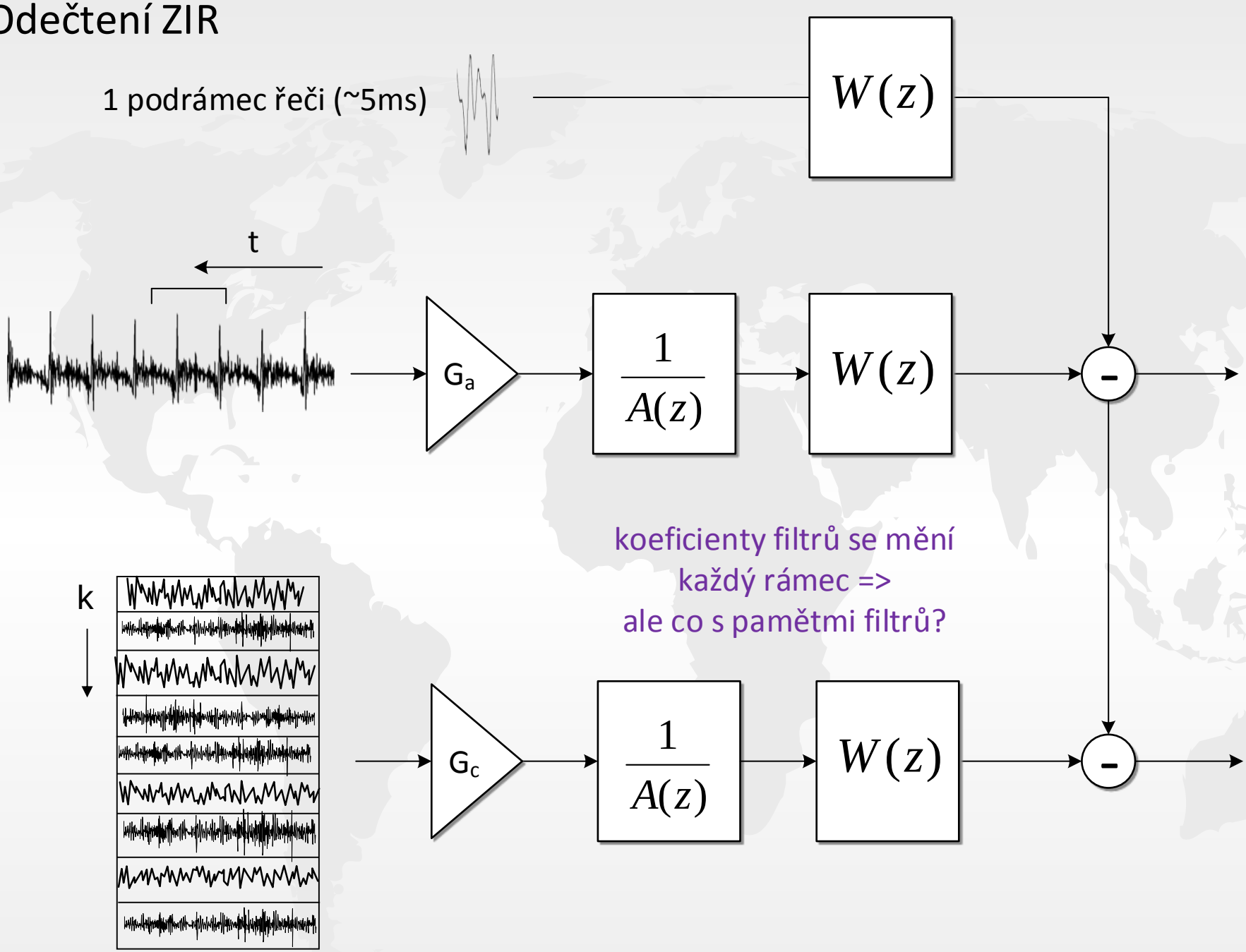
(nahrazení filtrů 1/A(z) a W(z) jejich impulzní odezvou) – viz dále



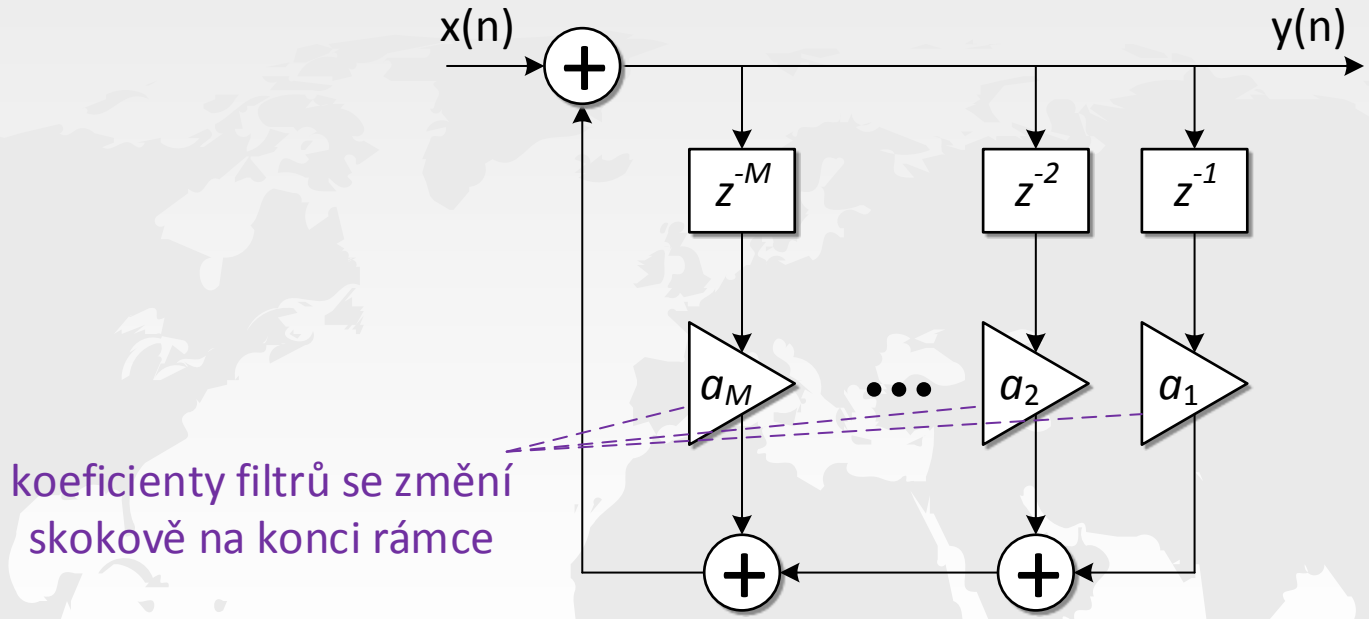
Přemístění filtru W(z)



Odečtení ZIR

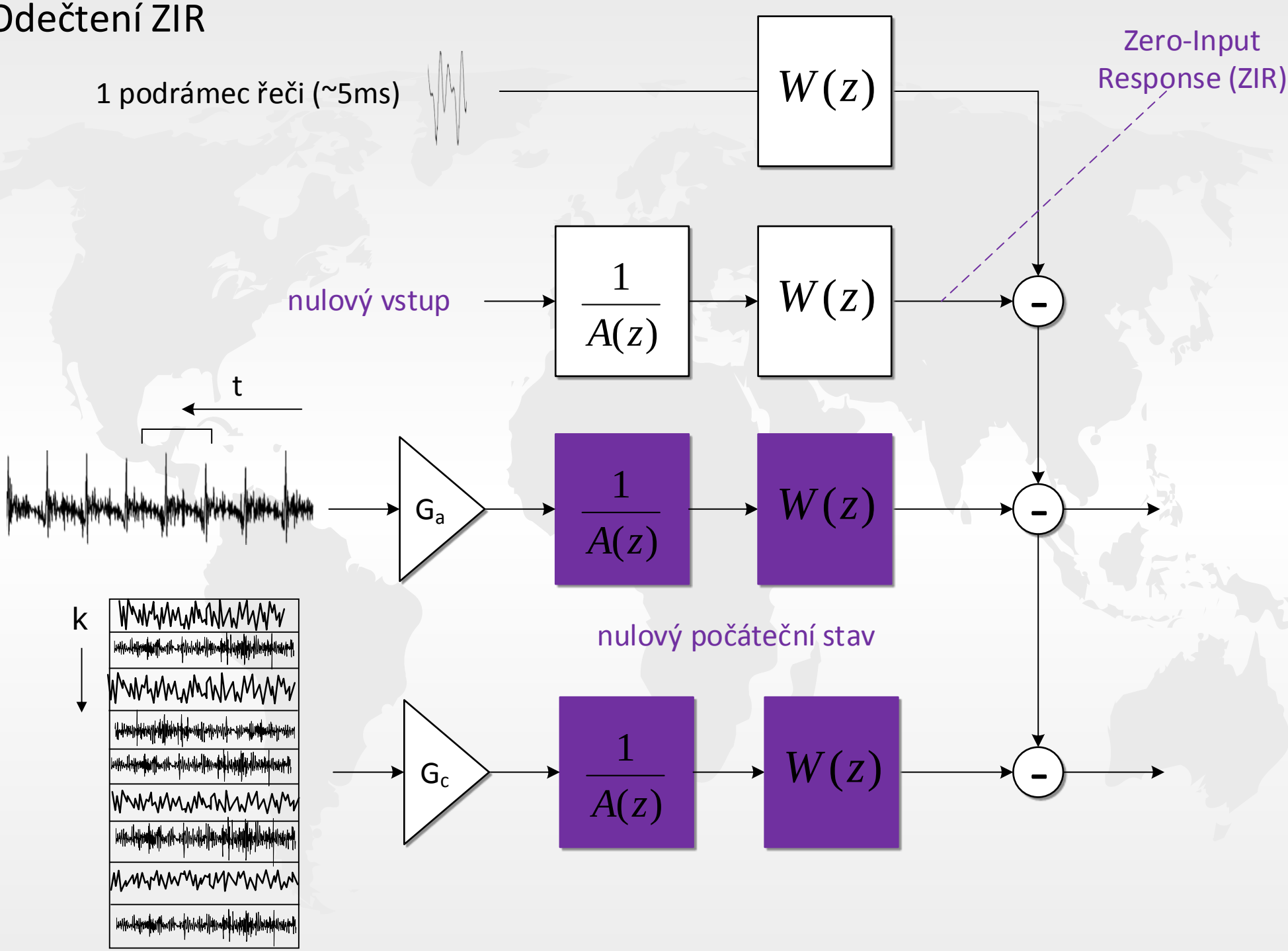


Odečtení ZIR

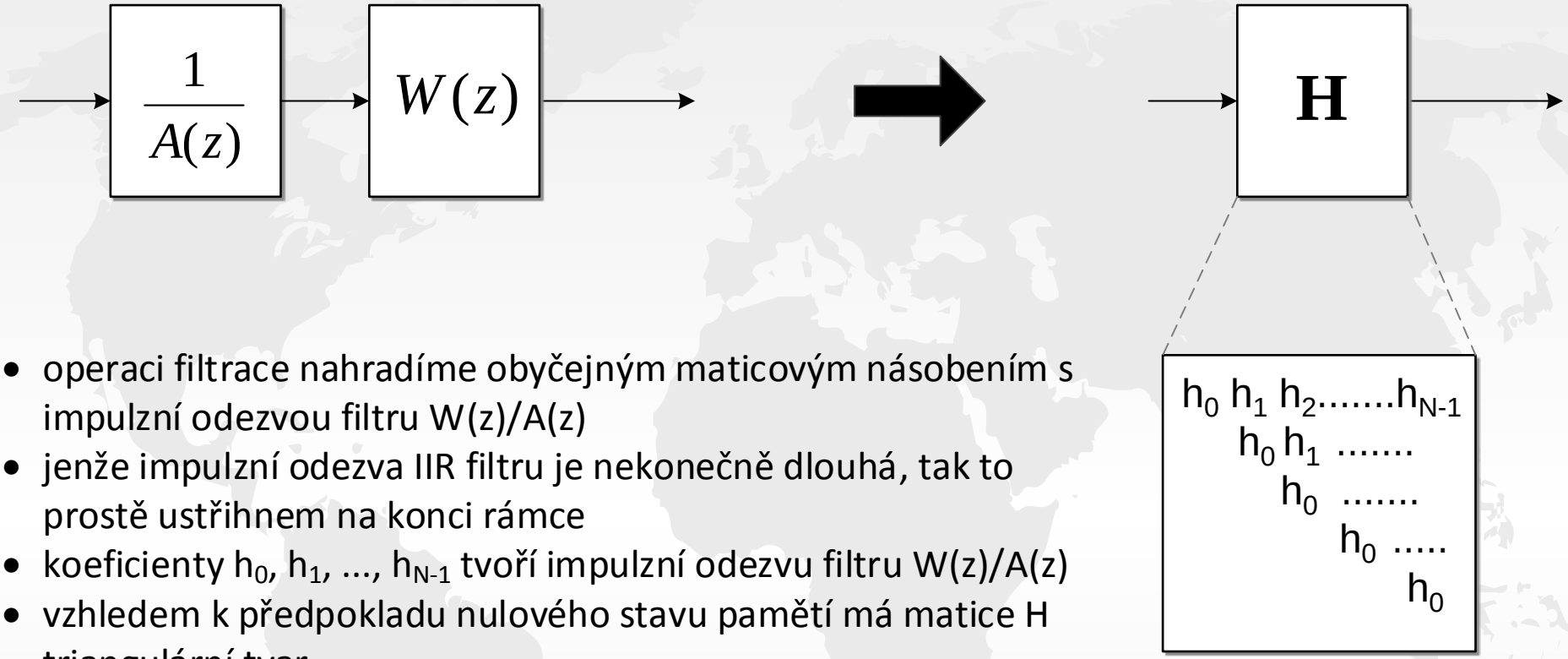


- pokud nic neuděláme, tak první vzorek v novém rámci bude filtrován nesprávně v důsledku špatných pamětí, v synt. signálu vzniknou skoky (discontinuity)
- musíme nechat filtr „doběhnout“ tím, že do něj dáme nulový signál a zaznamenáme jeho odezvu – tzv. zero-input response (ZIR)
- v novém rámci prostě paměti filtru vynulujeme a odečteme ZIR, tím chybu vynulujeme

Odečtení ZIR

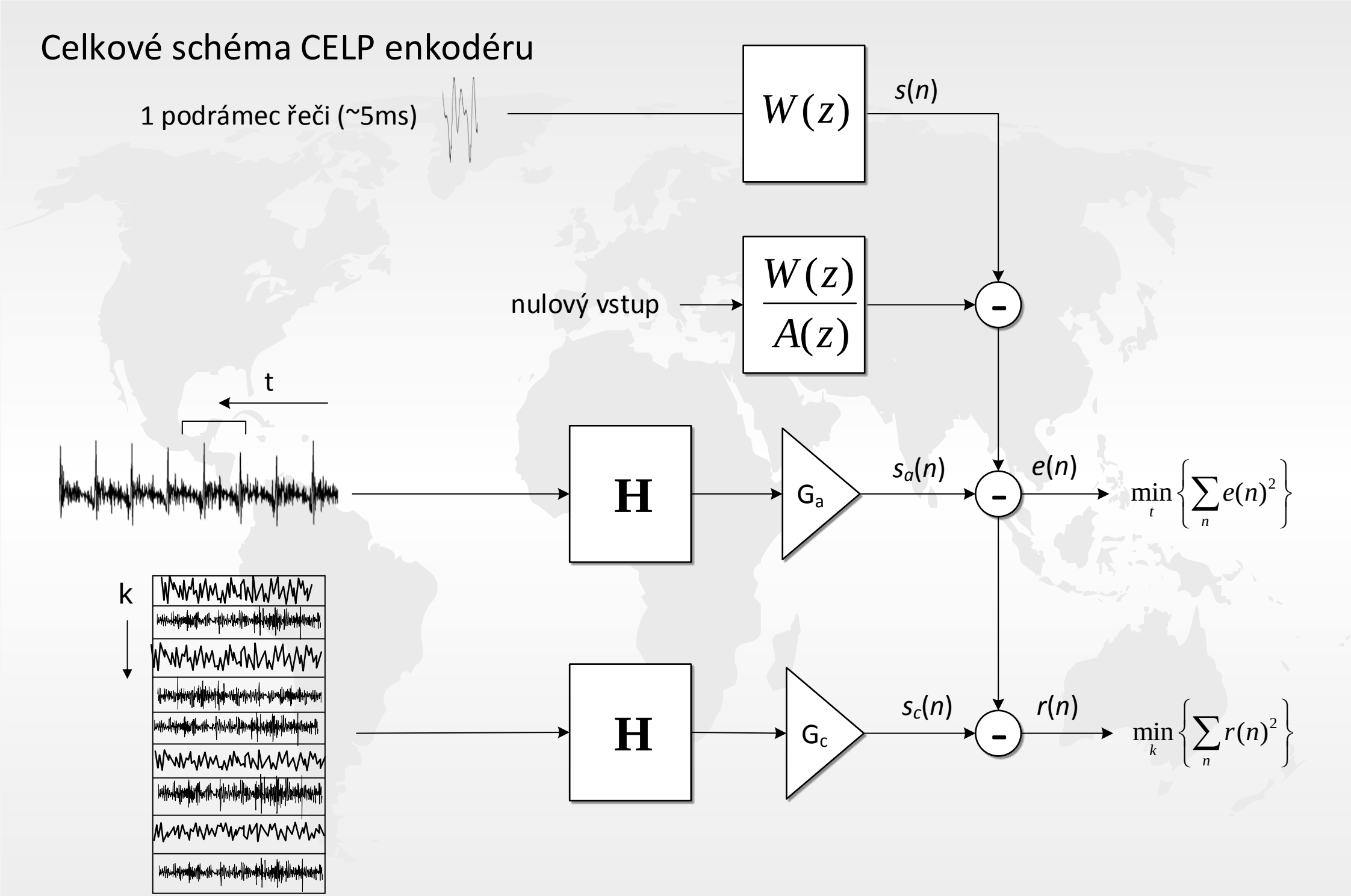


Nahrazení filtrů impulzní odezvou

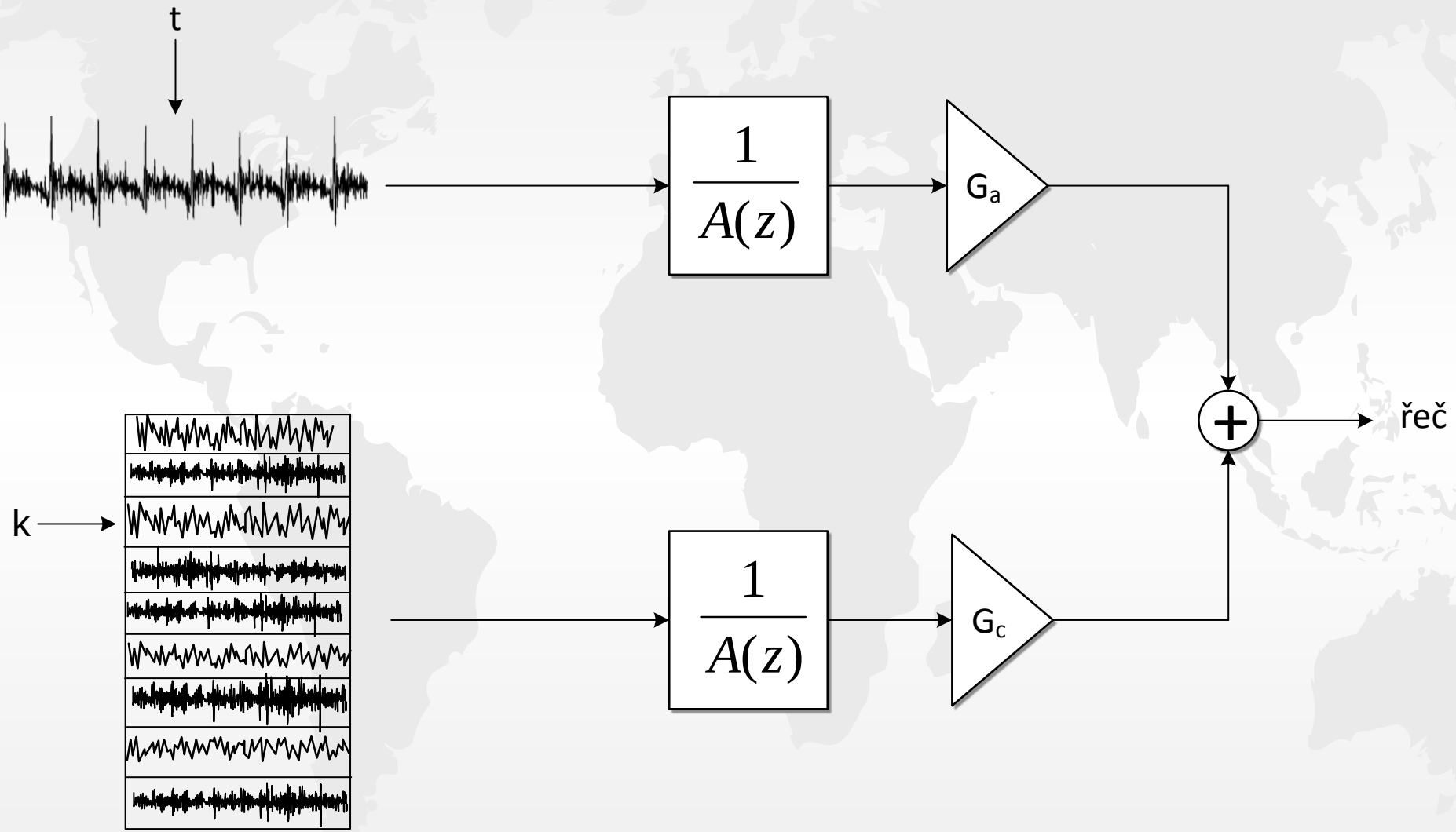


- operaci filtrace nahradíme obyčejným maticovým násobením s impulzní odezvou filtru W(z)/A(z)
- jenže impulzní odezva IIR filtru je nekonečně dlouhá, tak to prostě ustřihnem na konci rámce
- koeficienty h_0, h_1, ..., h_{N-1} tvoří impulzní odezvu filtru W(z)/A(z)
- vzhledem k předpokladu nulového stavu pamětí má matice H triangulární tvar

Celkové schéma CELP enkodéru



Celkové schéma CELP dekodéru



ACELP

ACELP

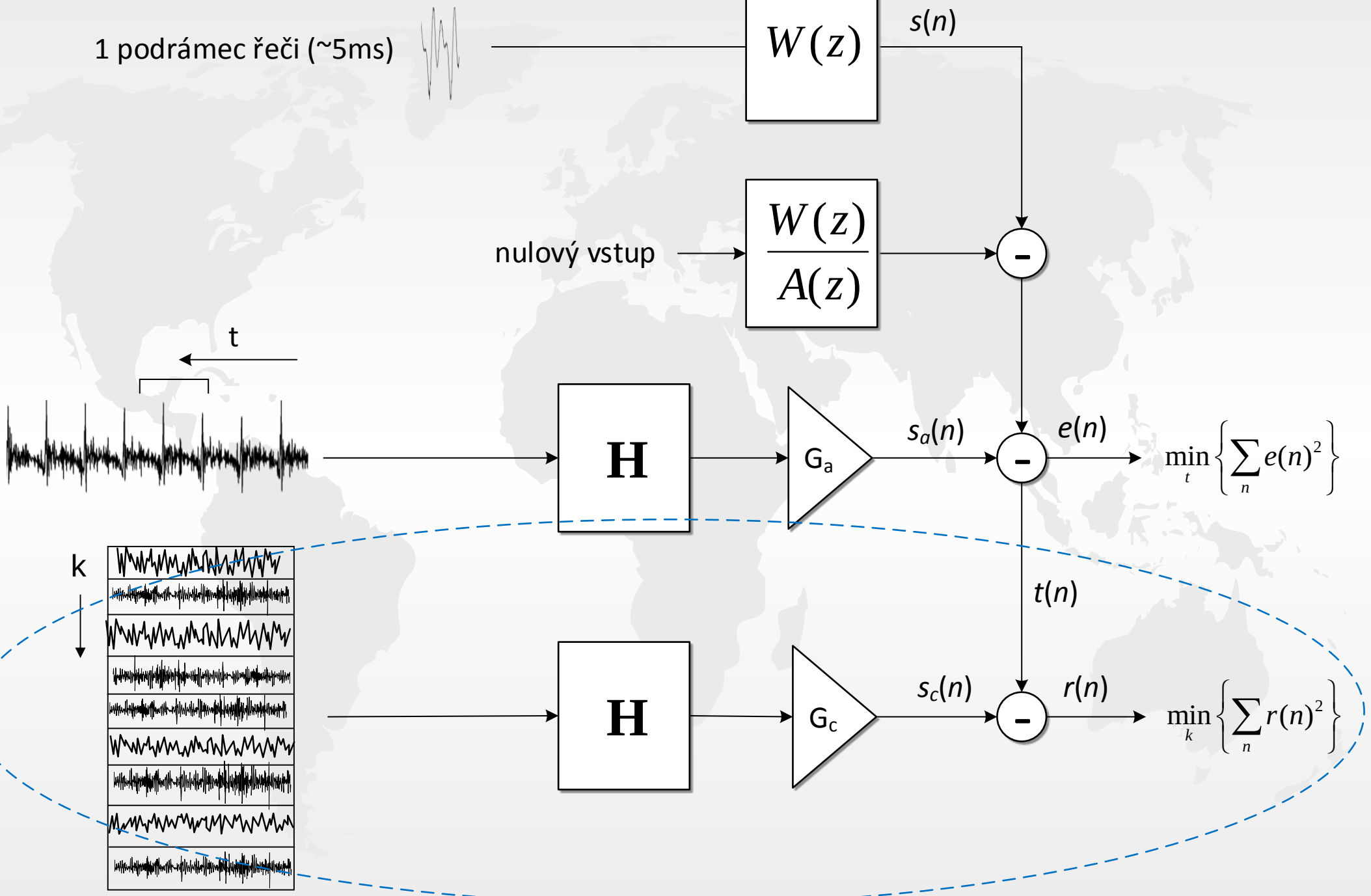
- ACELP® je patentovanou technologií VoiceAge Corp. a Université de Sherbrooke, CANADA
- vyvinuto v roce 1989 (Jean-Pierre Adoul, Claude Laflamme, Redwan Salami, Bruno Bessette)

UNIVERSITÉ DE SHERBROOKE

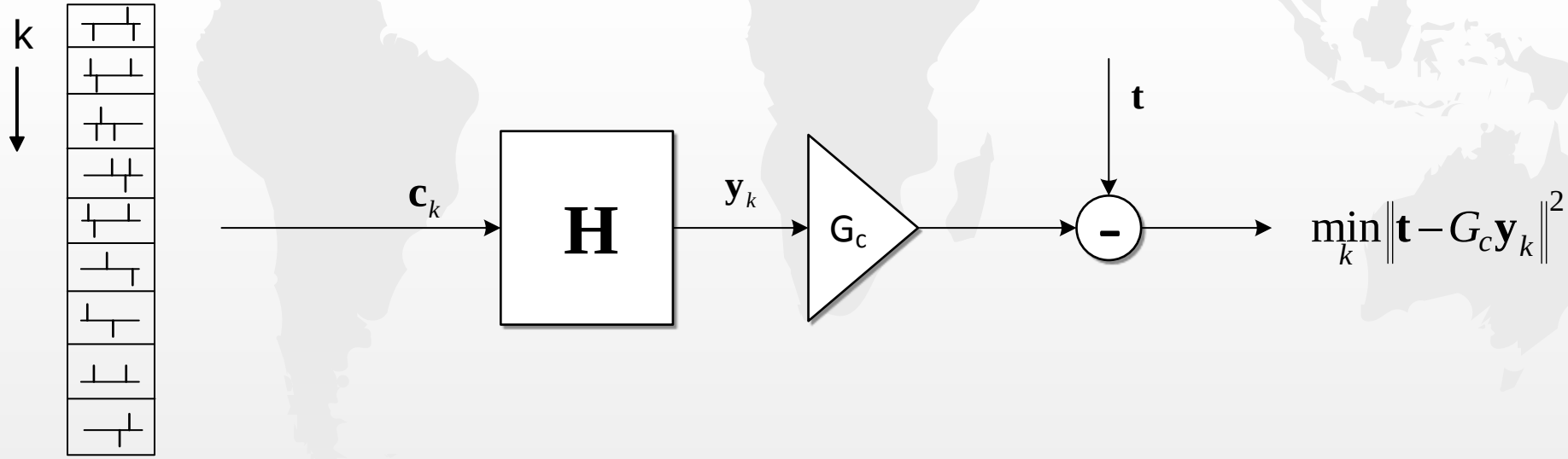
VoiceAge®
www.voiceage.com
The World's Premier Supplier of Speech and Audio Codecs

- kouzlo ACELPu spočívá v tom, že dokáže nahradit „obří“ fixní knihovnu signálů jednoduchou knihovnou s algebraickou strukturou, kde je jen několik málo pulzů v přesně definovaných pozicích, a tím zredukovat paměťovou a výpočetní náročnost
- technologii ACELP využívá cca
  - 2,4 miliard uživatelů mobilních telefonů na celém světě
  - 35 milionů uživatelů přehrávačů MP3
  - 500 milionů uživatelů internetových přehrávačů RealPlayer nebo MediaPlayer

Prohledávání fixní knihovny



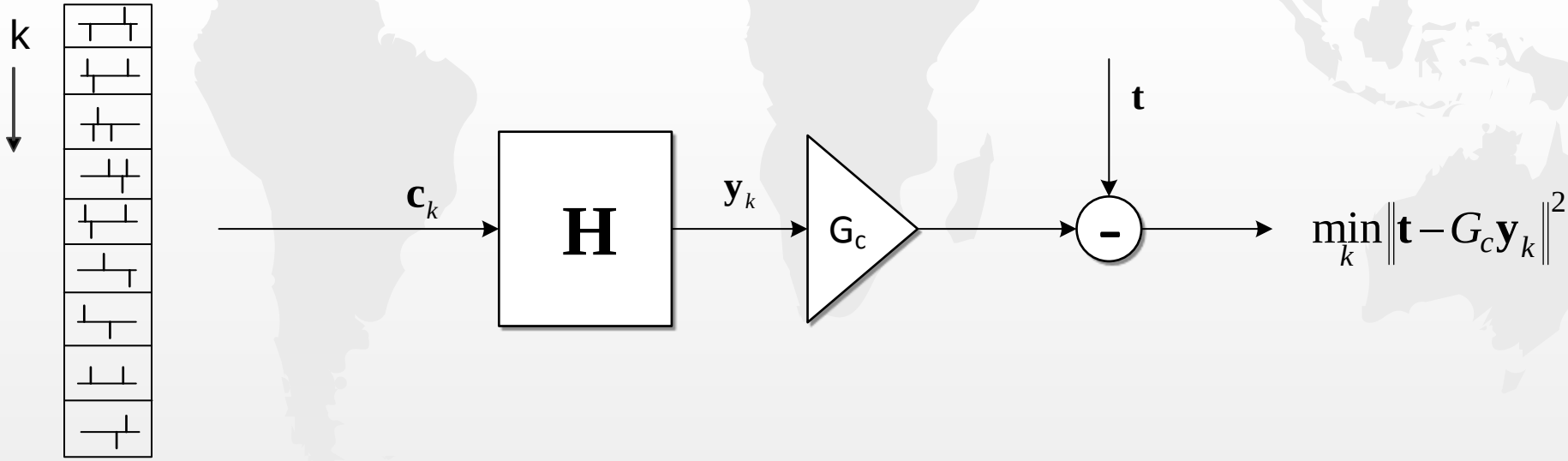
Zavedení algebraické knihovny



algebraická knihovna (až 80 bitů)

Prohledávání fixní knihovny

$$\min_k \left\| \mathbf{t} - G_c \mathbf{y}_k \right\|^2 \xrightarrow{\frac{\partial}{\partial G_c}}$$



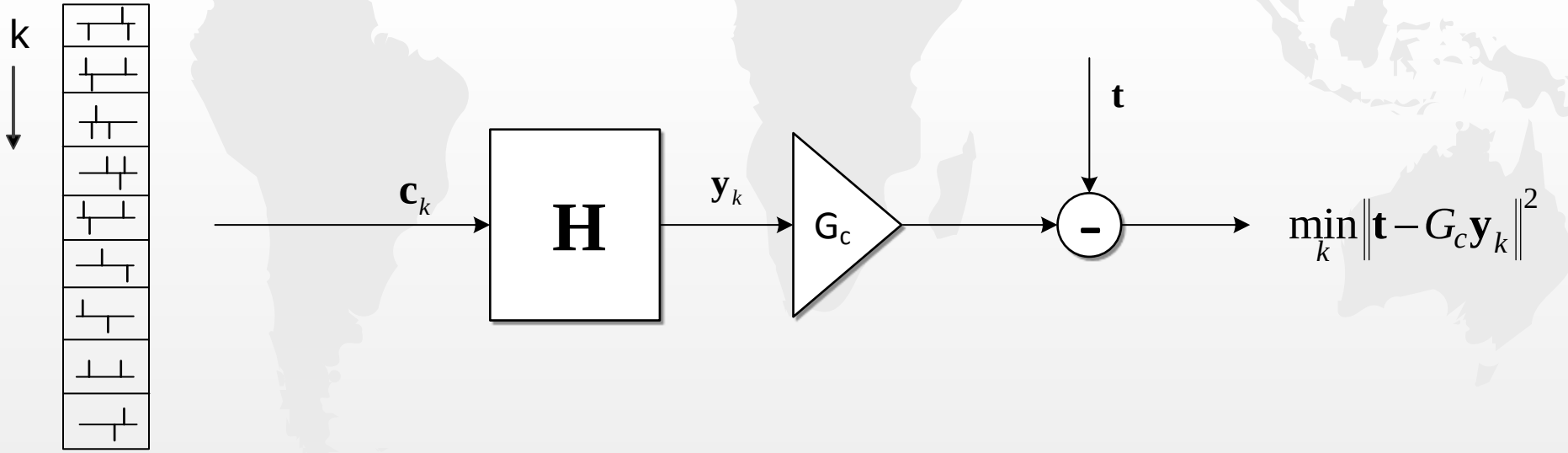
algebraická knihovna (až 80 bitů)

Prohledávání algebraické knihovny

$$\min_k \left\| \mathbf{t} - G_c \mathbf{y}_k \right\|^2 \xrightarrow{\frac{\partial}{\partial G_c}} \max_k \frac{\mathbf{t}^T . \mathbf{y}_k}{\mathbf{y}_k^T . \mathbf{y}_k}$$

korelace mezi cílovým (target) vektorem a testovaným vektorem

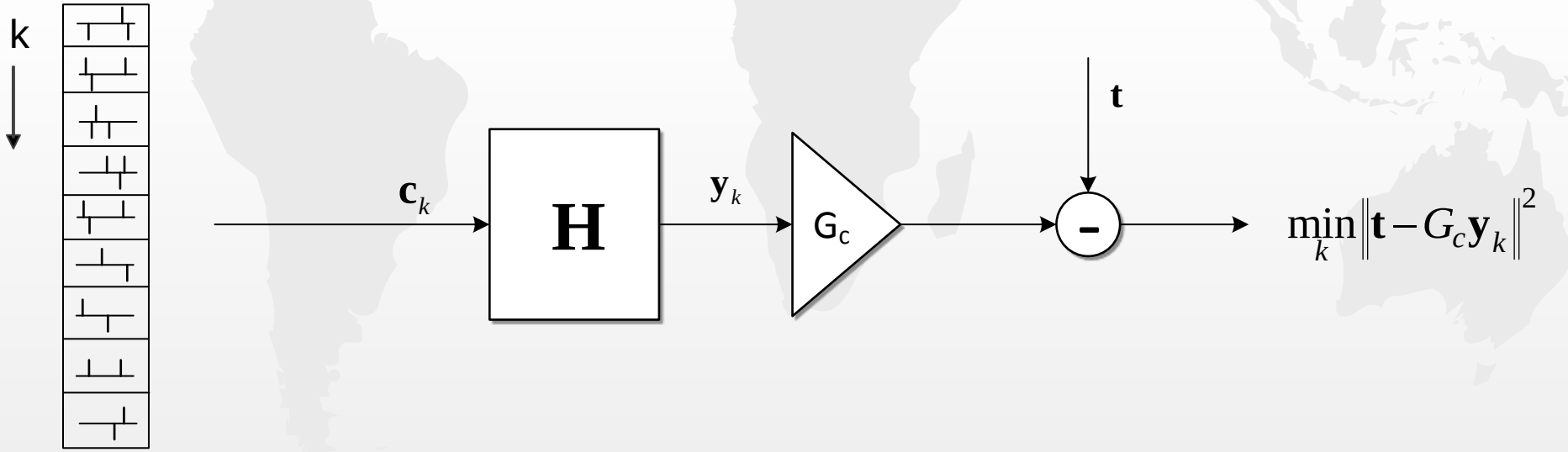
energie testovaného vektoru



algebraická knihovna (až 80 bitů)

Prohledávání algebraické knihovny

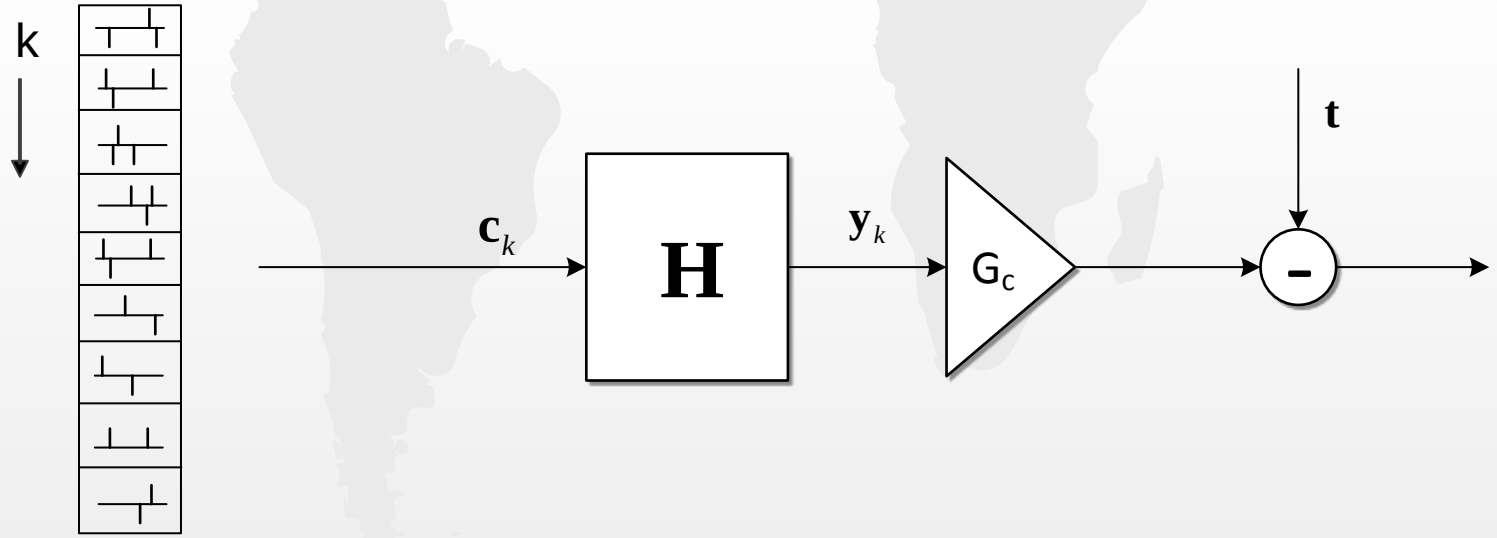
$$\min_k \left\| \mathbf{t} - G_c \mathbf{y}_k \right\|^2 \xrightarrow{\frac{\partial}{\partial G_c}} \max_k \frac{\left(\mathbf{t}^T . \mathbf{y}_k \right)^2}{\mathbf{y}_k^T . \mathbf{y}_k} \rightarrow \max_k \frac{\left(\mathbf{t}^T . \mathbf{H} . \mathbf{c}_k \right)^2}{\mathbf{c}_k^T . \mathbf{H}^T . \mathbf{H} . \mathbf{c}_k}$$



algebraická knihovna (až 80 bitů)

Prohledávání algebraické knihovny

$$\min_k \left\| \mathbf{t} - G_c \mathbf{y}_k \right\|^2 \xrightarrow{\frac{\partial}{\partial G_c}} \max_k \frac{\left(\mathbf{t}^T . \mathbf{y}_k \right)^2}{\mathbf{y}_k^T . \mathbf{y}_k} \rightarrow \max_k \frac{\left(\mathbf{t}^T . \mathbf{H} . \mathbf{c}_k \right)^2}{\mathbf{c}_k^T . \mathbf{H}^T . \mathbf{H} . \mathbf{c}_k} \rightarrow \max_k \frac{\left(\mathbf{d}^T . \mathbf{c}_k \right)^2}{\mathbf{c}_k^T . \mathbf{\Phi} . \mathbf{c}_k}$$



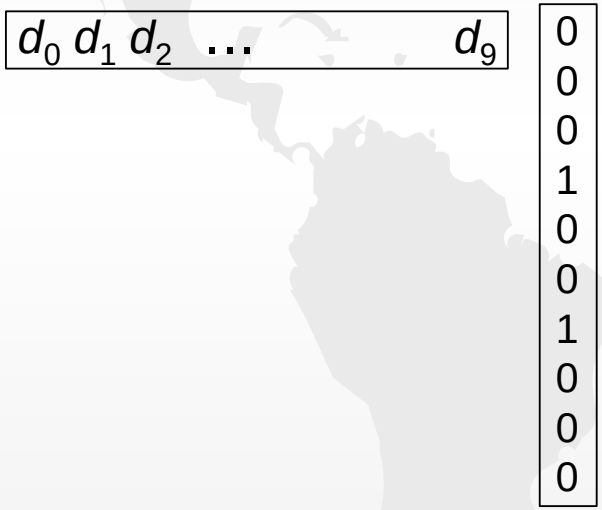
algebraická knihovna (až 80 bitů)

Prohledávání algebraické knihovny

$$\max_k \frac{\left(\mathbf{d}^T.\mathbf{c}_k\right)^2}{\mathbf{c}_k^T.\mathbf{\Phi}.\mathbf{c}_k}$$

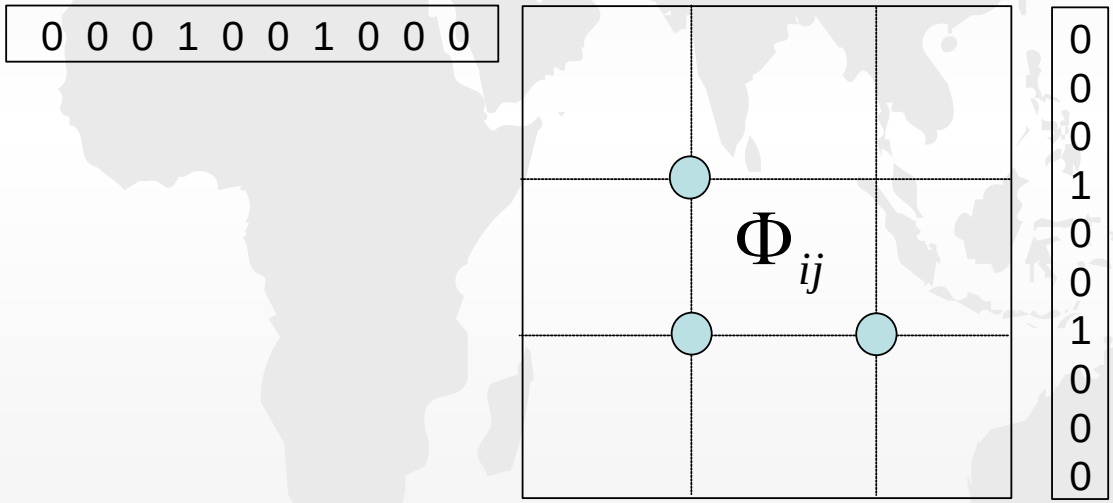
Lze prohledávat rychle, pokud $\mathbf{c}_k$ obsahuje jen velmi málo nenulových prvků s hodnotami +1 nebo -1

$\mathbf{d}^T.\mathbf{c}_k$



$= d_3 + d_6$

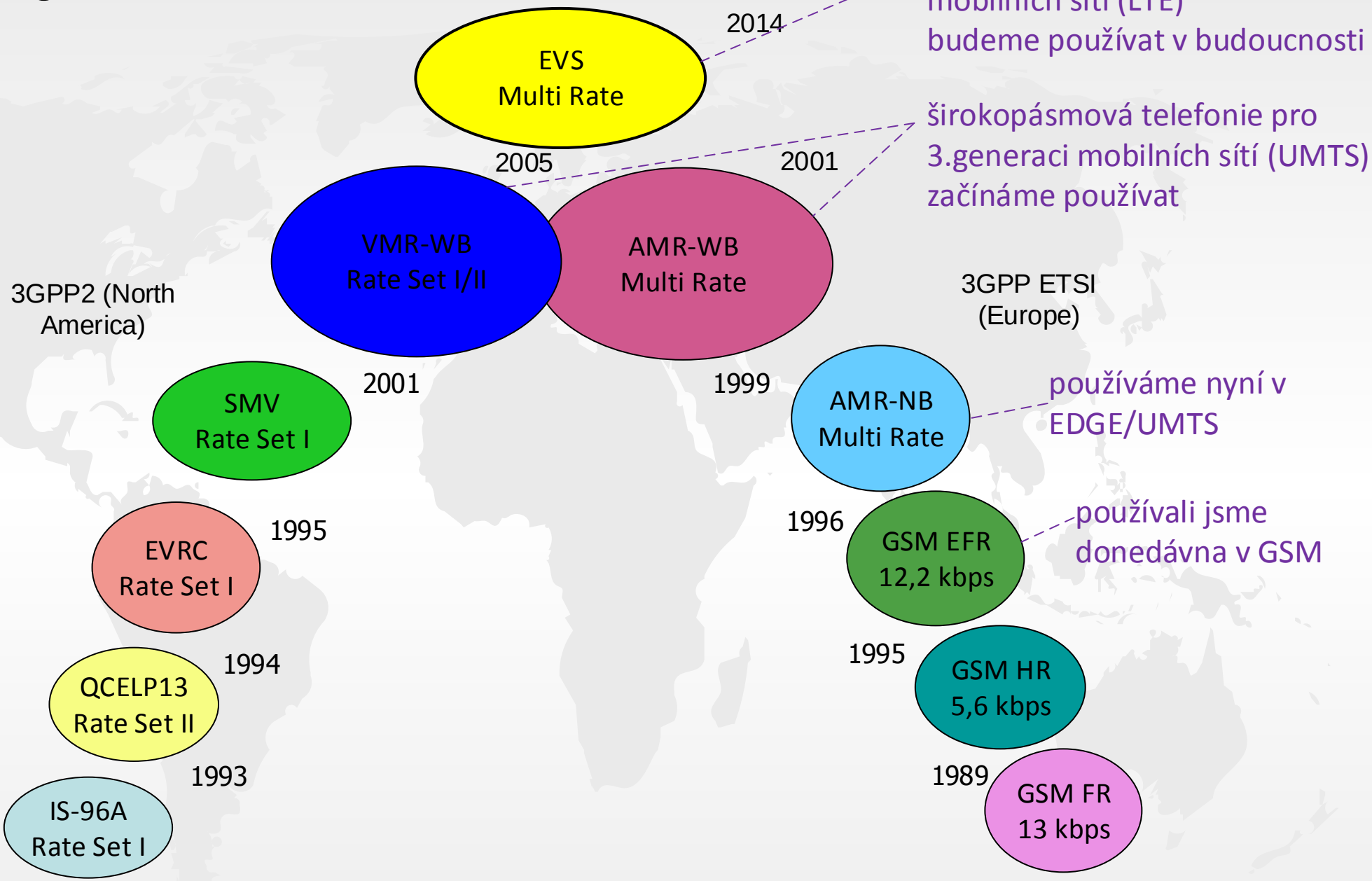
$\mathbf{c}_k^T.\mathbf{\Phi}.\mathbf{c}_k$



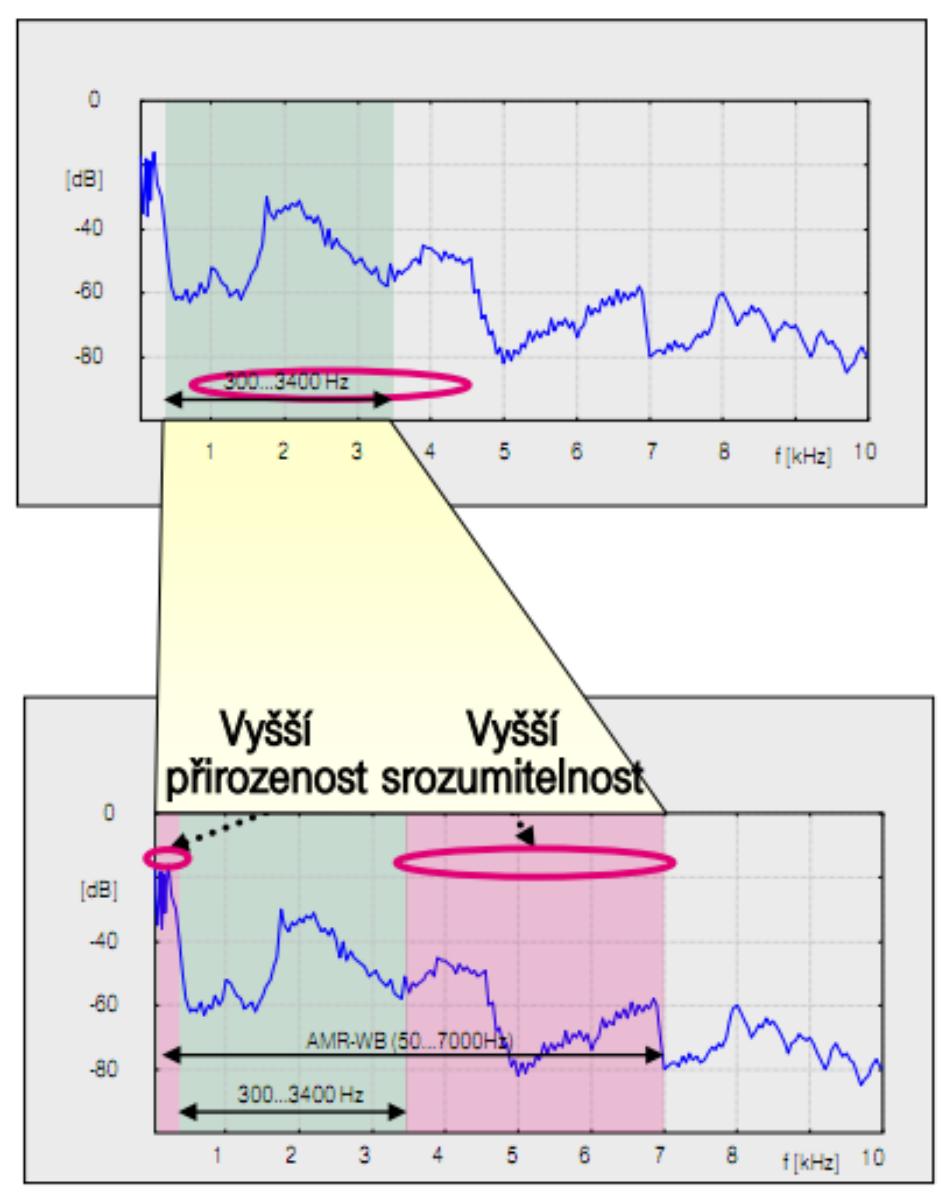
$= \Phi_{3,3} + \Phi_{6,6} + 2\Phi_{3,6}$

ACELP ve světě

Technologie ACELP v mezinárodních standardech



Od AMR-NB k AMR-WB (HD VOICE)



- HD voice demo na https://www.youtube.com/watch?v=Y4bb3b9PiRg


Konec

Základní schéma kodeku