Kódování řeči I.

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

FIT VUT Brno

Plán

- Dělení kodérů
- Vyhodnocování kvality
- Kódování “tvaru vlny” (waveform coding)
- Vokodéry
- Vektorové kvantování

Proč ?

Naprostá většina komunikací probíhá v dnešní době *číslicově*.

- co nejmenší počet bitů.
- co největší kvalitu.
- co nejmenší zpoždění.
- co největší odolnost proti chybám.
- co nejmenší výpočetní náročnost.

... jednotlivá kriteria jsou samozřejmě v rozporu.

Kódování — komerčně nejdůležitější součástí ASŘ (automatického zpracování řeči)

Standardizace

- CCITT (Centre Consultatif International Téléphonique et Télégraphique), ze kterého vznikla ITU-TSS (International Telecommunication Union — Telecom. Standardization Sector). Doporučení Gxxx. `http://www.itu.int`
- US DoD (Department of Defense). Federální standardy FSxxxx.
- ETSI (European Telecommunications Standards Institute), aktivní především v mobilní telefonii. `http://www.etsi.org`
- a další instituce, např. INMARSAT.

Dělení kodérů – Principiální dělení

- **“kódování tvaru vlny”** (waveform coders) - vzorek po vzorku. Vysoká kvalita, ale za cenu velkého bitového toku. I pro neřečové signály.
- **vokodéry** (vocoders) založeny na poznatcích o tvorbě a slyšení řeči člověkem (buzení + modifikace). Rámce. LP model. Složitější než waveform coders. Střední a nízké rychlosti. Jen řeč.
- (Hybridní) - někdy jsou takto nazývány algoritmy CELP (GSM). Model pro modifikační ústrojí, ale částečné kódování waveform pro buzení.
- **fonetické vokodéry** (phonetic vocoders) – delší řečovými úseky, než rámce (fonémy, automaticky natrénované jednotky). Založené na rozpoznávání v kodéru a syntéze v dekodéru. Dosud pouze laboratorní stadium, zatím žádná normalizace. Jen řeč, závislé na jazyku nebo i na mluvčím.

Dělení podle bitového toku

Bitový tok (bit rate) = počet bitů za sekundu pro kódování řeči (source coding). Při dalším kódování — channel coding.

- fixed-rate (klasické tlf. sítě)
- variable-rate (paketové).

označení	bitový tok
high rate	$>$ 16 kbit/s
medium rate	8 – 16 kbit/s
low rate	2.4 – 8 kbit/s
very low rate	$<$ 2.4 kbit/s

Dělení podle kvality

- **broadcast** (rozhlasová) – širší pásmo než telefonní, tok $>$ 64 kbit/s.
- **network** nebo **toll** (síťová) – klasická kvalita analogového telefonního signálu, pásmo 300–3400 Hz.
- **communications** (komunikační) – poněkud horší, avšak srozumitelná, zachovává charakter mluvčího.
- **synthetic** (syntetická) – nepřirozená, nezachovává charakter mluvčího, snížená srozumitelnost.

Vyhodnocování kvality

- objektivní
- subjektivní

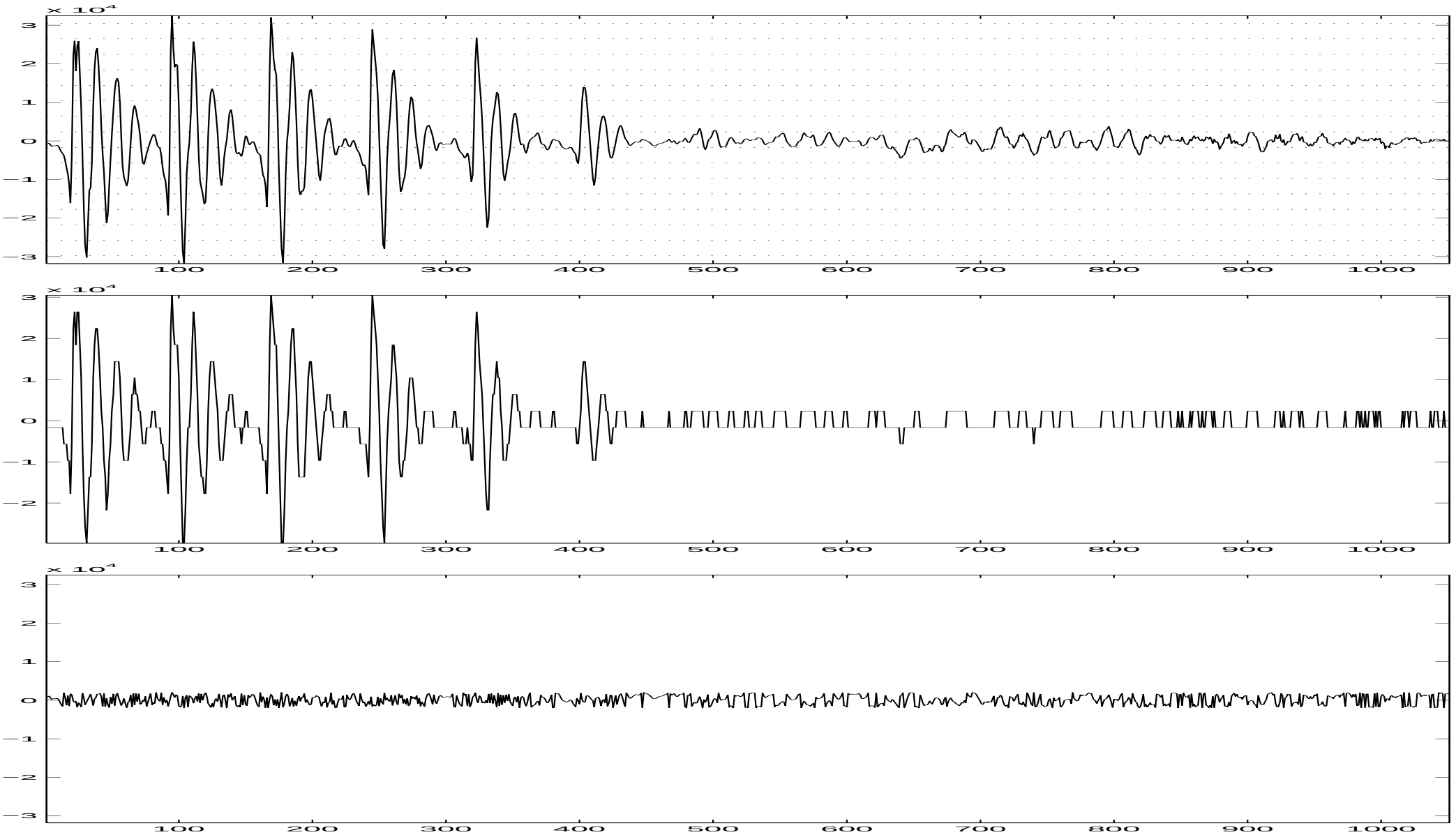
“Objektivní— vyhodnocování

odstup signálu od šumu (nebo poměru signálu k šumu, signal-to-noise ratio SNR):

$$SNR = 10 \log_{10} \left\{ \frac{\sum_{n=0}^{N-1} s^2(n)}{\sum_{n=0}^{N-1} [s(n) - \hat{s}(n)]^2} \right\}$$

Nevýhoda: je, že pohlíží na signály “globálně”

Příklad: kódování slabiky “as” pomocí 4 bitů (16 kvantizačních úrovní) SNR = 14.89 dB.


x 10^4
3
2
1
0
−1
−2
−3
100
200
300
400
500
600
700
800
900
1000
x 10^4
3
2
1
0
−1
−2
100
200
300
400
500
600
700
800
900
1000
x 10^4
3
2
1
0
−1
−2
−3
100
200
300
400
500
600
700
800
900
1000

segmentální poměr signálu k šumu (SEGSNR)

$$SEGSNR = \frac{10}{N_{ram}} \sum_{i=0}^{N_{ram}-1} \log_{10} \left\{ \frac{\sum_{n=0}^{l_{ram}-1} s^2(il_{ram}+n)}{\sum_{n=0}^{l_{ram}-1} [s(il_{ram}+n) - \hat{s}(il_{ram}+n)]^2} \right\} \qquad (1)$$

V tomto případě vycházejí SNR v jednotlivých rámcích (při "klasické" délce rámce 160 vzorků):

20.04 19.63 14.35 0.21 4.26 -0.54(!)

a SEGSNR (jejich průměrná hodnota) je **9.66 dB**, což je podstatně horší, avšak o realitě více vypovídající nežli SNR.

Logaritmická spektrální vzdálenost – logarithmic spectral distance

počítá zkreslení mezi originálním a kódovaným signálem ve *spektrální oblasti*. (. . . dá se velmi snadno počítat pomocí LPC-cepstrálních koeficientů – ne integrál, jen suma)

$$d_2 = \sqrt{\int_{-1/2}^{+1/2} |V(f)|^2 \, df}, \quad \text{where} \quad V(f) = 10 \log G(f) - 10 \log \hat{G}(f), \tag{2}$$

kde $G(f)$ a $\log \hat{G}(f)$ jsou spektrální hustoty výkonu originálnío a kódovaného rámce.

Subjektivní měření kvality

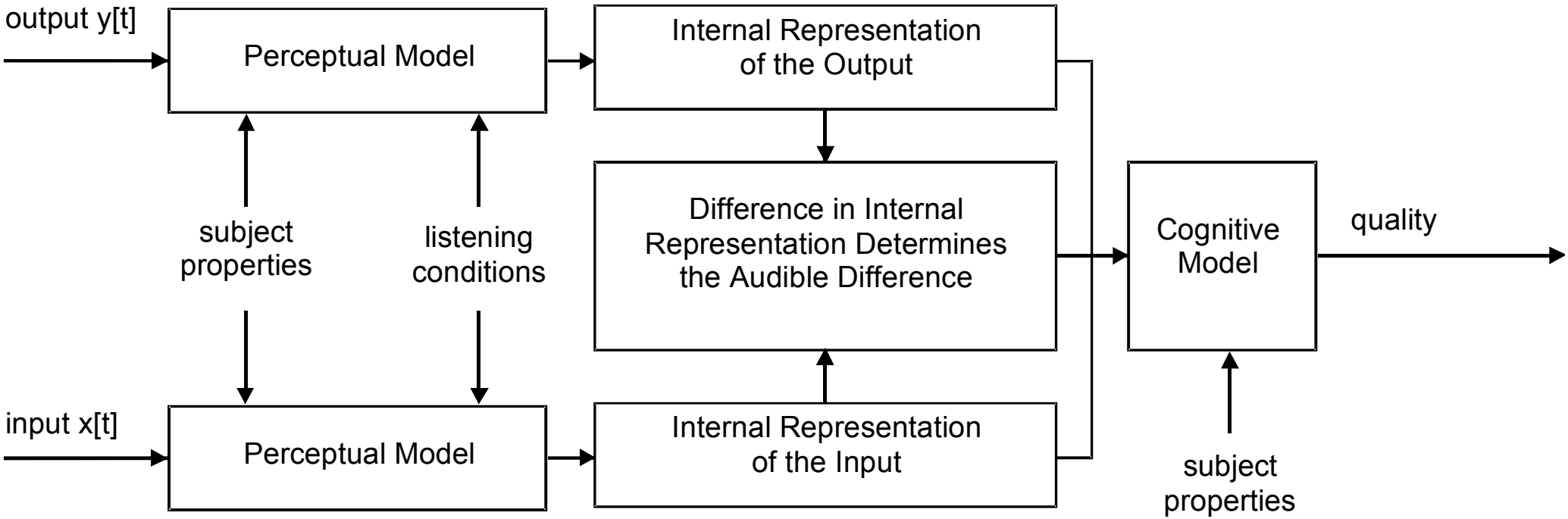
normalizované postupy vyžadují vždy skupinu posluchačů, kteří kvalitu posuzují.

- DRT (Diagnostic Rhyme Test) – měření srozumitelnosti pomocí párů podobných slov (např. meat×heat).
- DAM (Diagnostic Acceptability Measure) – soubor několika metod hodnotících kvalitu komunikačního systému.
- MOS (Mean Opinion Score) – skupina 12–64 posluchačů hodnotí kvalitu podle pětibodové stupnice. Posluchači jsou nejprve "kalibrování" signály se známými hodnotami MOS:

MOS	kvalita	poznámka
1	bad (unacceptable)	velmi rušivý šum a artefakty v signálu
2	poor	...
3	fair	něco mezi
4	good	...
5	excellent	nerozeznatelné od originálu, bez slyšitelného šumu

Měření kvality inspirované lidským slyšením

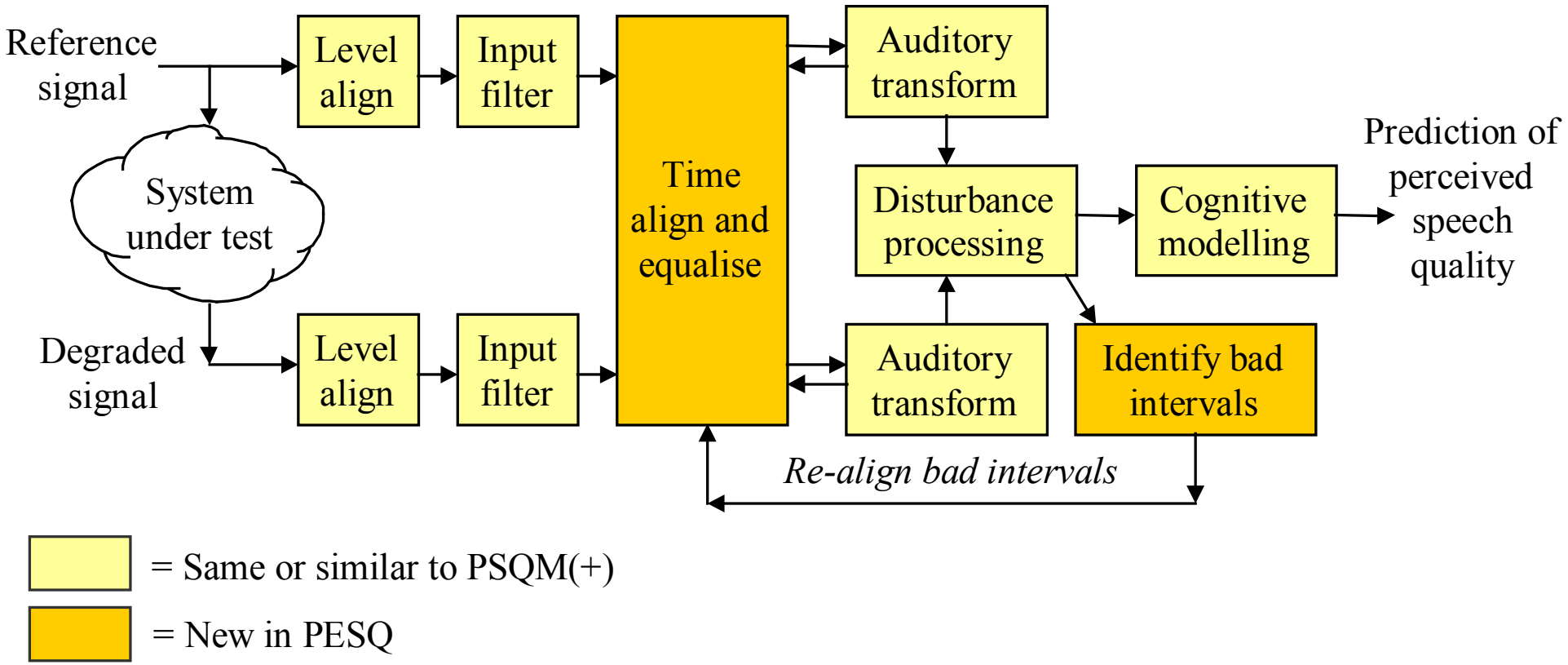
Perceptual Speech Quality Measure – PSQM



source: OPTICOM Whitepaper on "State-of-the-Art Voice Quality Testing", 2000

ITU standard 1996, P.861.

Perceptual Evaluation of Speech Quality – PESQ



source: OPTICOM Whitepaper on "State-of-the-Art Voice Quality Testing", 2000

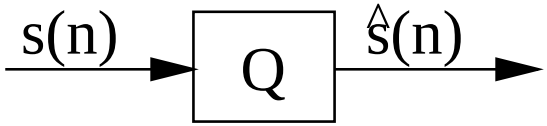
ITU standard P.862

A.W.Rix et al.: Perceptual evaluation of speech quality (PESQ) a new method for speech quality assessment of telephone networks and codecs, Proc. ICASSP 2001.

KÓDOVÁNÍ TVARU VLNY (WAVEFORM CODING)

Pulsní kódová modulace – Pulse code modulation (PCM)

historický název, nezávislé kvantování jednotlivých vzorků pomocí fixního počtu bitů



Lineární kvantování (Q krok konstantní):

$$SNR = 6B + K \tag{3}$$

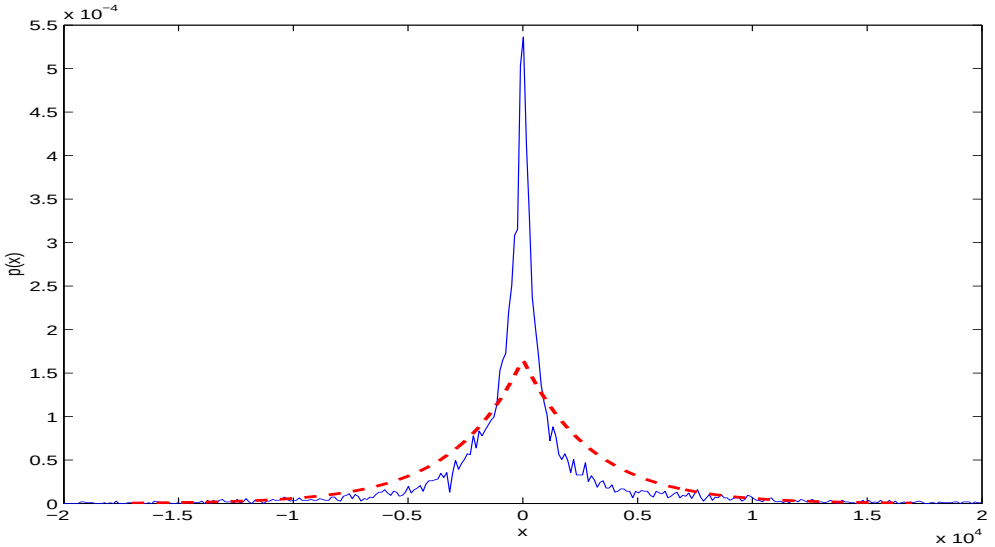
v [dB], kde B je počet bitů a K je konstanta závisející na charakteru signálu. Pro CD teoreticky: : 16×6=96 dB. Interpretace rovnice: přidáme-li 1 bit, SNR se zlepší o 6 dB.

Pro řečové signály ale není lin. kvantování to nejvhodnější:

1. řeč obsahuje mnoho "malých vzroků", její funkce hustoty rozdělení pravděpodobnosti (probability density function – PDF) m;že být aproximována Laplacovým rozložením:

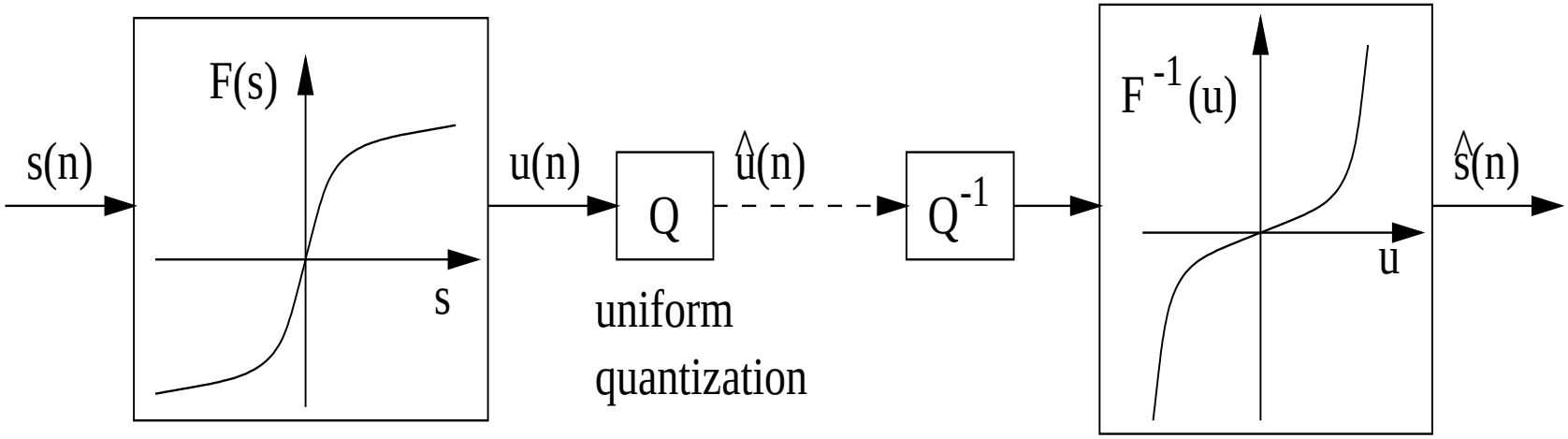
$$p(x) = \frac{1}{\sqrt{2}\sigma_x} e^{-\frac{\sqrt{2}|x|}{\sigma_x}}, \tag{4}$$

   kde σ_x je směrodatná odchylka. Příklad: česká věta, bez ticha:



2. z perceptuálních studií víme, že ucho má logaritmickou citlivost na apllitudu akustického tlaku.

logaritmická PCM, komprese v kodéru a expanse v dekodéru:



Nelinearita nemůže být přímo log: ($\log(0) = -\infty$), aproximace:

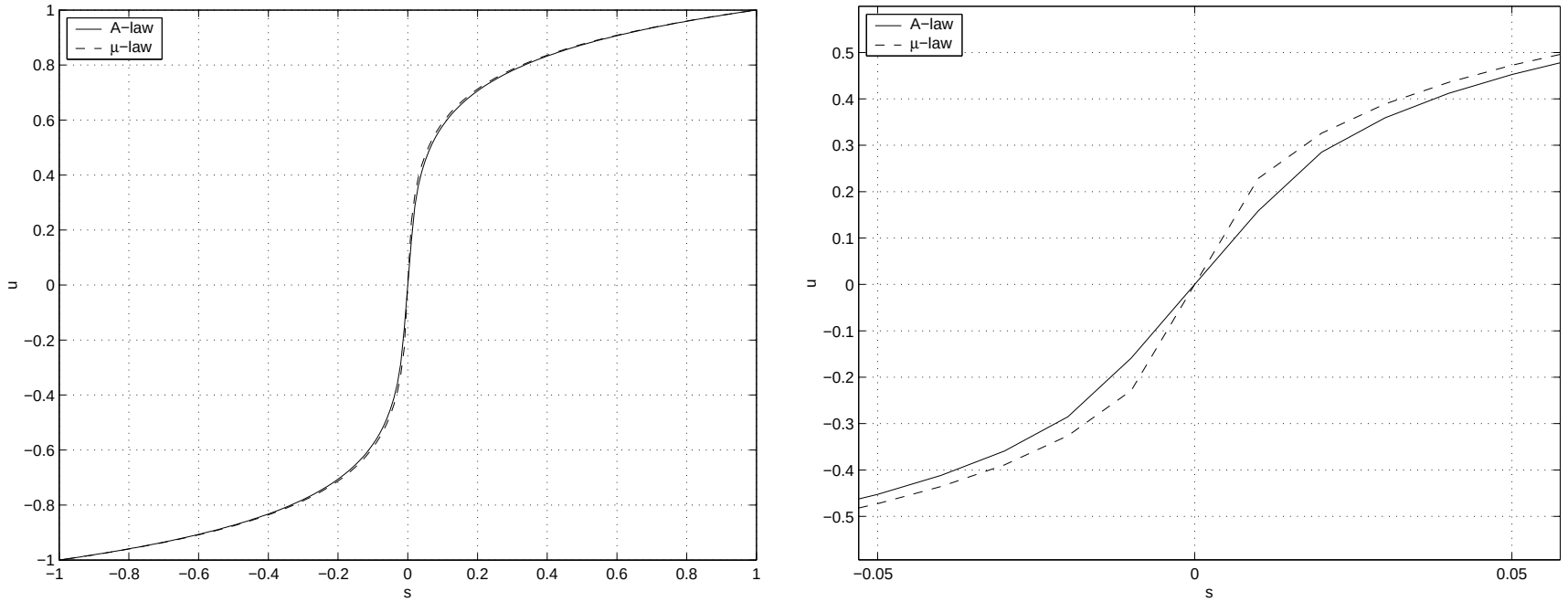
- Evropa: **A-law**:

$$u(n) = S_{max} \frac{1 + \ln A \frac{|s(n)|}{S_{max}}}{1 + lnA} \text{ sign } [s(n)], \quad \text{where} \quad A = 87.56. \tag{5}$$

- USA: μ-law:

$$u(n) = S_{max} \frac{\ln\left(1 + \mu \frac{|s(n)|}{S_{max}}\right)}{\ln(1 + \mu)} \text{ sign } [s(n)], \quad \text{where} \quad \mu = 255. \tag{6}$$

Srovnání A-law a μ-law:



$\Rightarrow$ obě jsou prakticky identická, obě zlepšují SNR pro malé signály o 12 dB. Pro telefonní aplikace, log-PCM s 8 bity má podobnou kvalitu jako 13 bitů lin. CCITT G.711.

Adaptivní pulsní kódová modulace (APCM)

Rozložení hladin se počítá z bloku několika vzorků. Informace o kvantovacích hladinách:

- se může vysílat do dekodéru jako přídavná informace, tzv. *feed-forward*.
- se může počítat také zpětně z několika minulých vzorků, které má k disposici i dekodér. V tomto případě není nutné informaci přenášet, tzv. *feed-back*.

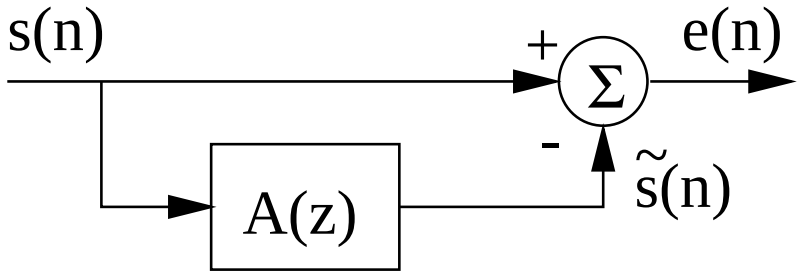
APCM se nepoužícvá samostatně, ale jako součást složitějších kodérů (např. full rate GSM: RPE-LTP).

Diferenční kódová modulace – Differential pulse-code modulation (DPCM)

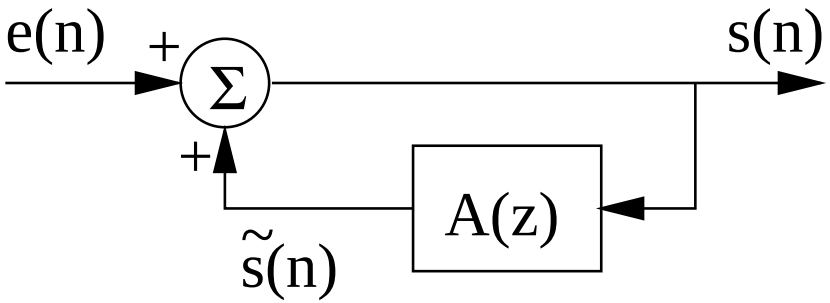
- využívá statistických závislostí mezi vzorky (jediný signál, kde jsou vzorky totálně nezávislé, je bílý šum, ten jen tak lehce nepotkáme).
- současný vzorek je odhadnut z několika předcházejících
- v případě dobrého odhadu bude mít rozdílový (chybový) signál malou energii a malou amplitudu $\Rightarrow$ méně bitů.

$$A(z) = \sum_{i=1}^{P} a_i z^{-i} \tag{7}$$

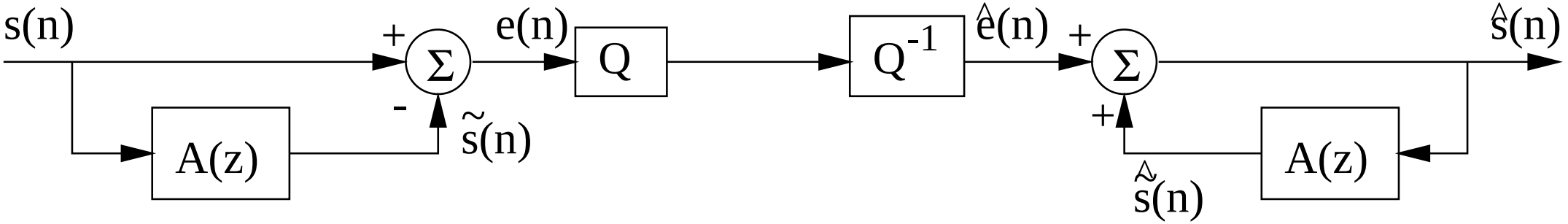
Chybový signál:



Dekodér: současný vzorek je předpovězen z předchozích a přidá se chyba:

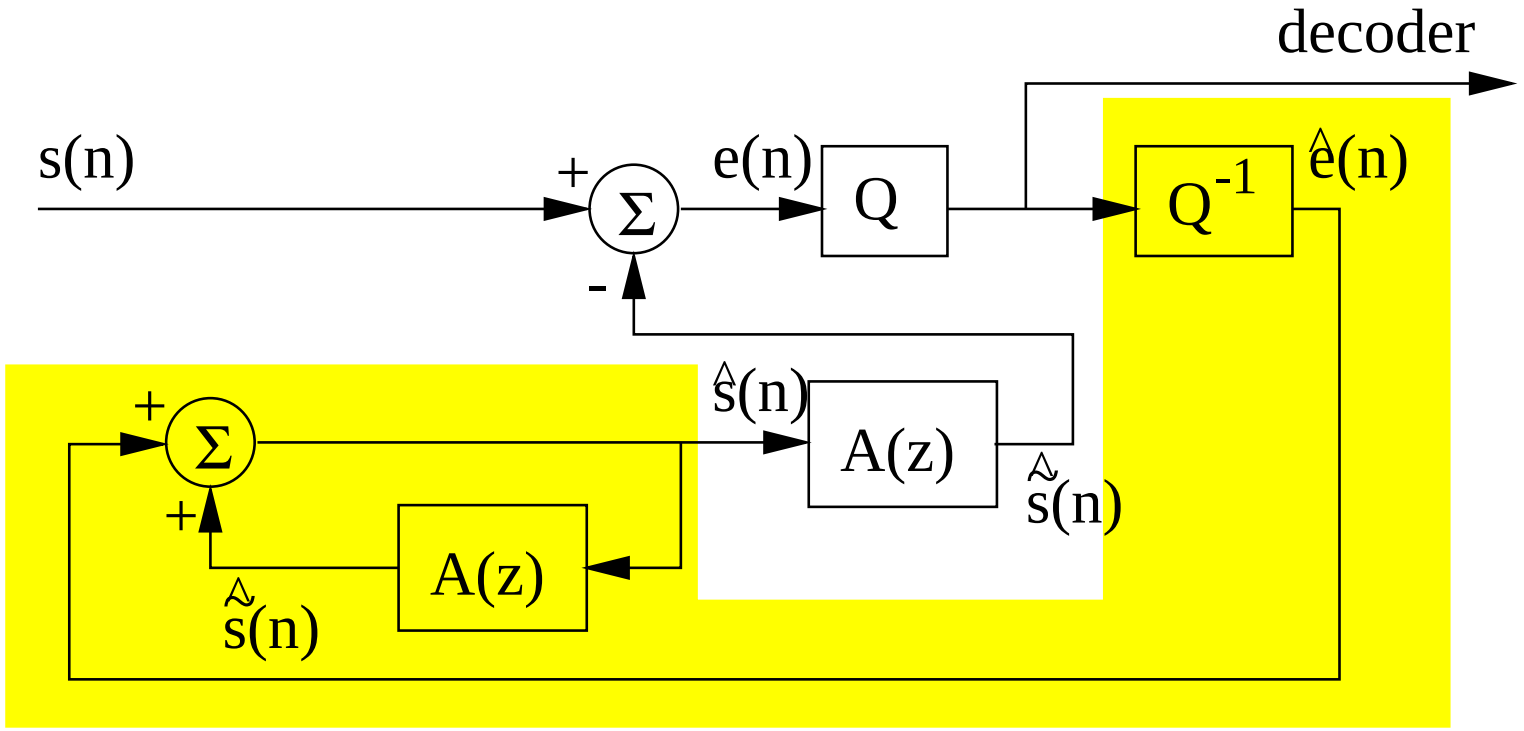


. . . jenže chybový signál je nutné **kvantovat**



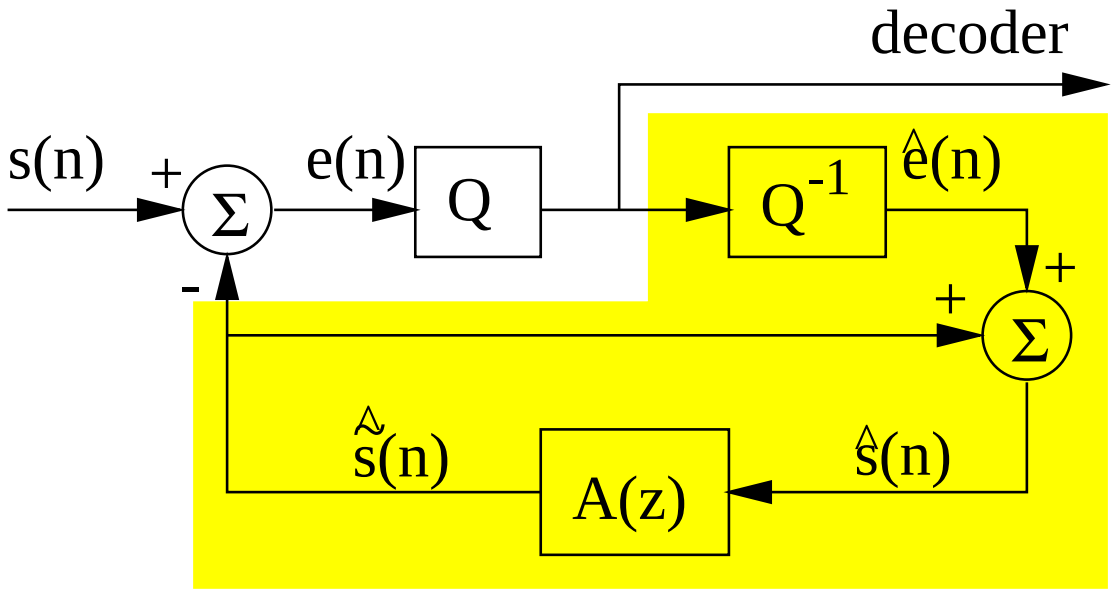
!!! V tomto případě by dekodér odhadoval současný vzorek z něčeho **jiného** než kodér !

- Kodér má: $s(n)$.
- Dekodér má $\hat{s}(n)$, který se díky kvantování $e(n)$, nerovná $s(n)$!
- Musíme dekodér "vestavět" do kodéru. Jeho výstup se využije k predikci.



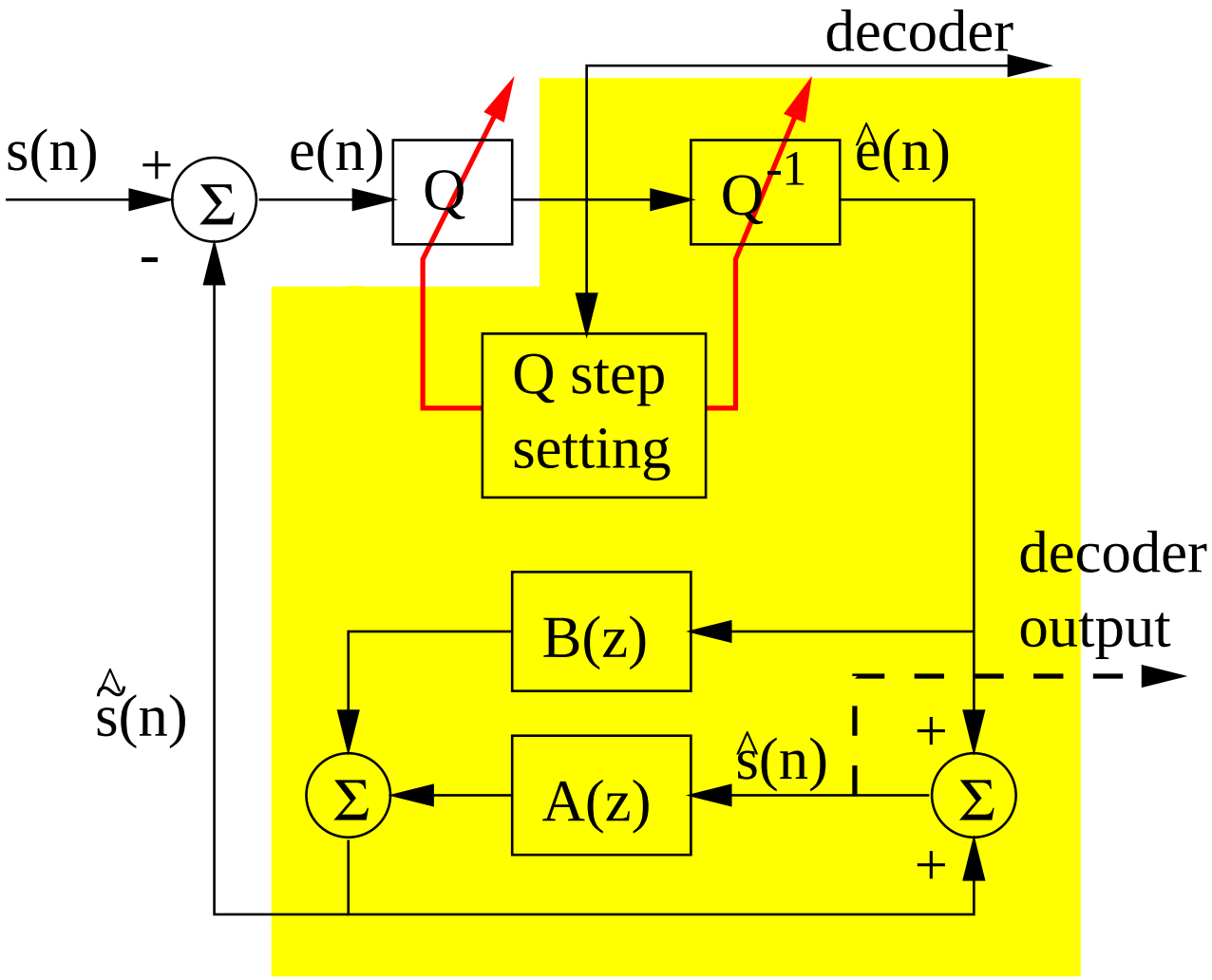
Je to dost komplikované a filtr $A(z)$ je tam dvakrát a filtruje ten samý signál !

Zjednodušení:



Kodér opět obsahuje celý dekodér (žlutá barva).

Adaptivní diferenční pulsní kódová modulace (ADPCM) G. 721



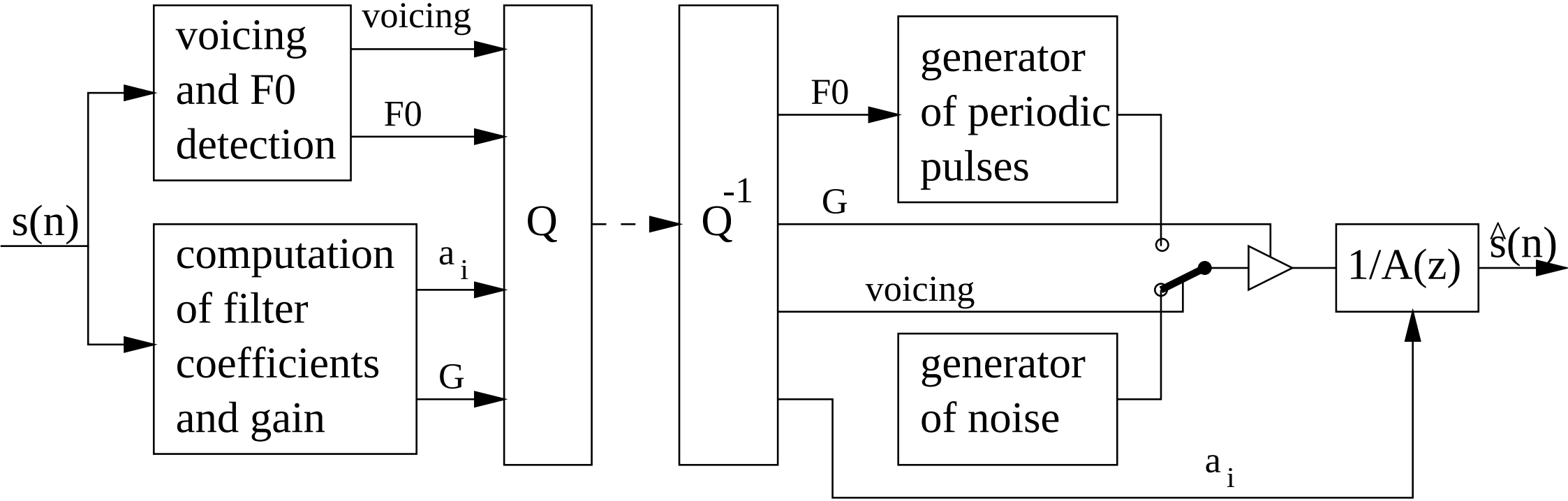
- rozdíl od DPCM - adaptace kvantovacího kroku.
- ADCPM je použito v G.721: log-PCM (64 kbit/s) $\Rightarrow$ 32 kbit/s.
- G.721 obsahuje 2 filtry pro výpočet chybového signálu $e(n)$: $A(z)$ (jako minule), $B(z)$ odhaduje vzorek chybového signálu z nekolika předešlých hodnot chybového signálu (IIR).
- Blok "Q-step setting" reguluje kvantovací krok.
- Dekodér (žlutý) je zase "schován" v kodéru.
- Informace o Q kroku a odad filtrů je založen na výstupu z kodéru (back-ward).

VOKODÉRY

- využívají poznatků o lidském řečovém ústrojí pro redukci bitového toku.
- uspokojivě zpracovávají pouze řeč, pokud jsou jim ke kódování předloženy jiné signály (např. hudba), výsledkem je většinou "cosi podobného řeči", samozřejmě zcela nesrozumitelného.
- využívají modelu buzení—filtr.

Kodér založený na lineárně prediktivním modelu - LPC

Vstupní signál rozdělen na rámce, z každého je odhadnuta sada koeficientů polynomu: $A(z) = 1 + \sum_{i=1}^{P} a_i z^{-i}$. Spočten gain tohoto filtru G a je detekována znělost a v případě znělého rámce perioda základního tónu (lag).



Buzení !!!

Příklad: US-DoD FS1015 standard: filtr 1800 bps, buzení 600 bps, 2.4 kbps. Hlavní nevýhodou bylo velmi zjednodušené modelování **buzení** $\Rightarrow$ nepřirozená řeč. Velké zlepšení v CELP kodérech.

Residual Excited Linear Prediction – RELP

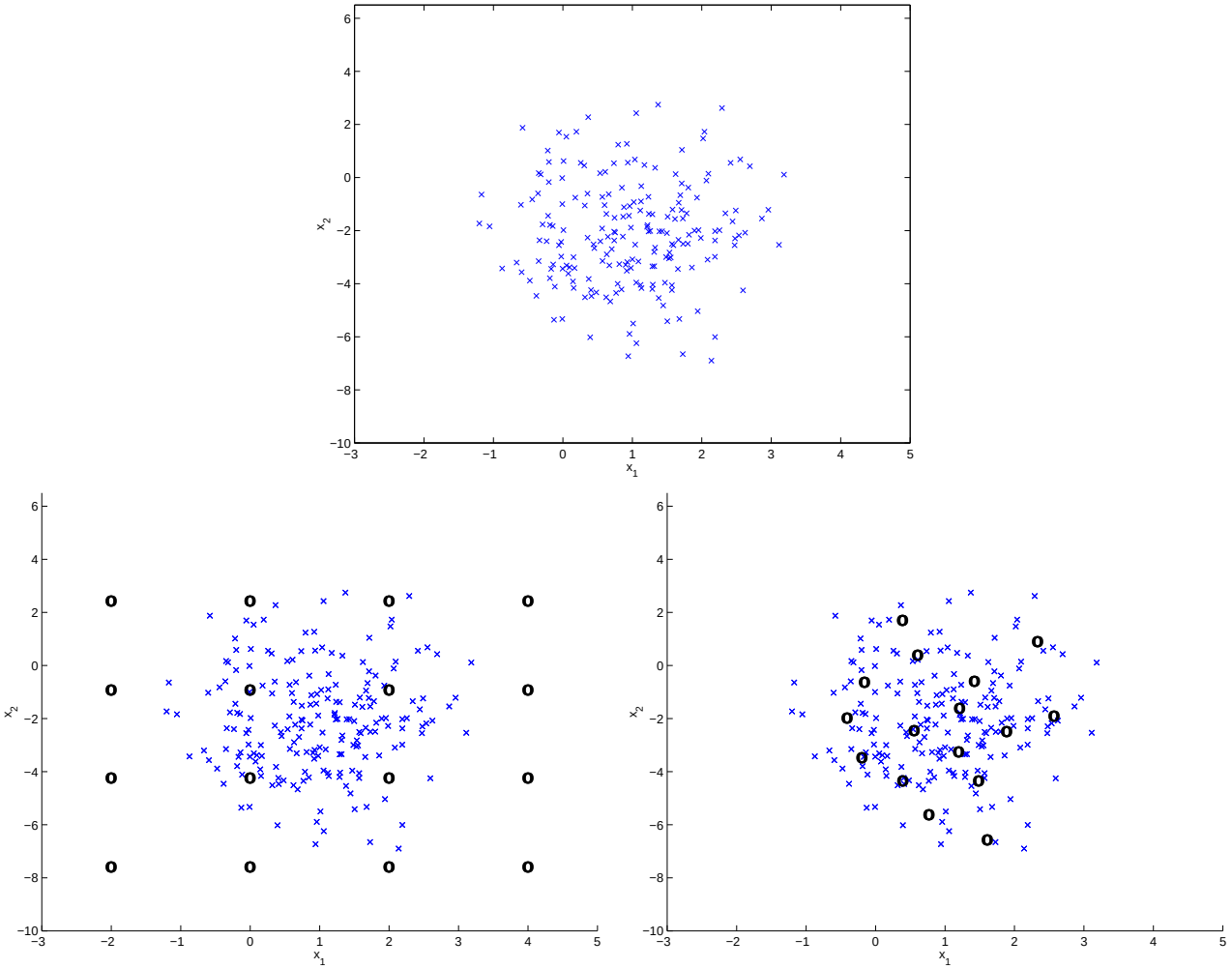
v každém rámci jsou odvozeny parametry filtru $A(z)$ a je vypočten chybový signál $e(n)$: $E(z) = A(z)S(z)$. Tento signál je přenesen do dekodéru, kde je filtrován filtrem $H(z) = \frac{1}{A(z)}$. Pokud nejsou koeficienty filtru a_i ani chybový signál $e(n)$ kvantovány, výsledkem je naprosto přesně vstupní signál $s(n)$.

Zvýšení bitového toku — Bad...

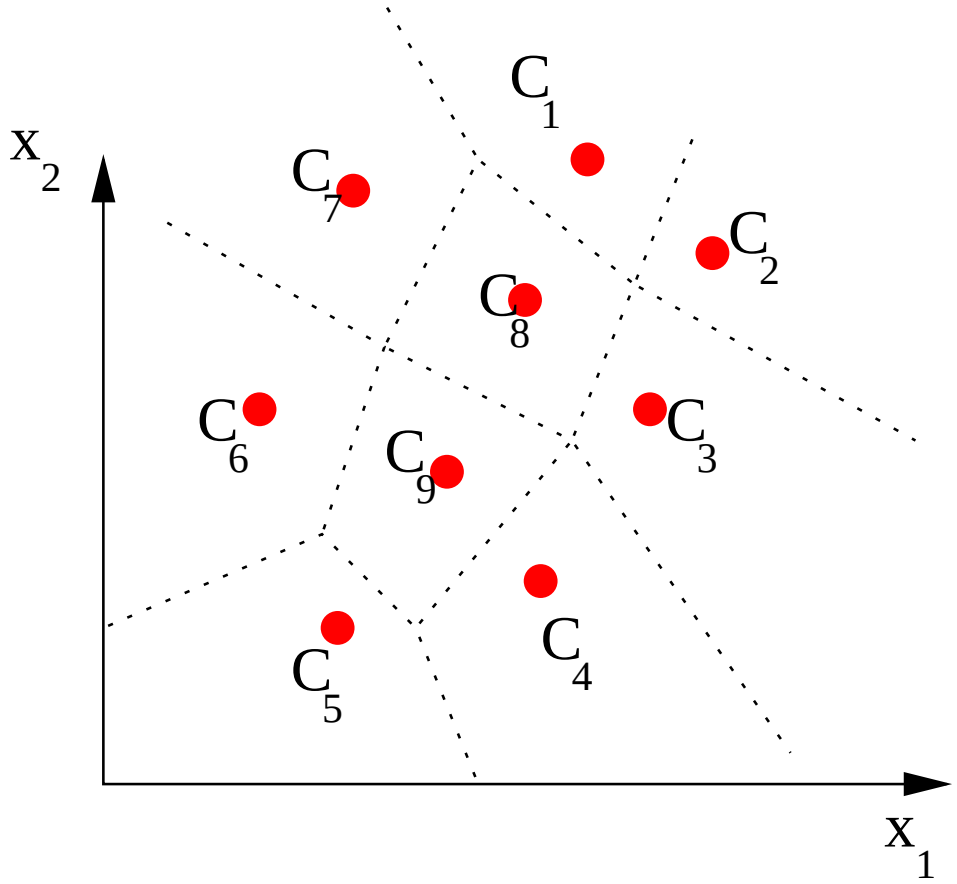
VEKTOROVÉ KVANTOVÁNÍ

Proč ?

- Kódování P - dimenzionálních vektorů je velmi nákladné (P floatů je $P \times 4$ byte...)
- Skalární kvantování jednotlivých složek je plýtvání bity.
- Lepší je umístit do P-rozměrného prostoru "typické" vektory a na ně "zaokrouhlovat".


x1
x2

Kódové vektory, centroidy, Voronoiovy regiony...



Trénování kódových vektorů — K-means

kódové vektory je nutné natrénovat na datech.

- máme k disposici N trénovacích vektorů.
- chceme kódovou knihu (codebook) $\mathbf{Y}$ o velikosti K.

- **Inicialisace:** $k = 0$, definujeme $\mathbf{Y}(0)$.
- **Step 1:** přiřazení vektorů jednotlivým buňkám "kódování":

$$Q[\mathbf{x}] = \mathbf{y}_i(k) \quad \text{if} \quad d(\mathbf{x}, \mathbf{y}_i(k)) \leq d(\mathbf{x}, \mathbf{y}_j(k)) \quad \text{pro} \quad j \neq i, \; j \in 1 \ldots K \tag{8}$$

Jako $d(\mathbf{x}, \mathbf{y}_j)$ je možné použít Euklidovu vzdálenost:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T (\mathbf{x} - \mathbf{y})} = \sqrt{\sum_{k=1}^{P} |x_k - y_k|^2}. \tag{9}$$

Po zakódování "patří" vektor $\mathbf{x}$ do buňky $C_i(k)$ (k je index "generace" kódové knihy).

- **Step 2:** Vyhodnocení kvality kódové knihy – totální vzdálenost (zkreslení):

$$D_{VQ} = \frac{1}{N} \sum_{n=1}^{N} d\left(\mathbf{x}(n), Q[\mathbf{x}(n)]\right). \tag{10}$$

- **Step 3:** pokud je relativní zlepšení zkreslení pod předem definovaným prahem:

$$\frac{D_{VQ}(k-1) - D_{VQ}(k)}{D_{VQ}(k)} \leq \varepsilon, \tag{11}$$

  STOP a prohlásíme k-tou generaci kódové knihy za výsledek. $\mathbf{Y} = \mathbf{Y}(k)$. Pokud stále zlepšení, pokračujeme:

- **Step 4:** Spočítáme nové centroidy buněk, stanou se z nich nové kódové vektory:

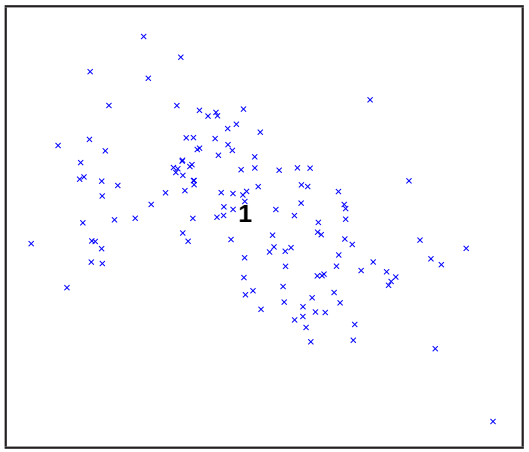
$$\mathbf{y}_i(k+1) = Cent(C_i(k)) = \frac{1}{M_i(k)} \sum_{\mathbf{x} \in C_i(k)} \mathbf{x}, \tag{12}$$

  kde $M_i(k)$ je počet trénovacích vektorů, které pro k-tou generaci kódové knihy spadly do buňky i. Jedná se o běžný aritmetický průměr. Inkrement: $k = k + 1$, GOTO Step 1.

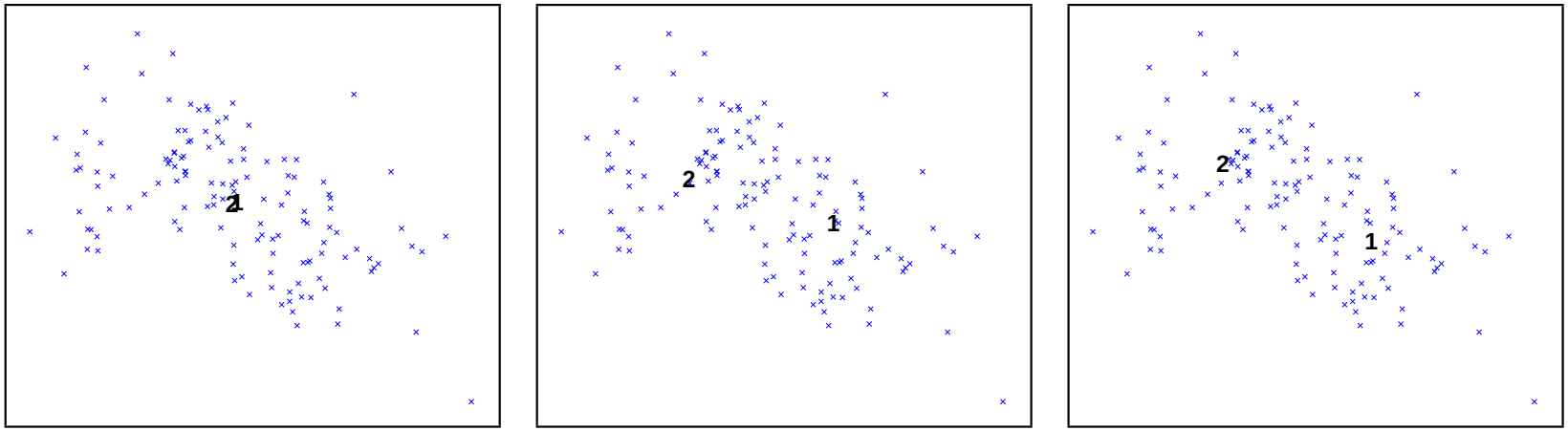
Linde-Buzo-Gray — LBG

- Algoritmus K-means má problémy s inicializací (jak inicializovat $\mathbf{Y}(0)$ – náhodně ? náhodným výběrem vektorů ?)
- může se stát, že při trénování na některý kód. vektor nezbydou žádné trénovací vektory $\Rightarrow$ dělení nulu $\Rightarrow$ crash :-(
- LBG řeší postupným štípáním vektorů v kódové knize:

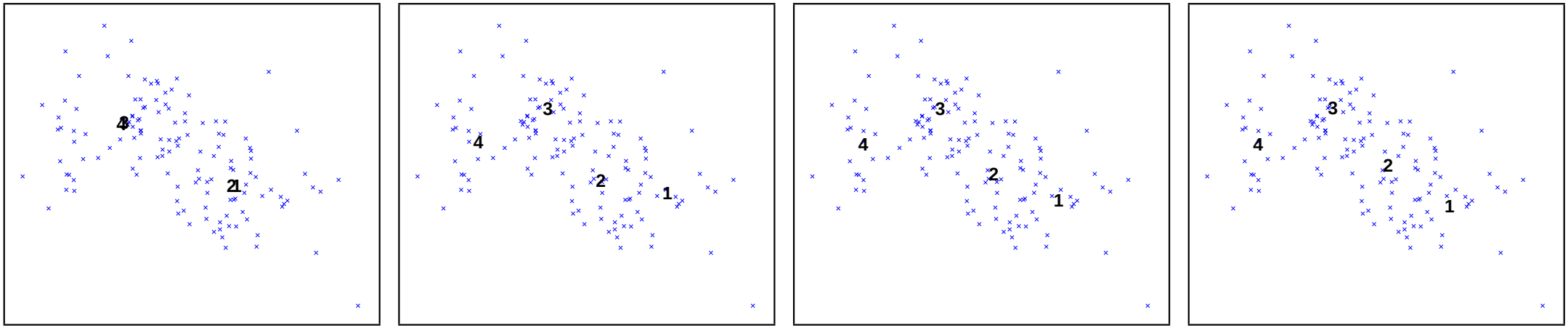
$r = 0, L = 1$



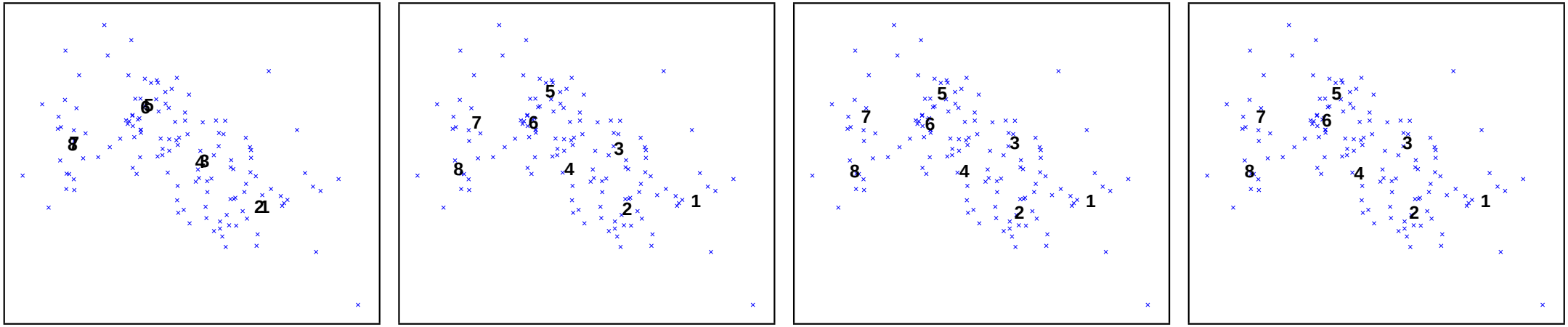
$r = 1, L = 2$



$r = 2, L = 4$



$r = 3, L = 8$



Kde se VQ používá

1. kódování koeficientů filtru a_i (nejčastějji jsou převedeny na něco vhodnějšího: PARCOR, LAR or LSF).
2. kódování buzení v CELP-like algoritmech, blocky vzorků po krátkodobém prediktoru a $A(z)$ a dlouhodobém prediktoru $B(z)$.

Varianty VQ

Trénování kódové, ale v případě velkých K i kódování jsou velmi výpočetně náročné
$\Rightarrow$ varianty.

- split-VQ: vektor je rozdělen na několik sub-vektorů s méně koeficienty. Použití několika menších codebooků. (typicky 3-3-4 pro P=10).
- algebraická VQ: kódové vektory nejsou rozmístěny libovolně, ale jejich pozice jsou deterministické (např. uniformě na hyper-ploše), není nutné porovnávat vstup se všemi kódovými vektory.
- náhodný codebook (pro trénování): pro velká K se kvalita natrénovaného codebooku blíží náhodnému $\Rightarrow$ trénování není potřeba!
- tree-structured VQ: zapamatujeme si generace při trénování LBG a předpokládáme, že když byl vektor $\mathbf{y}^k$ v generaci k kódové knihy přiřazen nějakému kódovému vektoru, pak pro generaci $k+1$ může náležet jen jeho dětičkám $\Rightarrow$ suboptimální, ale potřebuje jen $2\log_2 K$ srovnání na rozdíl od K.
- multi-stage VQ: jsou použity 2 codebooky, ve druhém se kvantuje chyba prvního. Při dekódování jsou kódové vektory z obou sečteny.