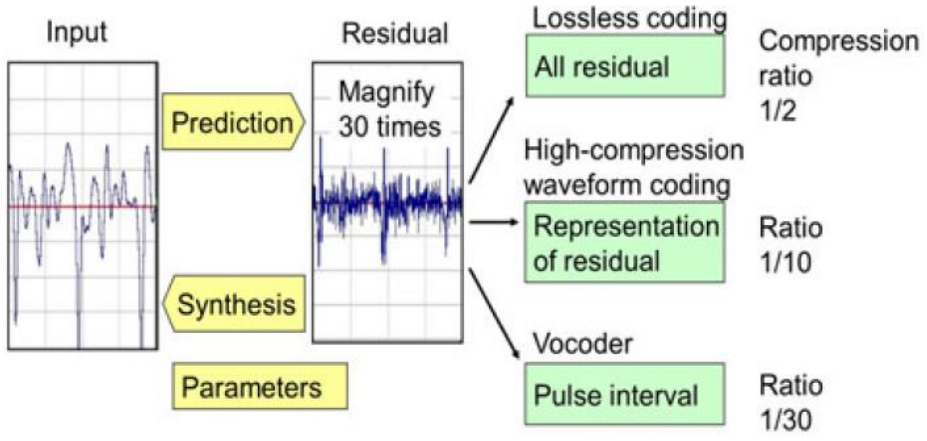
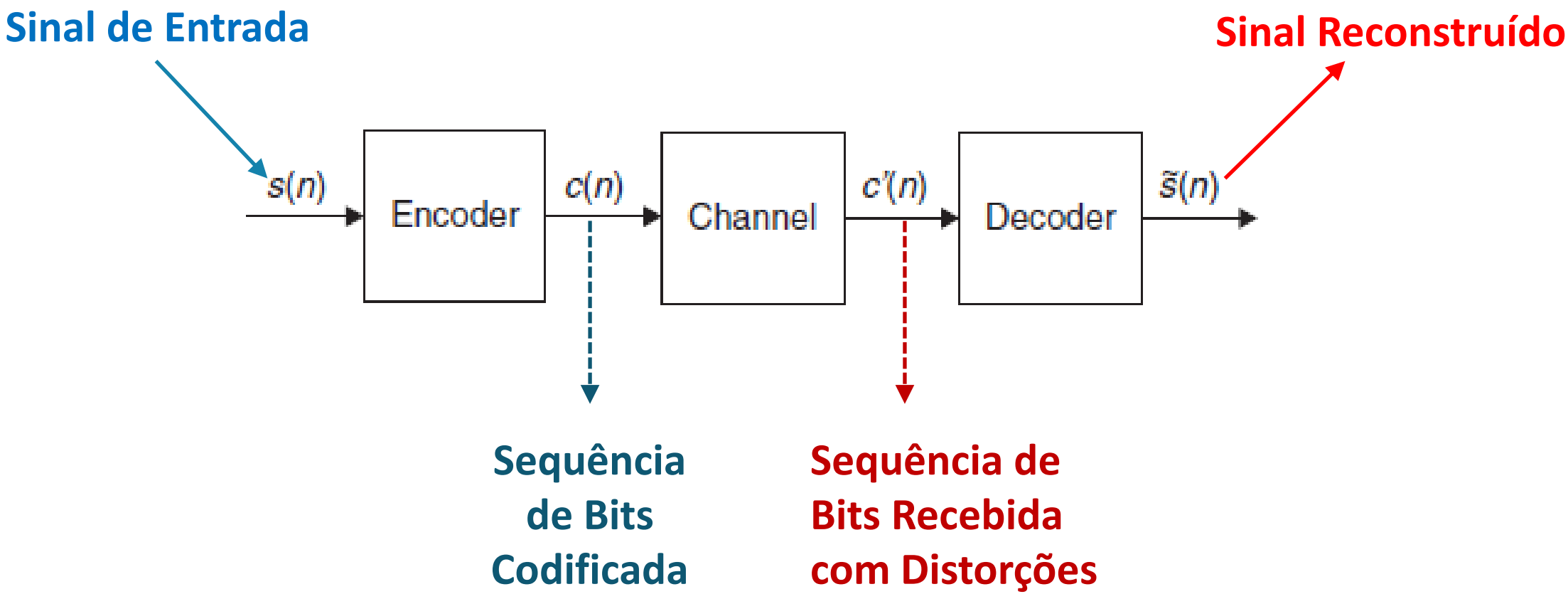
Codificação de Voz



ESTI019 – CODIFICAÇÃO DE SINAIS MULTIMÍDIA

PROFS. CELSO SETSUO KURASHIMA, KENJI NOSE E MÁRIO MINAMI

Modelo de Codificador/Decodificador



Atributos presentes numa comparação de Codificadores

Complexidade (alta, mais que dezena de MIPS ou MFLOPS)

Atraso (delay):

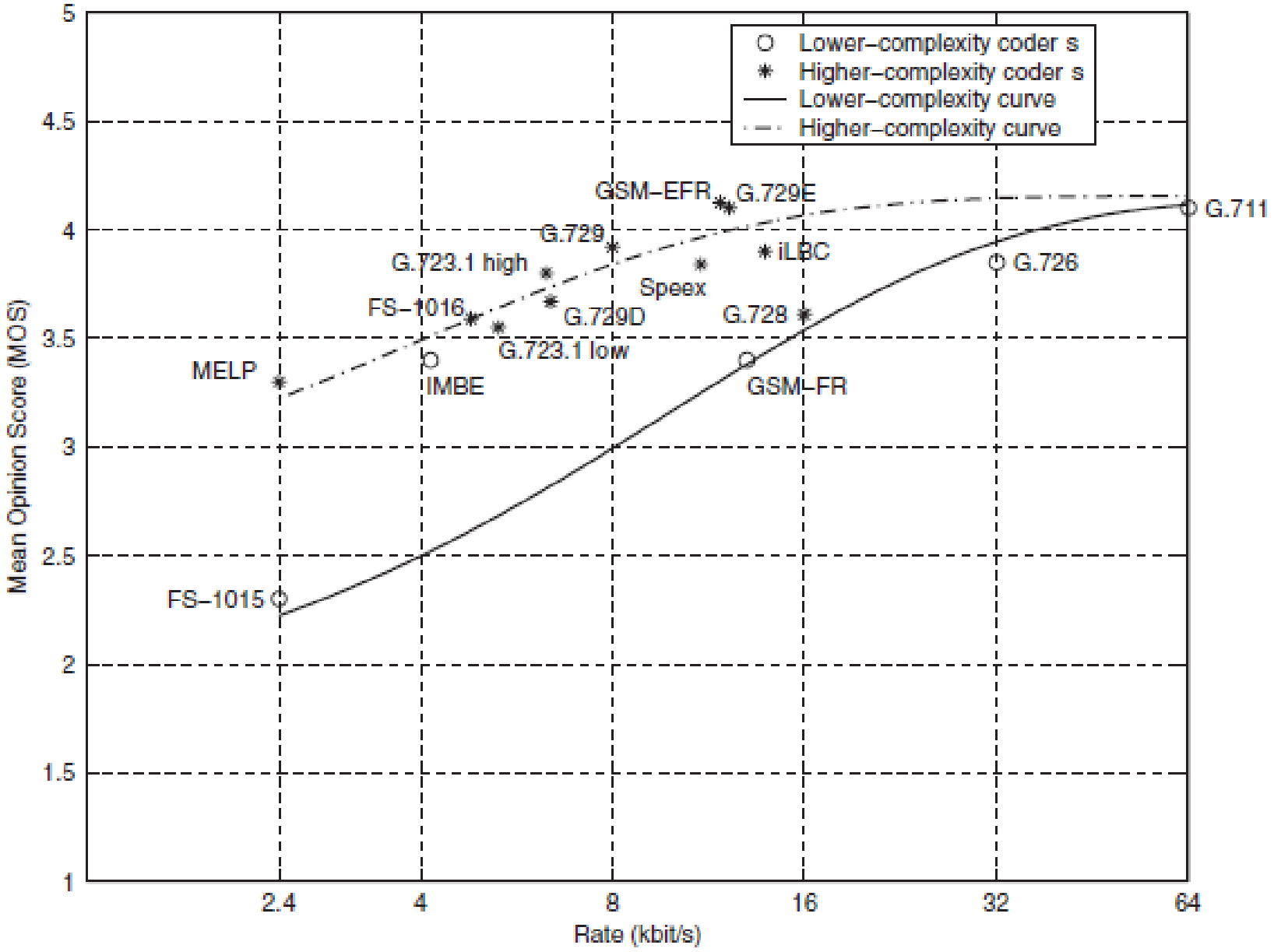
- Total máximo do sistema para full-duplex: 200ms
- Com eco: máximo de 25ms

Recuperação com Pacotes Perdidos

Escores Subjetivos de Qualidade de Voz

Descrição Baseada na Qualidade *Absolute Category Rating (ACR)*	**Descrição Baseada na Degradação *Degradation Category Rating (DCR)***	**Escore de Qualidade *MOS – mean Opinion Score***
Excelente	Imperceptível	5
Boa	Perceptível mas que não incomoda	4
Aceitável	Perceptível com leve incômodo	3
Fraca	Incômoda mas sem objeção	2
Ruim	Muito incômoda ou com várias objeções	1

Comparação de alguns Codificadores de Voz



Escore Subjetivo Comparativo de Qualidade Voz - *Comparison Category Rating (CCR)*

Descrição	Escore de Qualidade
Muito Melhor	+3
Melhor	+2
Um pouco melhor	+1
Semelhante	0
Um pouco pior	-1
Pior	-2
Muito Pior	-3

Medidas de Qualidade Objetivas ITU-T

Perceptual Speech Quality Measure (PSQM) (ITU-T 1996)

Perceptual Evaluation of Speech Quality (PESQ) (ITU-T 2001)

- correlaciona-se bem com medidas subjetivas,
- mesmo com atraso na rede (inclusive atraso variável)
- mas é intrusiva, pois necessida do sinal de referência
- Opção Não intrusiva: Recomendação P.563 (ITU-T 2004)

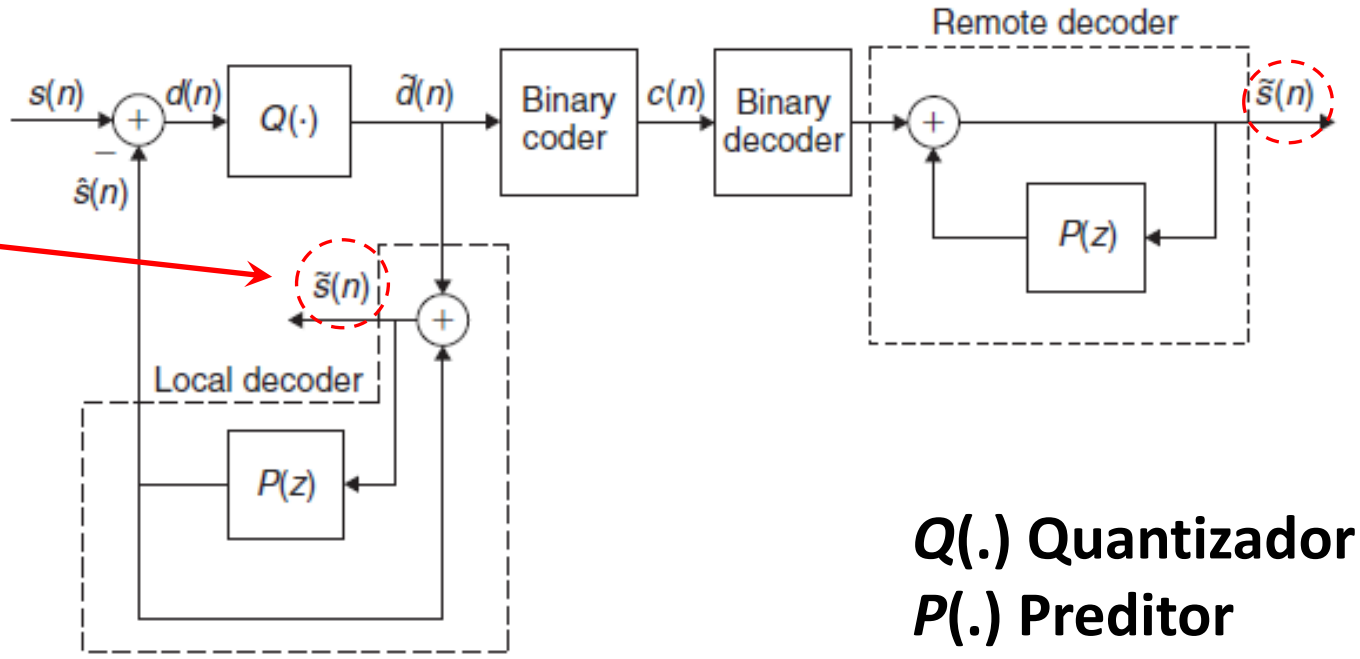
Wideband PESQ (ITU-T 2005)

Taxas de Bit para sinais de Áudio

Descrição	Largura de Banda	Frequência de Amostragem	Bits por Amostra	Taxa
Voz Faixa Estreita *Narrowband (NB) speech*	300 Hz–3.4 kHz	8.0 kHz	16	128 kbit/s
Voz Faixa Larga *Wideband (WB) speech*	50 Hz–7.0 kHz	16.0 kHz	16	256 kbit/s
Voz em Faixa Ultra Larga *Super-wideband speech*	50 Hz–14.0 kHz	32.0 kHz	16	512 kbit/s
Audio (formato CD)	10 Hz–20.0 kHz	44.1 kHz	16	706 kbit/s
Audio (formato DAT)	10 Hz–20.0 kHz	48.0 kHz	16	768 kbit/s

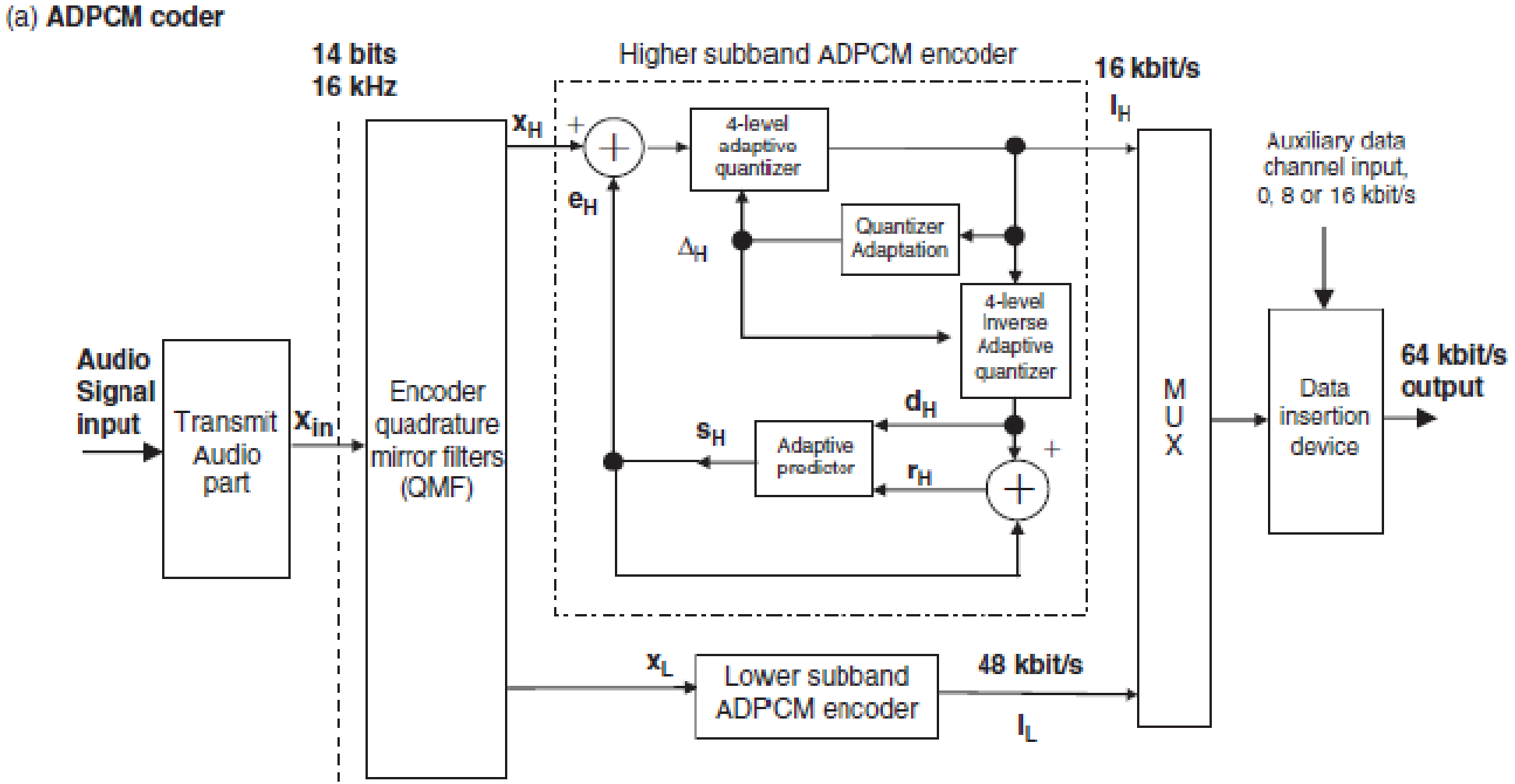
Codificador de Forma de Onda: PCM Diferencial (DPCM)

Produz no Codificador uma réplica do sinal recuperado

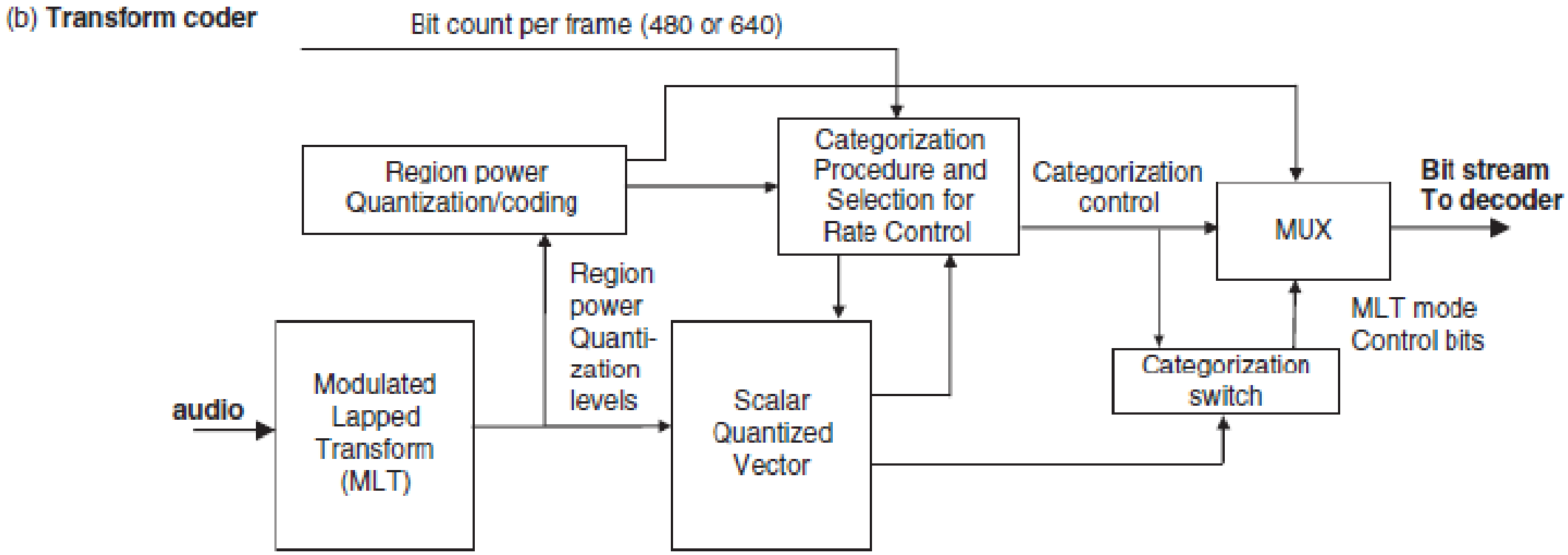


***Q*(.) Quantizador**
***P*(.) Preditor**

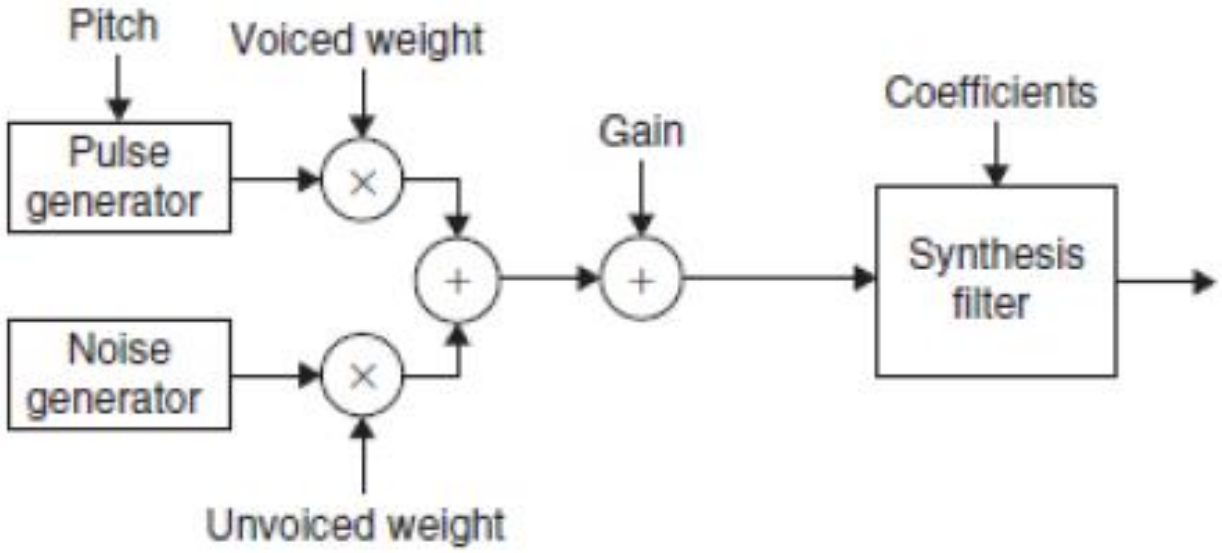
ADPCM – Adaptive Differential PCM, G722



G722.1 – Codificador por Transformada



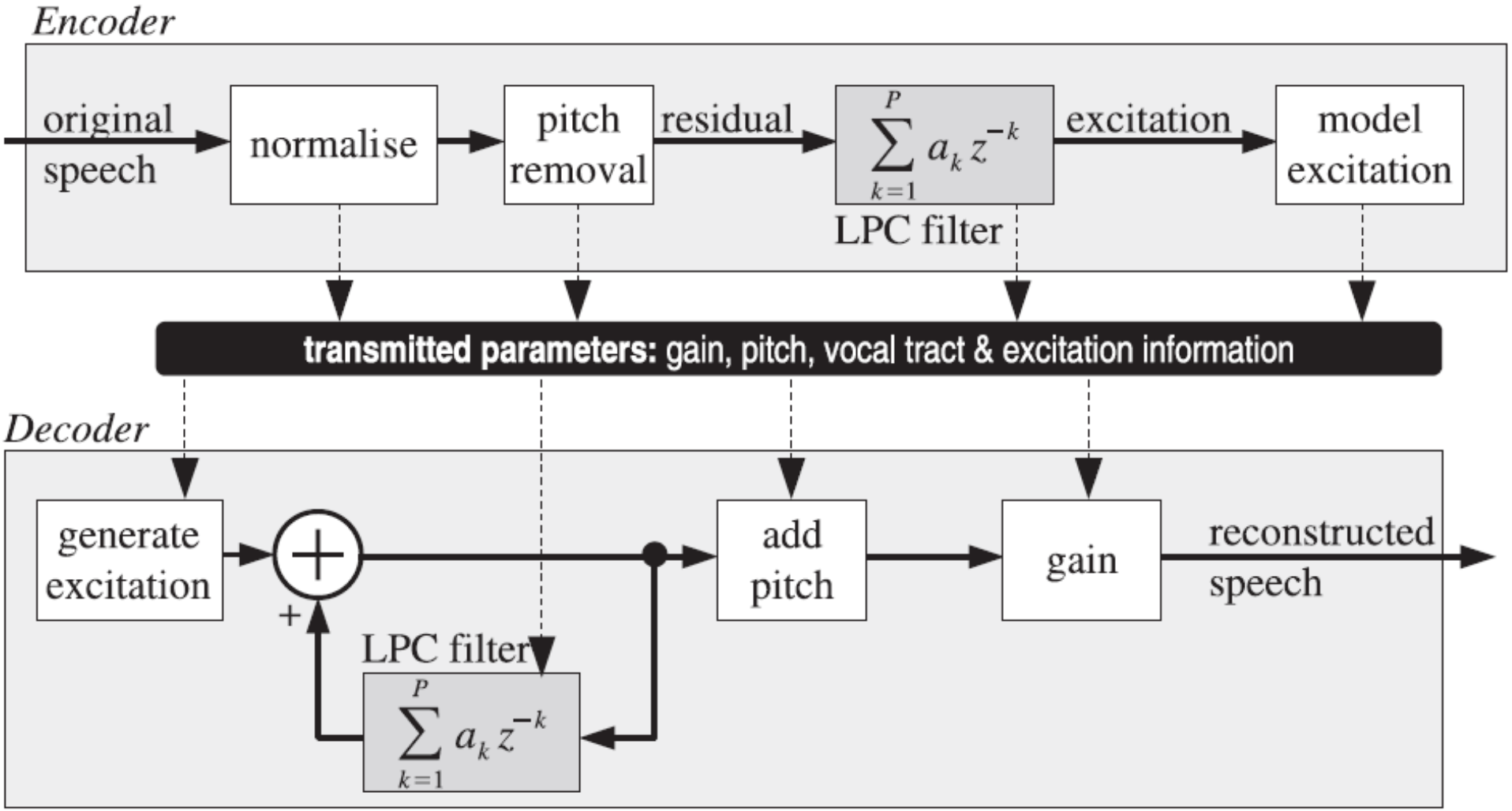
Codificação por Predição Linear (LPC)



Modelo Fonte + Filtro:

- Fonte de Excitação pode ser uma combinação linear entre uma sequência pseudo-aleatória ou pulsos periódicos de período igual ao Pitch (frequência fundamental)
- A amplitude depende do Ganho.
- Os coeficientes do Filtro são atualizados a cada configuração do trato vocal

Codificador e Decodificador LPC (p=10, FS1016)



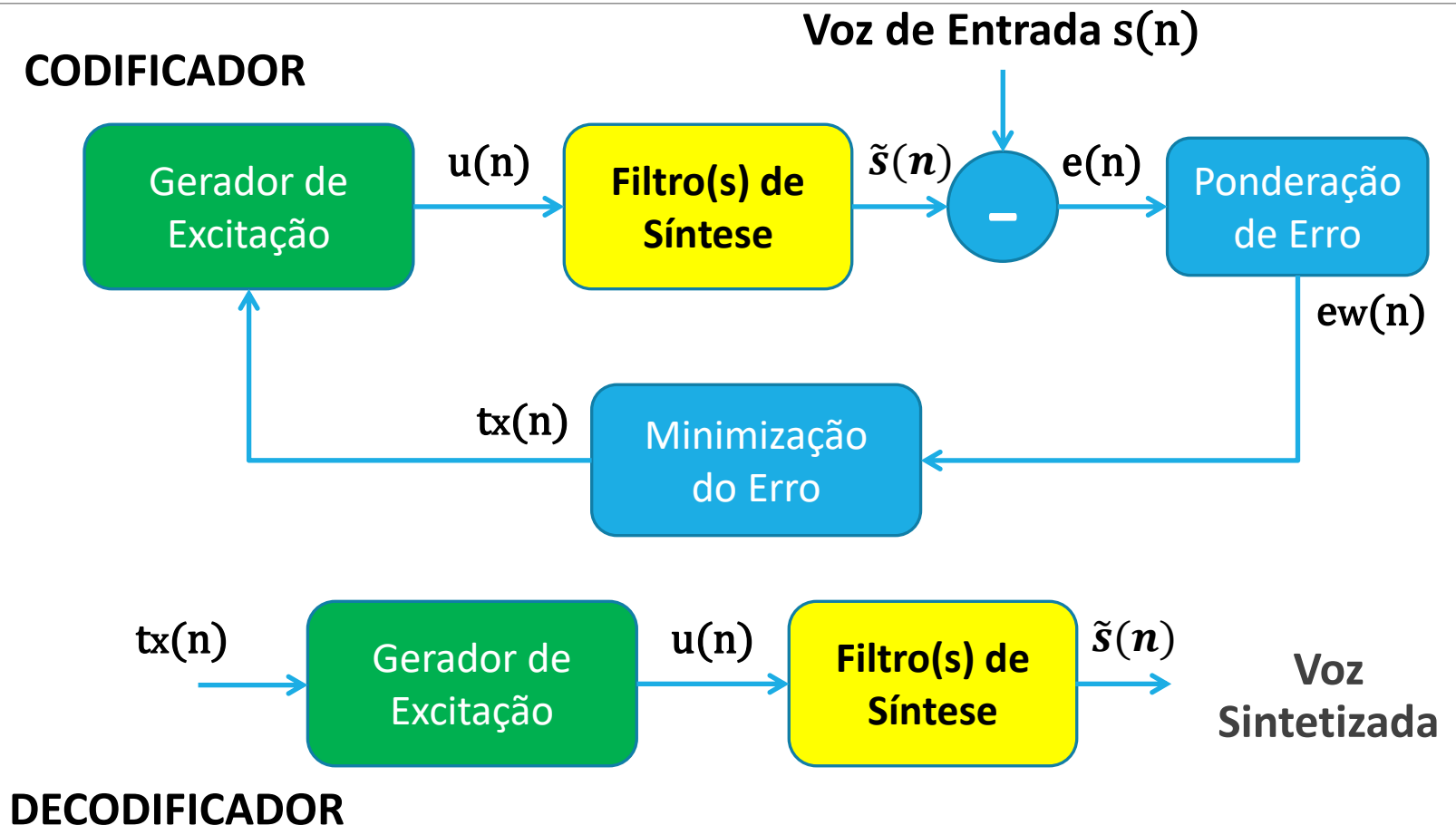
Codificação Análise-Por-Síntese (ApS):

Codificadores de Forma de Onda (p.ex. Delta e Sub-Bandas) perdem qualidade para taxas abaixo de 16 kbit/s

Codificadores por Predição Linear operam em taxas bem baixas (2 kbit/s) mas não possuem qualidade telefônica (FS1010 e FS1016).

Na tentativa de obtenção de uma codificação com boa qualidade telefônica (toll quality), em torno de 10kbit/s, surgiu o modelo de Análise-Por-Sintese (Atal)

Modelo Geral ApS Predição Linear



Modelo Análise-por-Síntese (ApS)

Filtro de Síntese

- Filtro variante no tempo só de pólos que modela a envoltória do espectro de curto prazo da voz
- Chamado também de filtro de correlação de curto prazo, pois os coeficientes são de predição linear.
- Podem ser dois (um de Longo Prazo em cascata)

Gerador de Excitação

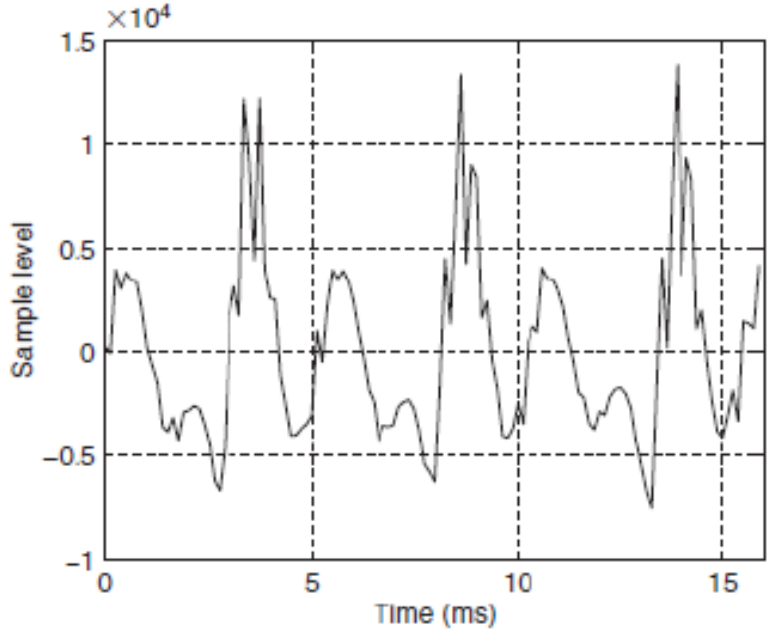
- Produz as sequências que alimentam o filtro, num Loop de Ponderação do Erro

Critério de Minimização do erro

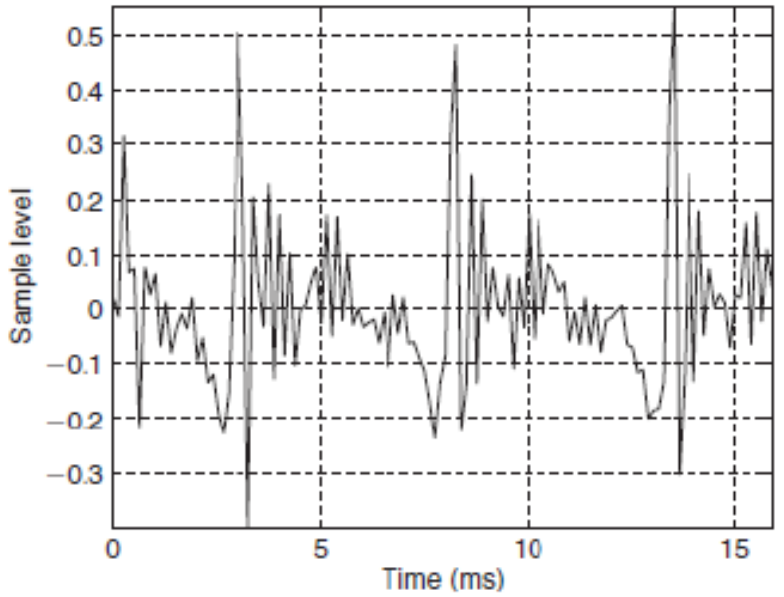
- Minimiza a diferença entre o sinal original e a Síntese

Predição Linear: Sinal e Resíduo

Forma de Onda do Sinal Original (a ser analisado)



Resíduo da Predição Linear de ordem 10.



Etapas de Codificação

1. Filtro de Síntese é calculado (10-30ms de voz) fora do loop de otimização
2. A sequência de excitação para o filtro é determinada pelo critério de erro poderado
3. Os coeficientes quantizados do filtro e da excitação são enviadas para o Receptor

Decodificação

A sequência de excitação recebida é filtrada pelo filtro recebido, para geração do sinal sintetizado

Preditor de Curto Prazo: Modela a envoltória do espectro da voz

Num segmento de tamanho de N amostras, a função de transferência do filtro só-de-pólos de ordem *p*, pode ser descrita por:

$$H(z) = \frac{1}{1-P(z)} = \frac{1}{1-\sum_{k=1}^{p} a_k z^{-k}} \qquad (1)$$

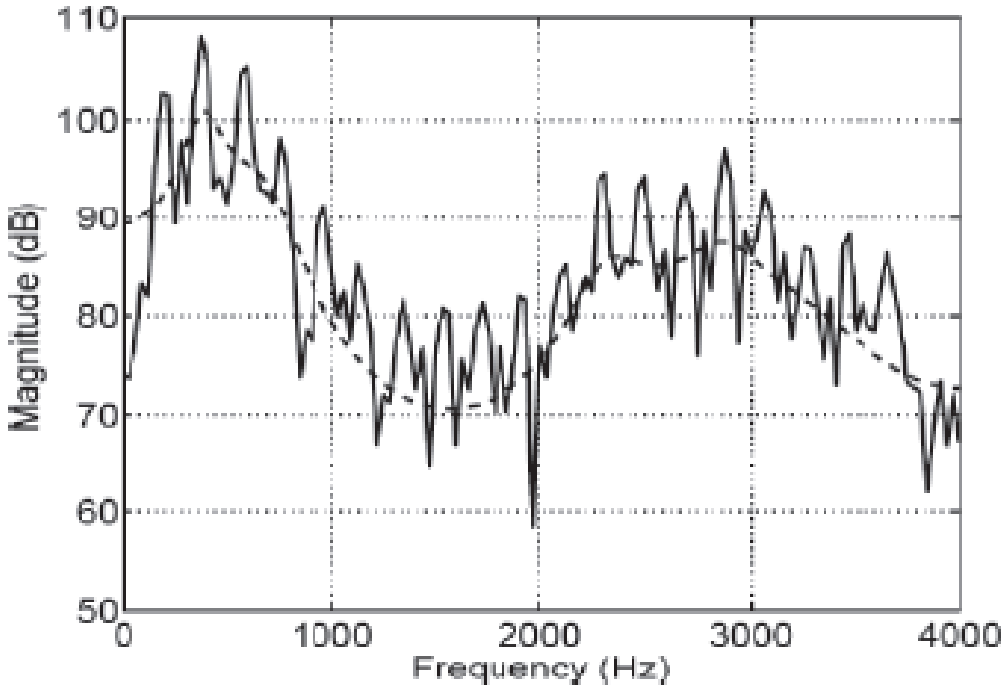
Onde:

$$P(z) = \sum_{k=1}^{p} a_k z^{-k} \qquad (2)$$

é o preditor de curto prazo

Os coeficientes $\{a_k\}$ são calculados pelo método da Predição Linear, por isso são chamados de parâmetros LPC, ou coeficientes de predição de ordem *p*.

Preditor de Curto Prazo: Modelagem da Envoltória Espectral



Densidade Espectral de Potência de um sinal de voz sonoro (Voiced) modelado por uma Predição Linear de Ordem 10 (LPC10)

Equações da Análise por Predição Linear

Amostra de voz no instante n aproximada por combinação linear de p amostras passadas:

$$\tilde{s}(n) = \sum_{k=1}^{p} a_k s(n-k), \qquad (3)$$

Onde $s(n)$ é a amostra de voz e $\tilde{s}(n)$ é a amostra predita no instante n.

Resíduo, ou Erro de Predição é definido como:

$$e(n) = s(n) - \tilde{s}(n) \qquad (4.1)$$

$$e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \qquad (4.2)$$

Pela TZ inversa:

$$E(z) = S(z)A(z) \qquad (5)$$

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \qquad (6)$$

$A(z)$ é Filtro Inverso, de $H(z)$ na equação (1)

Soluções de curto prazo, por Minimização do Erro Quadrático Médio (MMSE)

$$\sum_{k=1}^{p} a_k \varphi(i,k) = \varphi(i,0), \qquad i=1,\ldots,p \qquad (7)$$

Onde $$\varphi(i,k) = \sum_n s(n-i)s(n-k) \qquad (8)$$

O conjunto de p equações determinadas, com os limites da somatória em (8), temos:

1. $-\infty<n<\infty$, **método da Autocorrelação janelado**, com cálculo eficiente pelo algoritmo de Levinson-Durbin (equações normais, Yule-Walker). Sempre possui solução estável para H(z).
2. $0 \leq n \leq N-1$, **método da Covariância**, preciso, mas nem sempre estável para H(z). Uma variação é denominada de método da Covariância estabilizado.
3. Solução recursiva, **método Lattice** (Algoritmo de Burg), boa para detecção de sinais impulsivos.

Algumas Considerações:

i. Ordem de predição p geralmente entre 8 e 16.

ii. O valor inferior está relacionado ao tempo da onda sonora ir e voltar ao percorrer o trato vocal (~1ms); o limite superior deve-se à saturação do ganho de predição e à eficiência de cálculo.

iii. Pela rapidez e estabilidade, o método da autocorrelação é muito empregado. No algoritmo de Levinson-Durbin, uma recursão se atualiza com um parâmetro intermediário k_i, denominado ***coeficiente de reflexão, ou Coeficiente de Correlação Parcial***:

$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)} \qquad (9)$$

Representação Matricial (Equações Levinson ou Yule-Walker)

$$\mathbf{R}a = -r \qquad (10)$$

Onde:

- **R** é a matrix (pxp) Toeplitz de autocorrelação, com elementos
$$[R]_{ij} = R(i-j) \qquad (11)$$
- r é o vetor de autocorrelações com elementos $r(i) = R(i)$
$$r[|m|] = \sum_{n=0}^{M-|m|} x[n]x[n+|m|] \qquad (12)$$
- a é o vetor dos coeficientes da Predição Linear

Cuja solução é obviamente

$$a = -\mathbf{R}^{-1}r \qquad (13)$$

➢ Resolvida de forma eficiente computacionalmente pelo algoritmo de Levinson-Durbin

Exemplo: Predição Linear de ordem p=2.

Sejam as amostras: `x(i) = {-1.6117  0.1983  -0.9080  1.4386  0.8024}`

Para i = 0, 1, 2 e usando a expressão (12), profundidade M = 2.

$r(0) = x(0).x(0) + x(1).x(1) + x(2).x(2) = (-1.6117)^2 + (0.1983)^2 + (-0.9080)^2 = 3.4614$

$r(1) = x(0).x(1) + x(1).x(2) = (-1.6117*0.1983) + (0.1983*-0.908) = -0.4997$

$r(2) = x(0).x(2) = (-1.6117*-0.9080) = 1.4634$

No vídeo estava trocado este valor!

$$\mathbf{r'} = [r(1)\ r(2)]^T$$

$$\mathbf{R} = \text{Toeplitz}(r(0),r(1)) = \begin{bmatrix} r(0) & r(1) \\ r(1) & r(0) \end{bmatrix} = \begin{bmatrix} 3.4614 & -0.4997 \\ -0.4997 & 3.4614 \end{bmatrix}$$

Equações Normais, usando (10):

$$\mathbf{R}\boldsymbol{a} = -\mathbf{r}'$$

$$\begin{bmatrix} 3.4614 & -0.4997 \\ -0.4997 & 3.4614 \end{bmatrix}\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = -\begin{bmatrix} -0.4997 \\ 1.4634 \end{bmatrix} \Rightarrow \begin{bmatrix} 3.4614 & -0.4997 \\ -0.4997 & 3.4614 \end{bmatrix}\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0.4997 \\ -1.4634 \end{bmatrix}$$

Como,

$$\mathbf{R}^{-1} = \begin{bmatrix} 0.2951 & 0.0426 \\ 0.0426 & 0.2951 \end{bmatrix}$$

Sinais trocados no vídeo

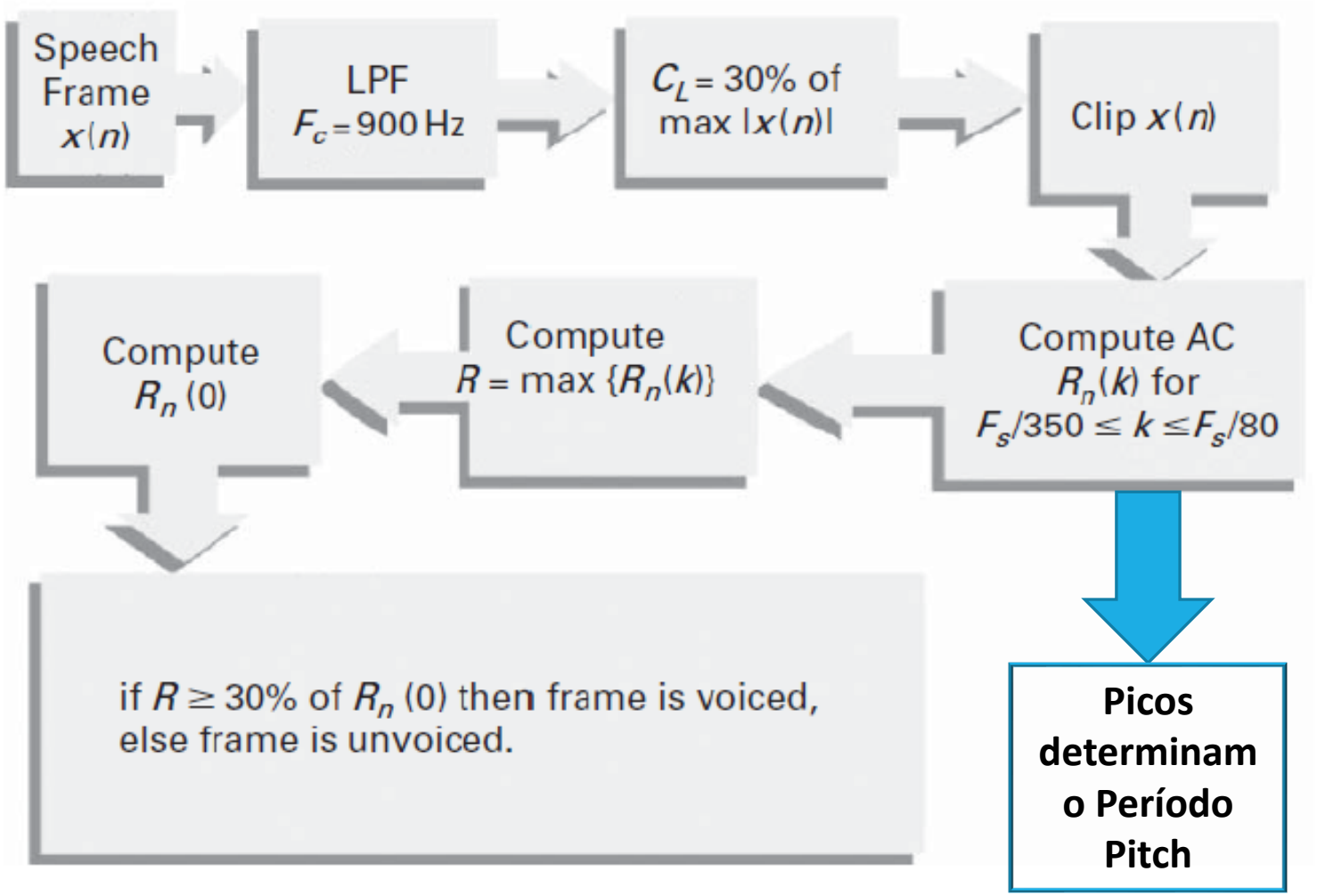
Logo:

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0.2951 & 0.0426 \\ 0.0426 & 0.2951 \end{bmatrix}\begin{bmatrix} 0.4997 \\ -1.4634 \end{bmatrix} \Rightarrow \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0.0851 \\ -0.4105 \end{bmatrix}$$

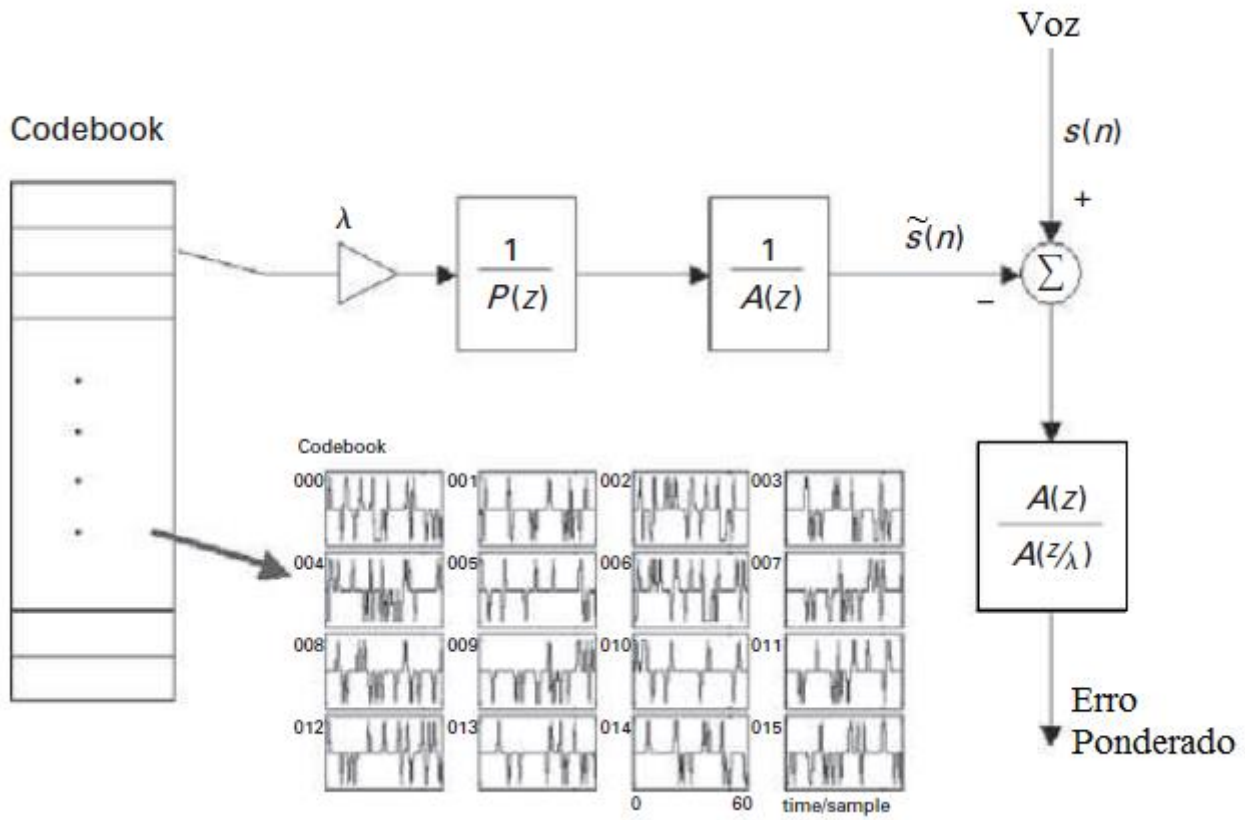
O Preditor de Longo Prazo (LTP) e chave V/UV

Efetua a predição do período de Pitch, modelando a estrutura fina do espectro

Acerta a ponderação da Fonte de Excitação(U/UV) Voiced/Unvoiced (Sonora/Surda)



Busca de Solução LP ótima:
Busca Exaustiva nas possíveis Excitações *(Codebook)*



Onde:

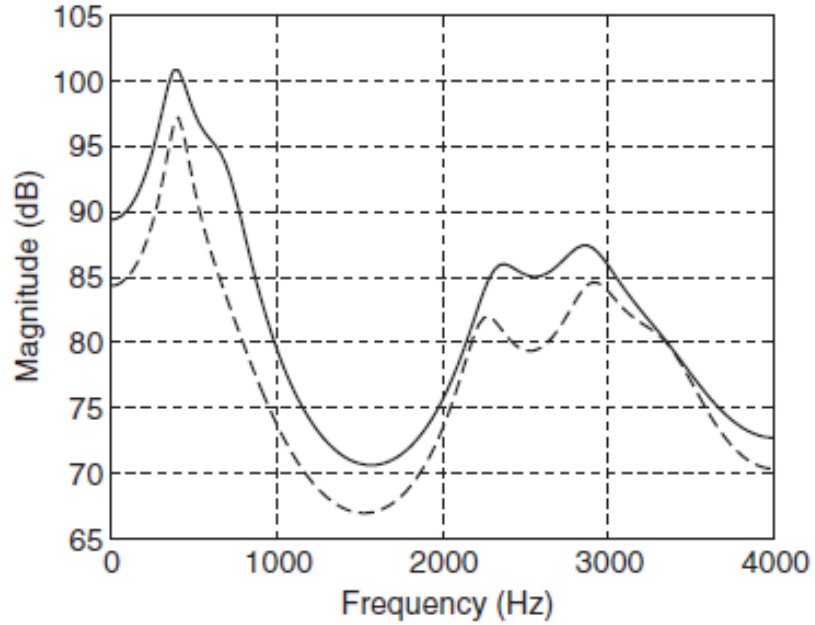
Codebook: Livro de Códigos de Possíveis Excitações Quantizadas

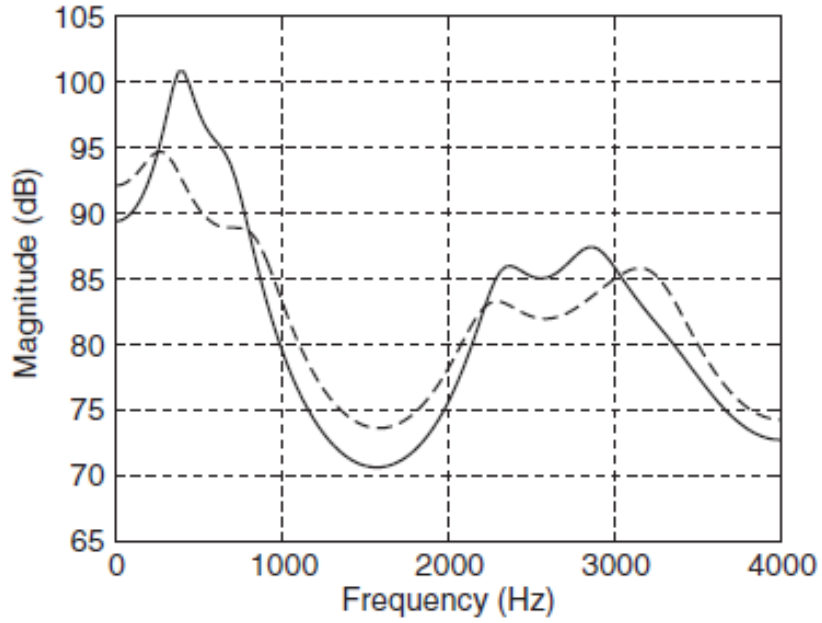
$P(z)$, Filtro de Predição de Longo Prazo (LTP), recupera u(n) de e(n) via 1/P(z)

1/A(z) Filtro de Síntese, para obtenção de $\tilde{s}(n)$, e comparar com $s(n)$

A(z)/A(z/λ) Filtro de Ponderação para determinar o erro de codificação:

- Atribui menor erro nas regiões onde o sinal possui baixa densidade de potência e atribui maior erro onde a densidade de potência do sinal é alta, ou seja, com melhores condições de mascarar o erro.



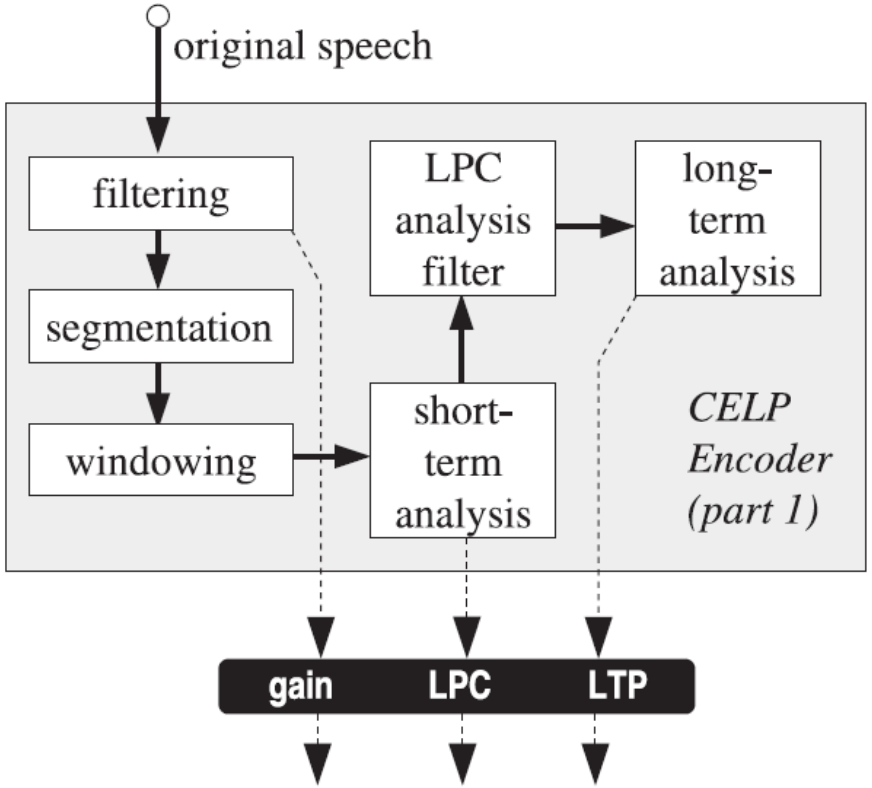


PONDERAÇÃO (λ)

A envoltória do erro de reconstrução spectral (pontilhada) fica abaixo do sinal no Loop ApS

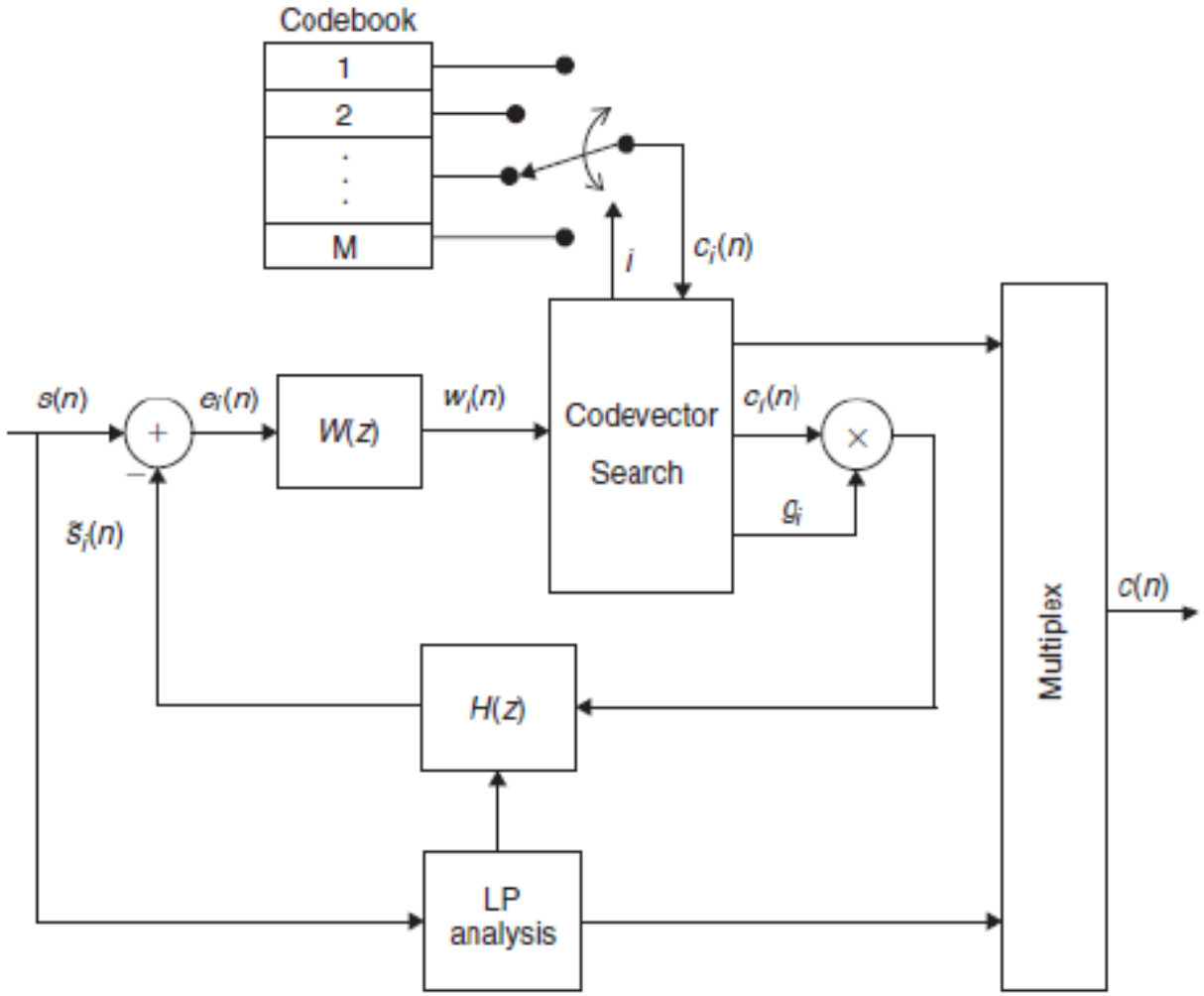
Em outros momentos a envoltória do erro de reconstrução spectral pode ficar acima do sinal (entre 900 e 2kHz)

CELP - Code-excited linear prediction
Solução ótima para LP Análise-Por-Síntese



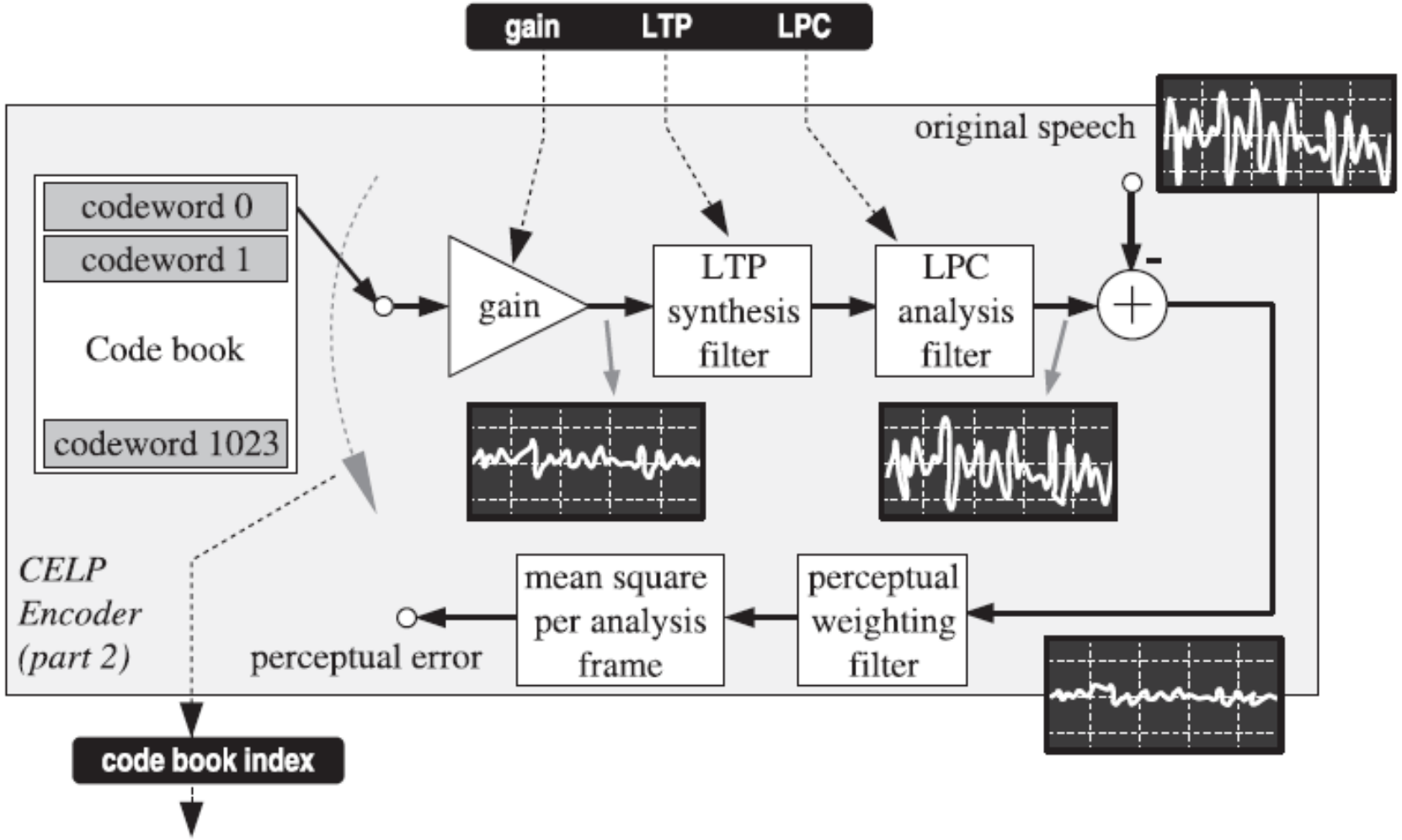
Cálculo dos Parâmetros de Predição Linear

CELP – Code Excited Linear Prediction

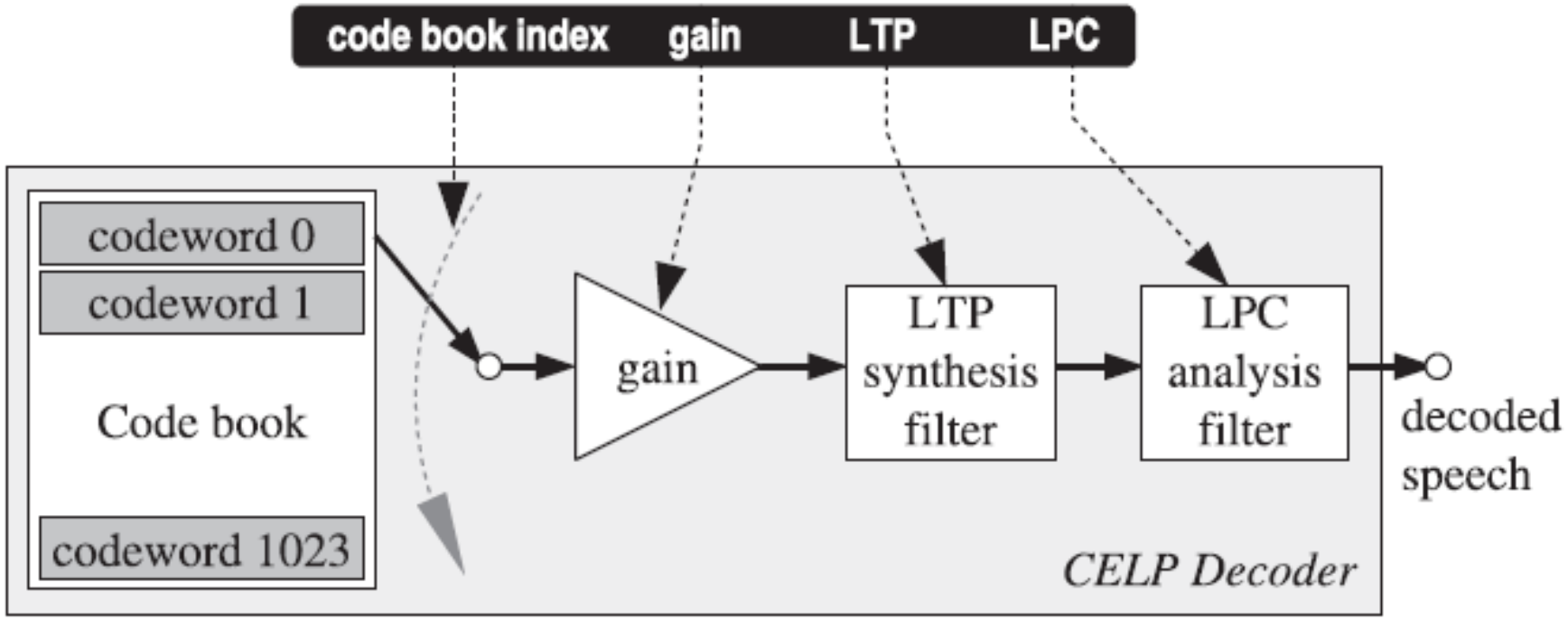


- Codificador CELP codifica o sinal s(n) na sequência de bits c(n)
- O Codificador inclui um loop de Análise-Por-Síntese, alimentado por um sinal de erro e(n)
- O Codebook modela os tipos de excitação

Code-excited linear prediction (CELP) – Loop ApS



Code-excited linear prediction (CELP) - Decoder



Transformação Coeficientes LP na quantização

Para Coeficientes de Reflexão, ou de correlação parcial: k_i, para modelagem de tubos acústicos do trato vocal:

$|k_i < 1|$, procedimento step-down (Levinson-Durbin)

Para θ_i arco-seno dos k_is: $\theta_i = \arcsin(k_i)$

Log-Area Ratio (LAR), $lar_i = \frac{1-k_i}{1+k_i}$

Line Spectral Pair (LSP) ou Line Spectral Frequency (LSF), raízes dos polinômios, modelo de glote totalmente aberta ou fechada:

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

$$\mathrm{Q}(z) = A(z) - z^{-p}A(z^{-1})$$

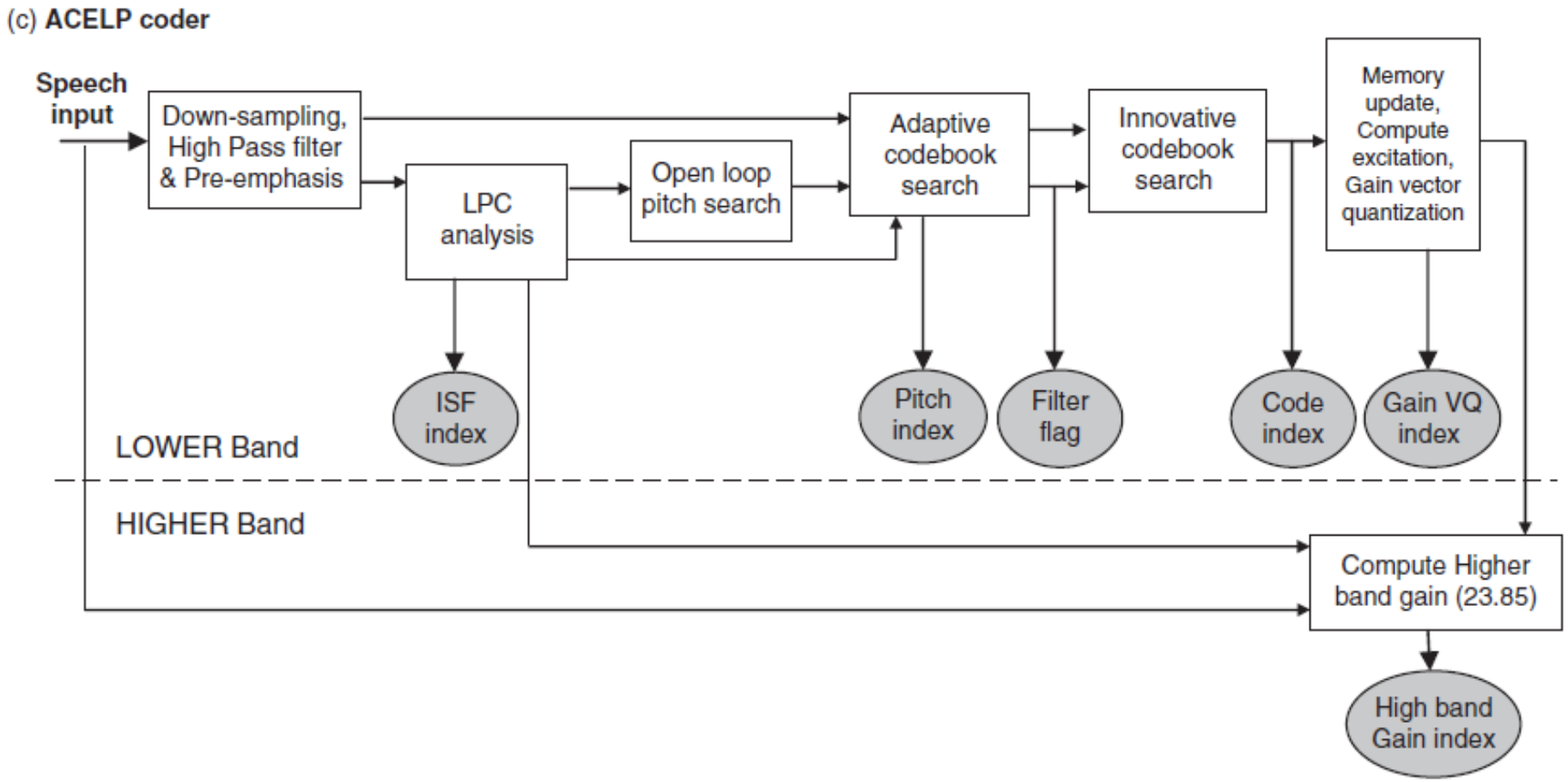
Nada é perfeito: custo computacional

Busca Exaustiva gera atraso demasiado, inviabilizando implementação ótima.

LD-CELP (low-Delay), buscam soluções viáveis, sub-ótimas:

- IS-54, TDMA, usa o VSELP (Vector-Sum Codebook), quadros de 20ms
- G723.1 – MPE, Multi-Pulse Excitation, quadros de 30ms, LPC10 em coeficientes LSPs
- G729, G722.2 (AMR-WB), ACELP – Algebraic CELP, Codebooks esparsos, isto é, com muitos coeficientes nulos, e os demais +1 ou -1, para minimizar as operações (não necessita multiplicações).

G722.2 (AMR-WB) - ACELP



Questões teóricas da P2b (duas semanas)

1. Qual a diferença entre uma medida de qualidade subjetiva de voz e outra objetiva? Cite dois exemplos de cada, explicando como se diferenciam entre si.

2. Explique as diferenças entre os codificadores G722, G722.1 e G722.2.

3. O que é a Predição Linear? Explique um modelo de Codificação de Voz por LPC tanto no tempo quanto na frequência.

4. Explique no que consiste a Codificação Análise por Síntese no CELP.

5. O que é um Codebook, e como ele é gerado?

6. Cite algumas versões de codificadores com LP.

Questões da P2b:

Sejam os 4 últimos pares de números do seu RA = $R_1R_2R_3R_4$, p.ex. se RA = 11075113, R_1= 11, R_2= 07, R_3= 51, R_4= 13.

---

5. Sejam as amostras com janela retangular, do sinal $x(i)$, f_a = 8 kHz, para i = 0, 1,...,15:

$x(i) = \{ -0.1587 \quad 0.R_1 95 \quad -0.25R_2 \quad 0.0R_3 1 \quad 1.R_4 62 \quad -1.6117 \quad 0.1983 \quad -0.9080$

$1.4386 \quad 0.8024 \quad -1.6272 \quad 0.0459 \quad -0.5898 \quad 0.5891 \quad 0.1483 \quad -0.1429\}$

a) (0,5) Determine a autocorrelação do intervalo i = 1, 2, 3 e 4, para m = 0, 1, 2 e 3, para o cálculo de um modelo LPC de ordem p=3.

b) (0,5) Escreva a formulação do problema dos coeficientes de predição com os valores da matriz de autocorrelação (equações normais, ou Yule-Walker);

c) (0,5) Resolva o problema pela inversão da matriz de autocorrelação, calculando os coeficientes LPC,

$$\underline{\boldsymbol{a}} = \{a_1, a_2, a_3\}^{\mathrm{T}}.$$

6. Selecione um trecho de um fonema de uma vogal do seu nome e outro de uma vogal diferente do seu sobrenome com uma janela de Hamming de 40ms, fs = 22050Hz e, calcule o espectro em magnitude (dB x frequência) e sobreponha o espectro modelado LPC, com p=12. Envie os arquivos gravados com o seu nome e sobrenome.

a) (0,5) Plote as formas de onda analisadas (janeladas), os espectros delas e as envoltórias LPC de cada espectro (dica, veja o slide 20)

b) (0,5) Forneça os coeficientes LPC de ordem 12 e o ganhos de predição linear (resíduo do modelo, derivado de (4.2)) de cada análise.

Referências

- RAMIREZ, M. A.; MINAMI, M**., Technology and Standards for Low-Bit-Rate Vocoding Methods**, in: Handbook of Computer Networks: LANs, MANs, WANs the Internet and Global, Cellular, and Wireless Networks, Volume 2, Wiley, Ch89, 2008.
- RAMIREZ, M. A.; MINAMI, M. and SREENIVAS, T. V., **Models for Speech Processing,** in: Signals and Images: Advances and Results in Speech, Estimation, Compression, Recognition, Filtering, and Processing, CRC Press, 2015.
- HWANG, J.-N., **Multimedia Networking: From Theory to Practice**, Cambridge University Press, Ch2, 2009.
- MCLOUGHLIN, I., **Applied Speech and Audio Processing: With MATLAB Examples**, Cambridge University Press, Ch5, 2009.