Technion - Israel Institute of Technology
Department of Electrical Engineering

Signal and Image Processing Laboratory

DATA EMBEDDING IN SPEECH AND AUDIO SIGNALS

Ariel Sagi

Supervisor: Prof. David Malah

April-2004

OUTLINE

- Background
- Scalar Costa Scheme
- Data-embedding in speech and audio signals
- Application: speech bandwidth extension
- Summary and future research

BACKGROUND

- Data-embedding Vs. Watermarking

- Data-embedding system requirements
  - **Transparency, Robustness, Rate**

- Applications
  - Additional payload, embedding data in an analog signal, …

- Existing methods for data embedding
  - **Spread Spectrum** watermarking schemes with correlation based detection suffer significantly from host signal interference
  - **Informed Embedding:** Considering the host signal as side-information to the encoder
    - Quantization index modulation (QIM), Dither modulation (DM) – Chen & Wornell, 1998
    - Scalar Costa scheme (SCS) – Eggers & Girod, 2000

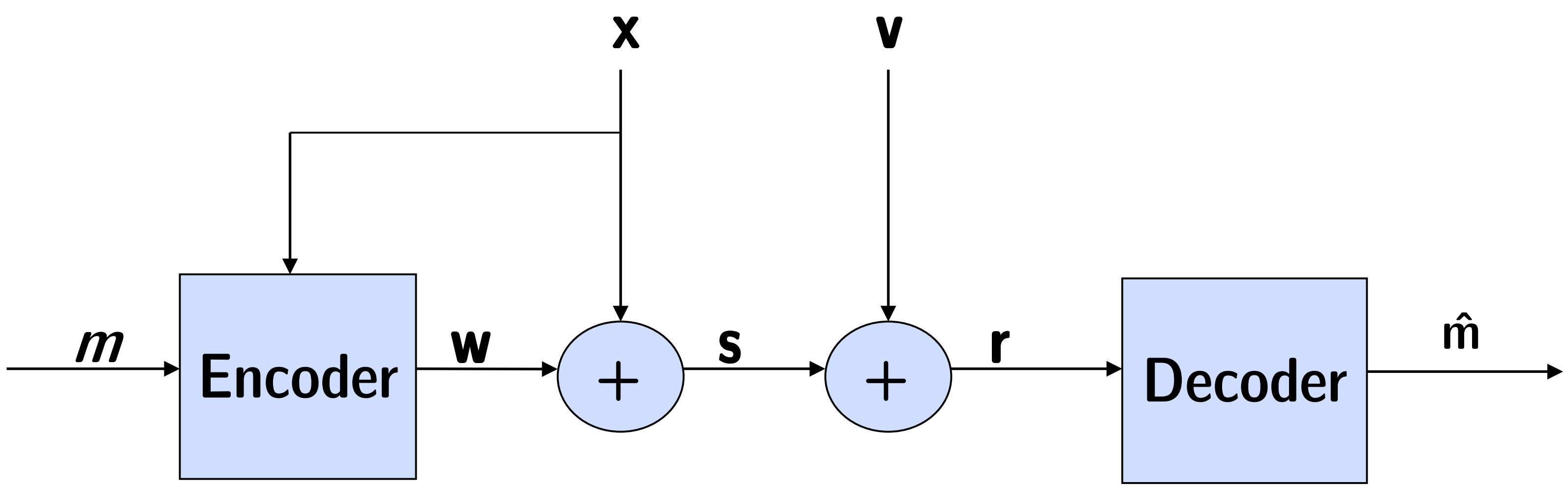
GOALS

1. Combining **informed embedding** principles with a **perceptual model** for speech and audio signal

2. Developing methods for parameter estimation, and to test the methods under degradations caused by a **telephone channel**

3. Demonstrating a possible use of embedded-data in speech, for **speech bandwidth extension**

Block diagram



Notations

m	message
$\mathbf{w}$	watermark signal
$\mathbf{x}$	host signal
$\mathbf{s}$	combined signal
$\mathbf{v}$	noise
$\mathbf{r}$	received signal
$\hat{m}$	decoded message

Definitions

$$\mathrm{WNR} = 10\log_{10}\left(\frac{\sigma_w^2}{\sigma_v^2}\right) \text{ [dB]}$$

$$\mathrm{SWR} = 10\log_{10}\left(\frac{\sigma_x^2}{\sigma_w^2}\right) \text{ [dB]}$$

- **Ideal Costa Scheme**
  Costa, 1983: "Writing on Dirty Paper", proved that for IID Gaussian host signal and IID Gaussian noise host signal interference can be completely avoided

$$C_{\mathrm{ICS}} = \frac{1}{2}\log_2\left(1 + \frac{\sigma_w^2}{\sigma_v^2}\right)$$

- **Scalar Costa Scheme**
  Eggers & Girod suggested a **suboptimal practical** embedding rule, that uses **dithered uniform scalar quantizers**

  - Encode message m in $\mathbf{d} = d_1, d_2, \ldots, d_n$, where $d \in \{0, 1, \ldots, D-1\}$
  - Embed $\mathbf{d} = d_1, d_2, \ldots, d_n$ in $\mathbf{x} = x_1, x_2, \ldots, x_n$

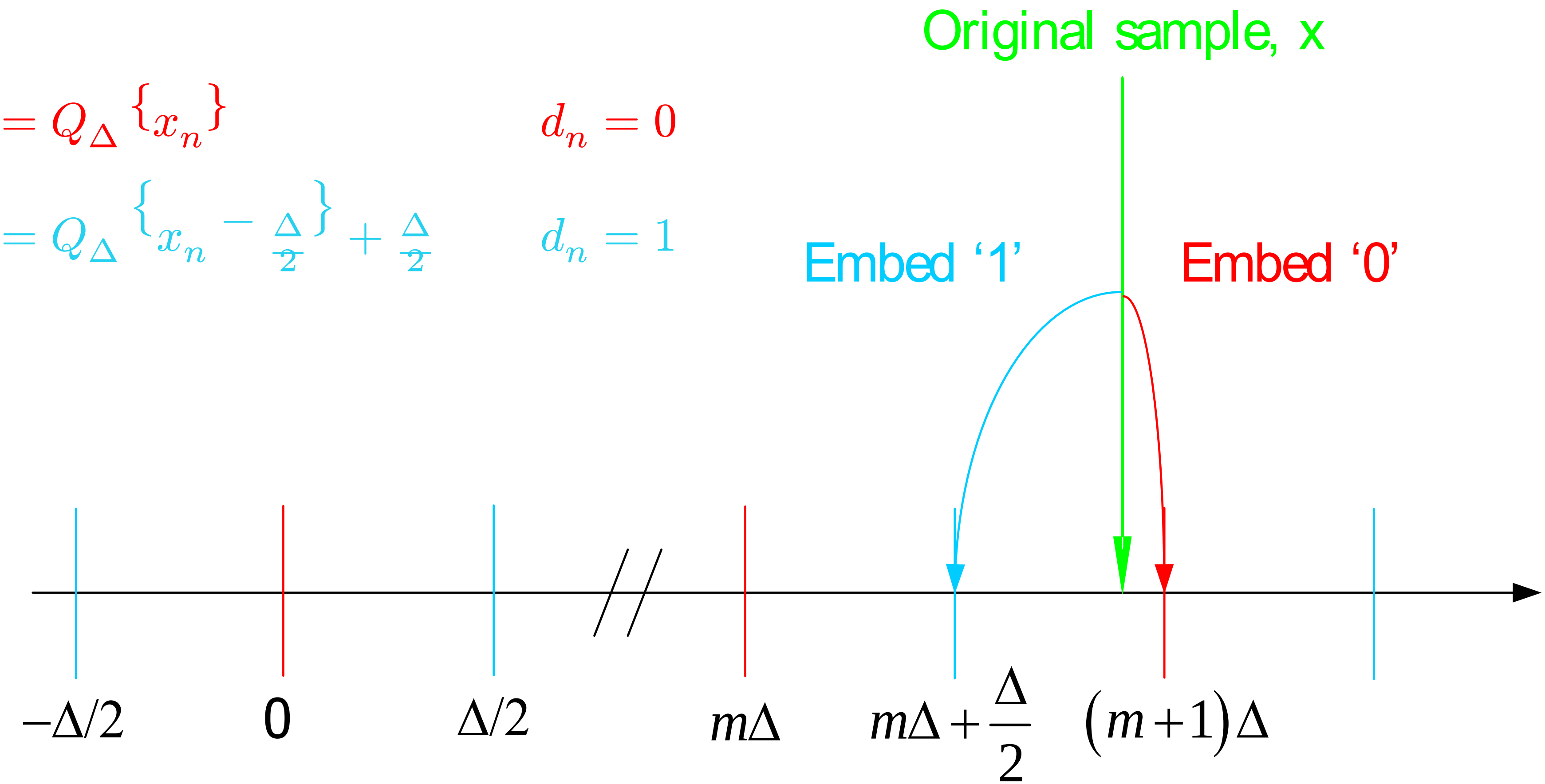
$$\boxed{s_n = (1-\alpha)x_n + \alpha\left(Q_\Delta\left\{x_n - \Delta\left(\frac{d_n}{D}\right)\right\} + \Delta\left(\frac{d_n}{D}\right)\right)}$$

Example: $\{\alpha = 1, D = 2\}$

$$s_n = Q_\Delta\left\{x_n - \Delta\left(\frac{d_n}{2}\right)\right\} + \Delta\left(\frac{d_n}{2}\right)$$

$s_n = Q_\Delta\{x_n\}$ $\quad$ $d_n = 0$

$s_n = Q_\Delta\{x_n - \frac{\Delta}{2}\} + \frac{\Delta}{2}$ $\quad$ $d_n = 1$



- The signal y_n is defined by

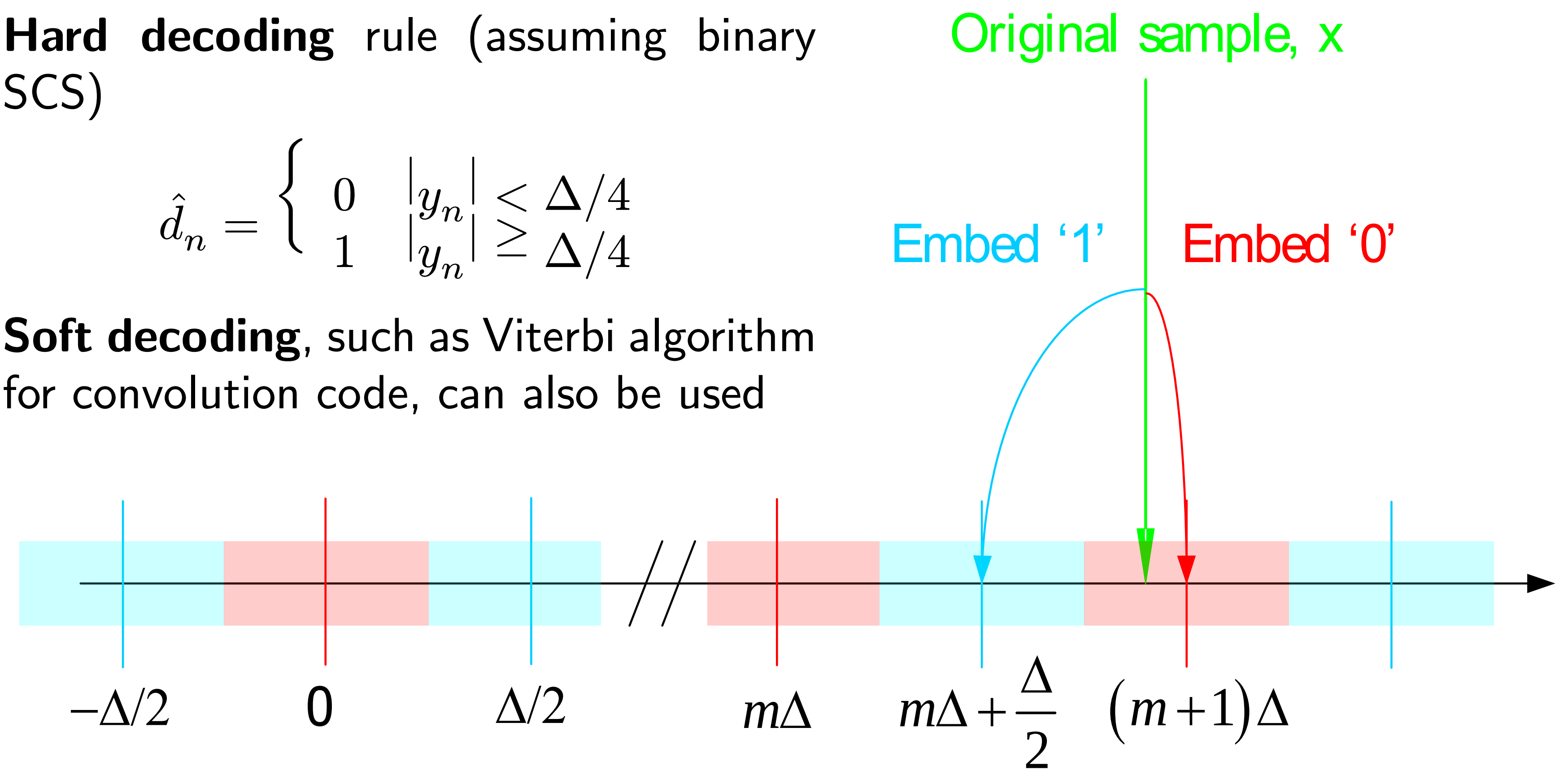
$$y_n = Q_\Delta\{r_n\} - r_n$$

and therefore $|y_n| \le \Delta/2$

- **Hard decoding** rule (assuming binary SCS)

$$\hat{d}_n = \begin{cases} 0 & |y_n| \le \Delta/4 \\ 1 & |y_n| \ge \Delta/4 \end{cases}$$

- **Soft decoding**, such as Viterbi algorithm for convolution code, can also be used



DATA-EMBEDDING PARAMETERS

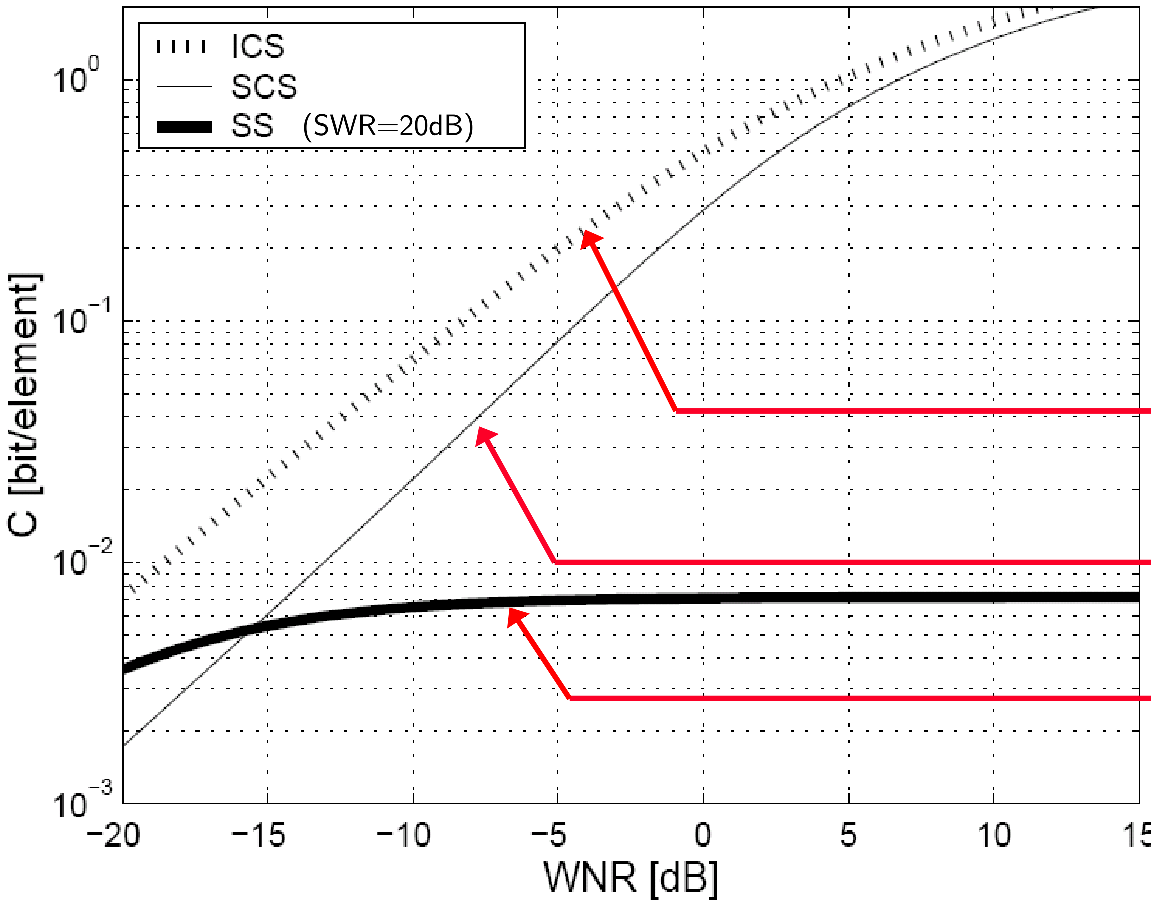
- The mean squared error distortion caused by data-embedding

$$\boxed{\sigma_w^2 = \frac{\alpha^2 \Delta^2}{12}}$$

- An approximative analytical expression for the optimum value of parameters

$$\begin{aligned} \alpha_{\text{SCS,approx}} &= \sqrt{\frac{\sigma_w^2}{\sigma_w^2 + 2.71\sigma_v^2}} \\ \Delta_{\text{SCS,approx}} &= \sqrt{12(\sigma_w^2 + 2.71\sigma_v^2)} \end{aligned}$$

SCS THEORETICAL PERFORMENCE

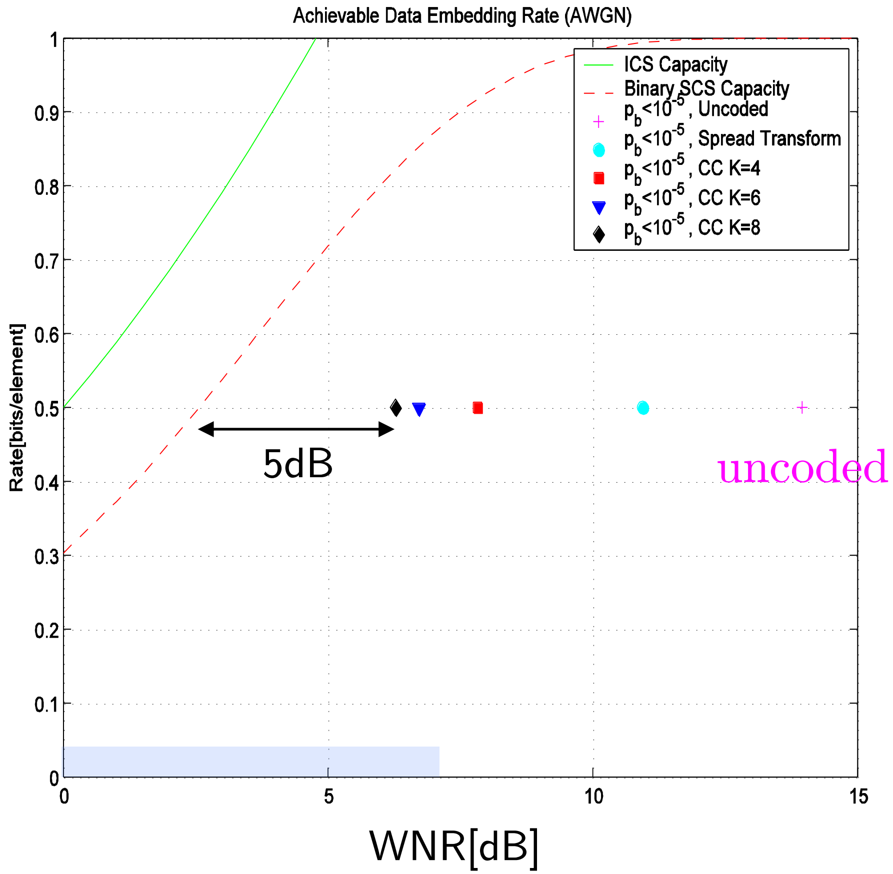


$$C_{\mathrm{ICS}} = \frac{1}{2}\log_2\left(1+\frac{\sigma_w^2}{\sigma_v^2}\right)$$

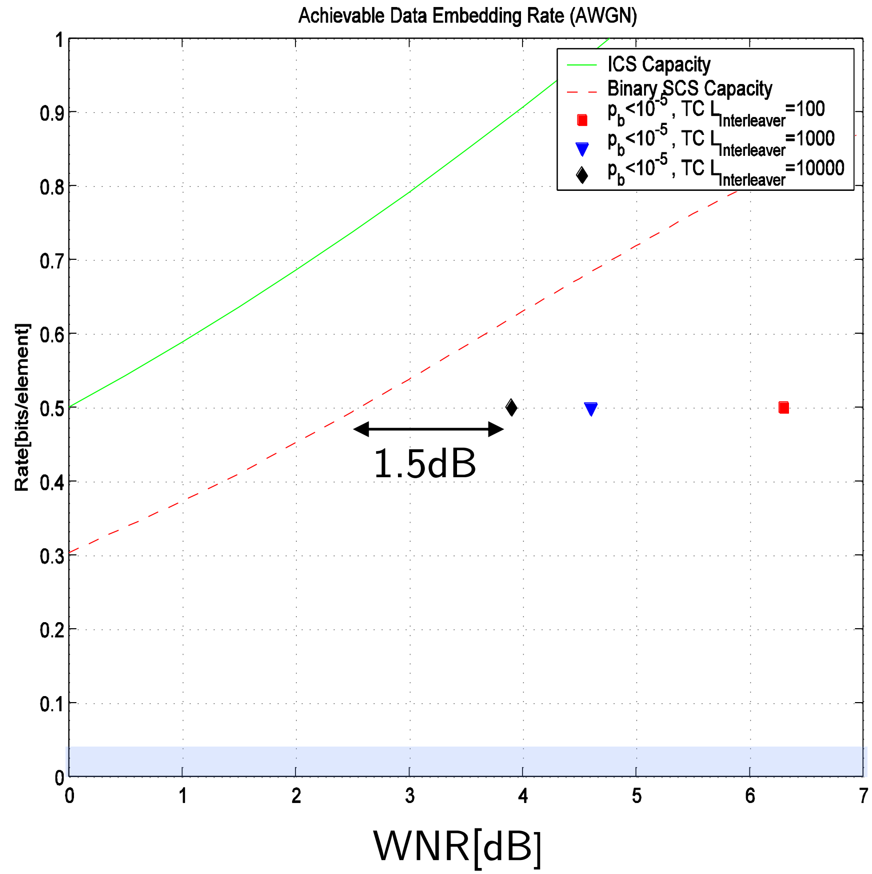
$$C_{\mathrm{SCS}}$$

$$C_{\mathrm{SS}} = \frac{1}{2}\log_2\left(1+\frac{\sigma_w^2}{\sigma_x^2+\sigma_v^2}\right)$$

- Results of using error correction coding. Code is chosen according to the application
  - Convolution codes, Block codes, Turbo codes
- Demonstrations (white Gaussian host signal, white Gaussian channel noise)

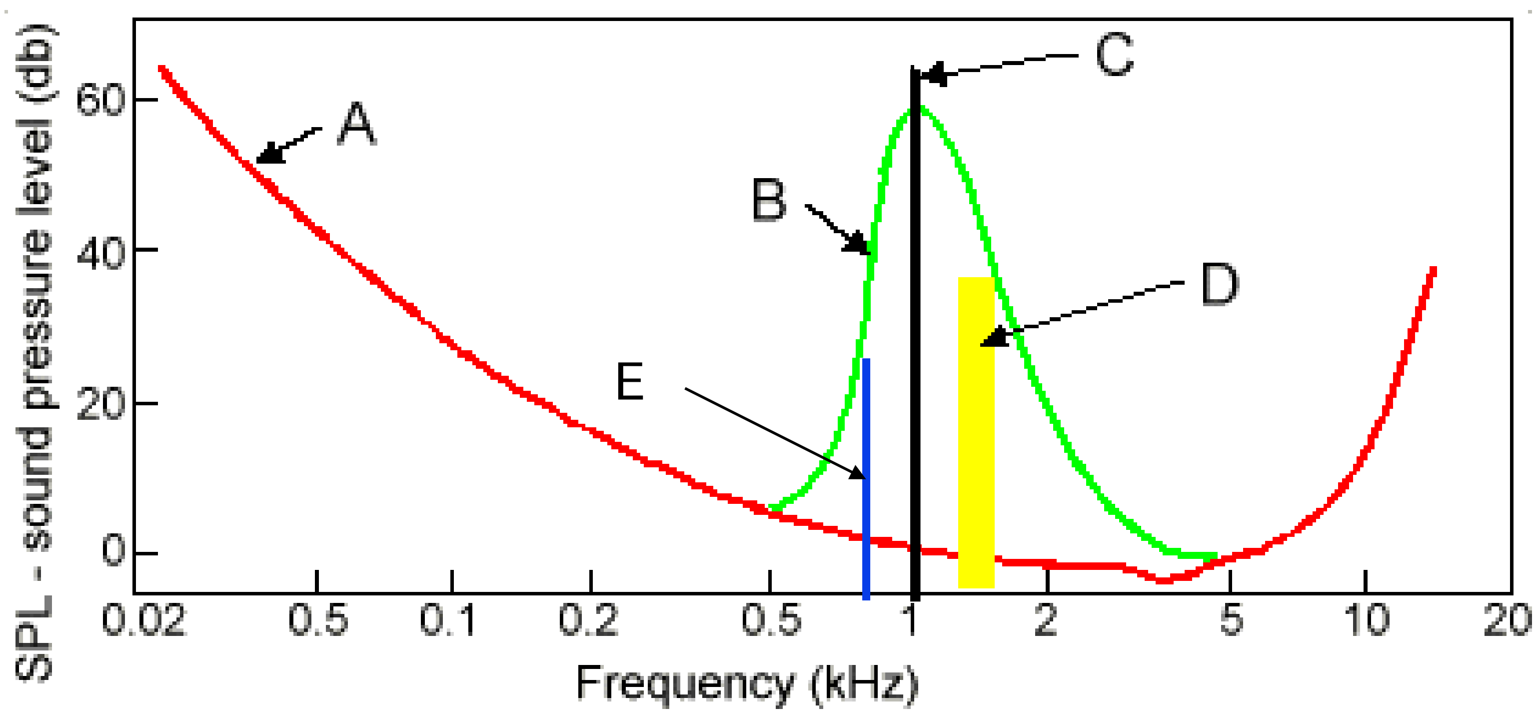


Convolution code, k=4,6,8



Turbo code, L=100,1000,10000

- Many advantages can be obtained by using the hearing system characteristics. These are used in speech and audio processing:
  - Compression, **Data-embedding**, Enhancement

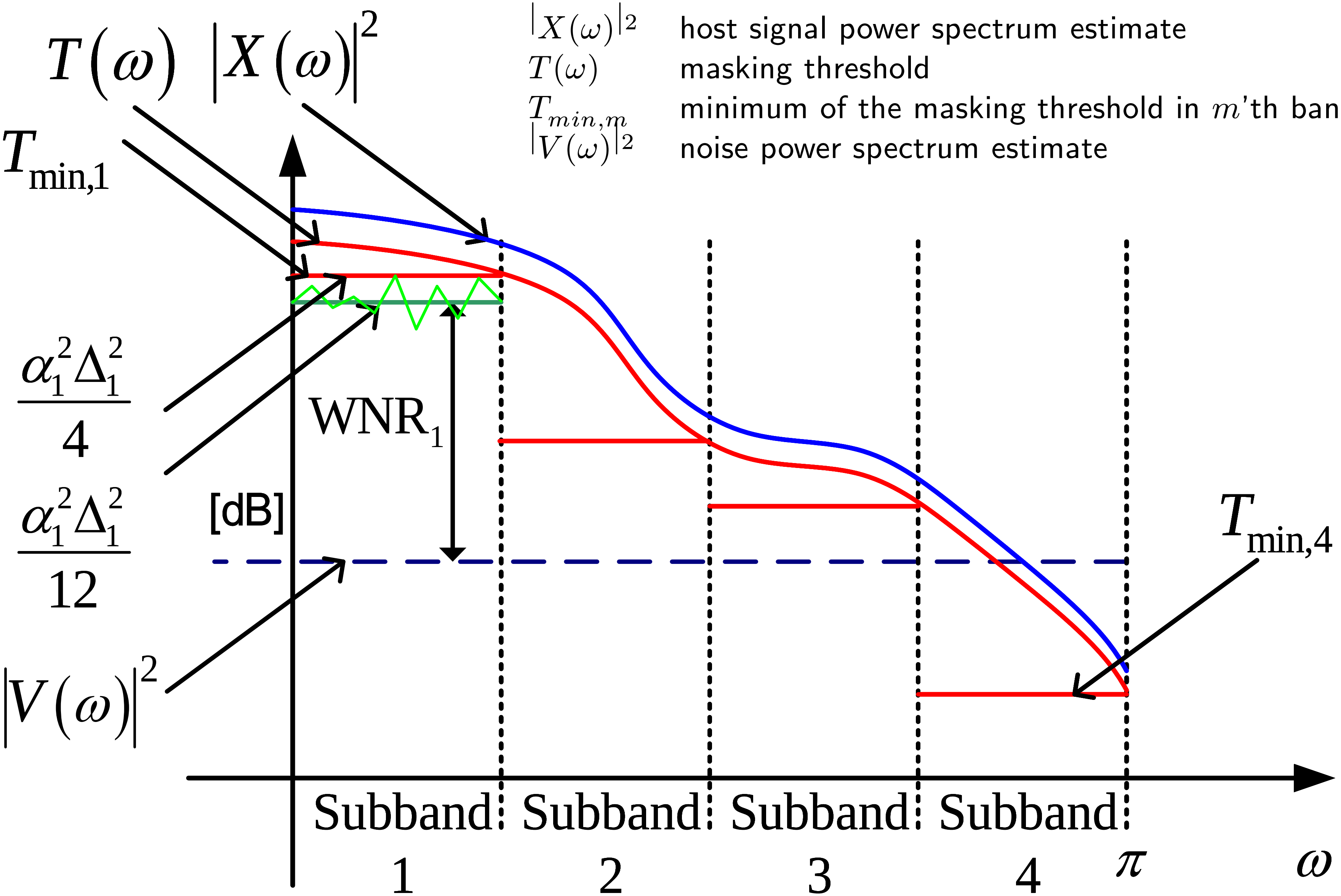


A Normal threshold of hearing

B Modified threshold due to tone C

D Band of noise rendered inaudible by the presence of tone C

E Tone E rendered inaudible by the presence of tone C



$|X(\omega)|^2$ host signal power spectrum estimate
$T(\omega)$ masking threshold
$T_{min,m}$ minimum of the masking threshold in m'th band
$|V(\omega)|^2$ noise power spectrum estimate

- The subband average embedding-distortion can be expressed by

$$\sigma_{w,m}^2 = \frac{\alpha_m^2 \Delta_m^2}{12} = \frac{10^{T_{min,m}/10}}{3}$$

- **Scale factor determination**
  Given a model or estimation of the subband noise variance $\sigma_{v,m}^2$, the scale factor α_m is given by

$$\alpha_m = \sqrt{\frac{\sigma_{w,m}^2}{\sigma_{w,m}^2 + 2.71\sigma_{v,m}^2}}$$

- **Quantization-step determination**
  The subband quantization-step value is given by

$$\Delta_m = \frac{2}{\alpha_m} 10^{T_{min,m}/20}$$

- To improve the robustness and computational complexity, Δ_m is quantized, in the log domain, to one of $\{\Delta^0, \Delta^1, \ldots, \Delta^{J-1}\}$

- **Discrete Cosine Transform**

  The masking threshold function should be transformed to the DCT domain

- **Discrete Fourier Transform**

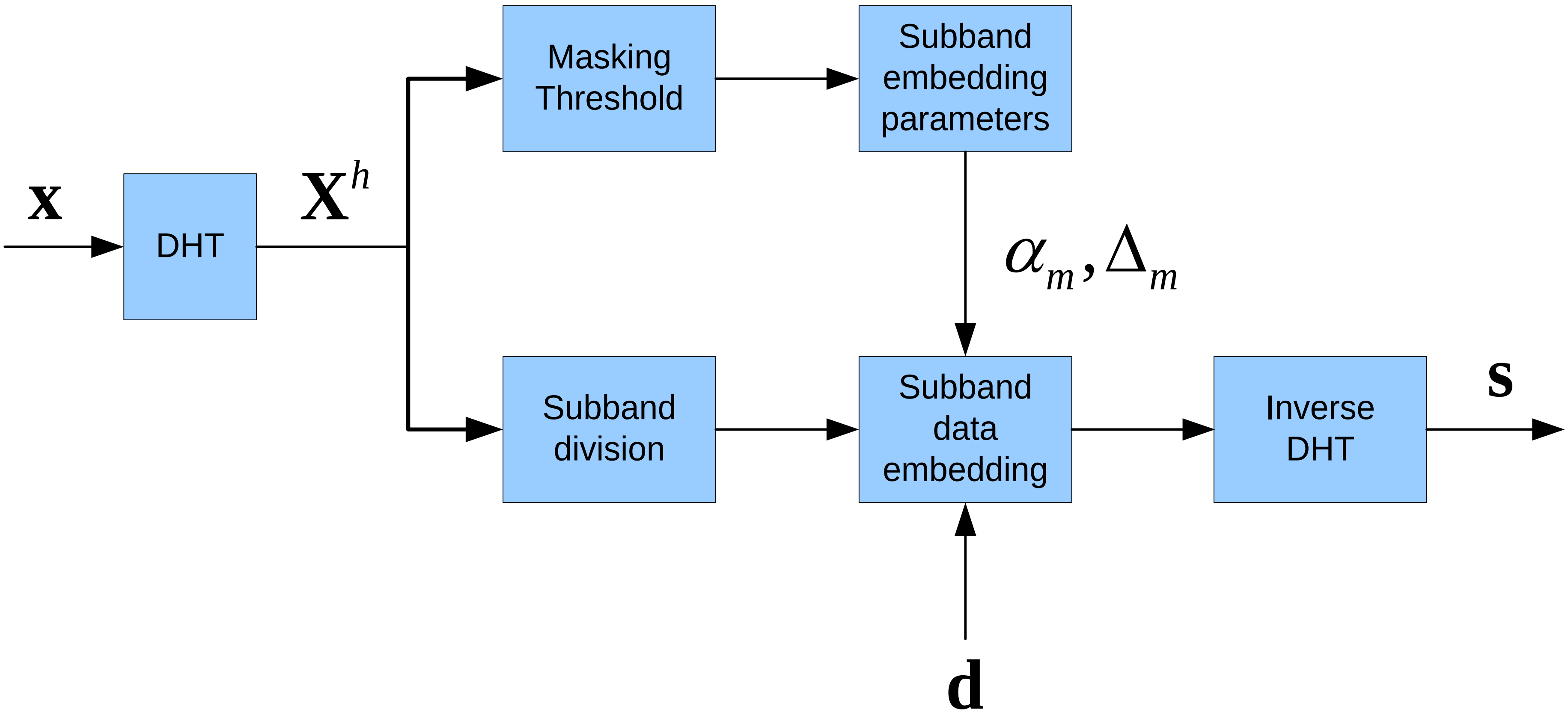
  The DFT is a complex valued transform

- **Discrete Hartley Transform**

$$X_k^h = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x_n \mathrm{cas}\left(\frac{2\pi}{N} nk\right); \quad k = 0, 1, \ldots, N-1$$

where $\mathrm{cas}(x) \triangleq \cos(x) + \sin(x)$

- Block diagram



DEMONSTRATION (1/2)

- Data-embedding in speech
  - Embedding only in frames detected by a voice activity detector
  - 2 subbands per frame of 256 samples (32ms), 32 bits per subband

- Transparency
  - Evaluated by PESQ, MOS scale [0-4.5]

- Averaged results (TIMIT 520 sentences, 22 minutes of speech)
  - MOS=3.9 WNR=18.3dB

- Female Speaker

  Original narrowband speech   Speech with embedded-data

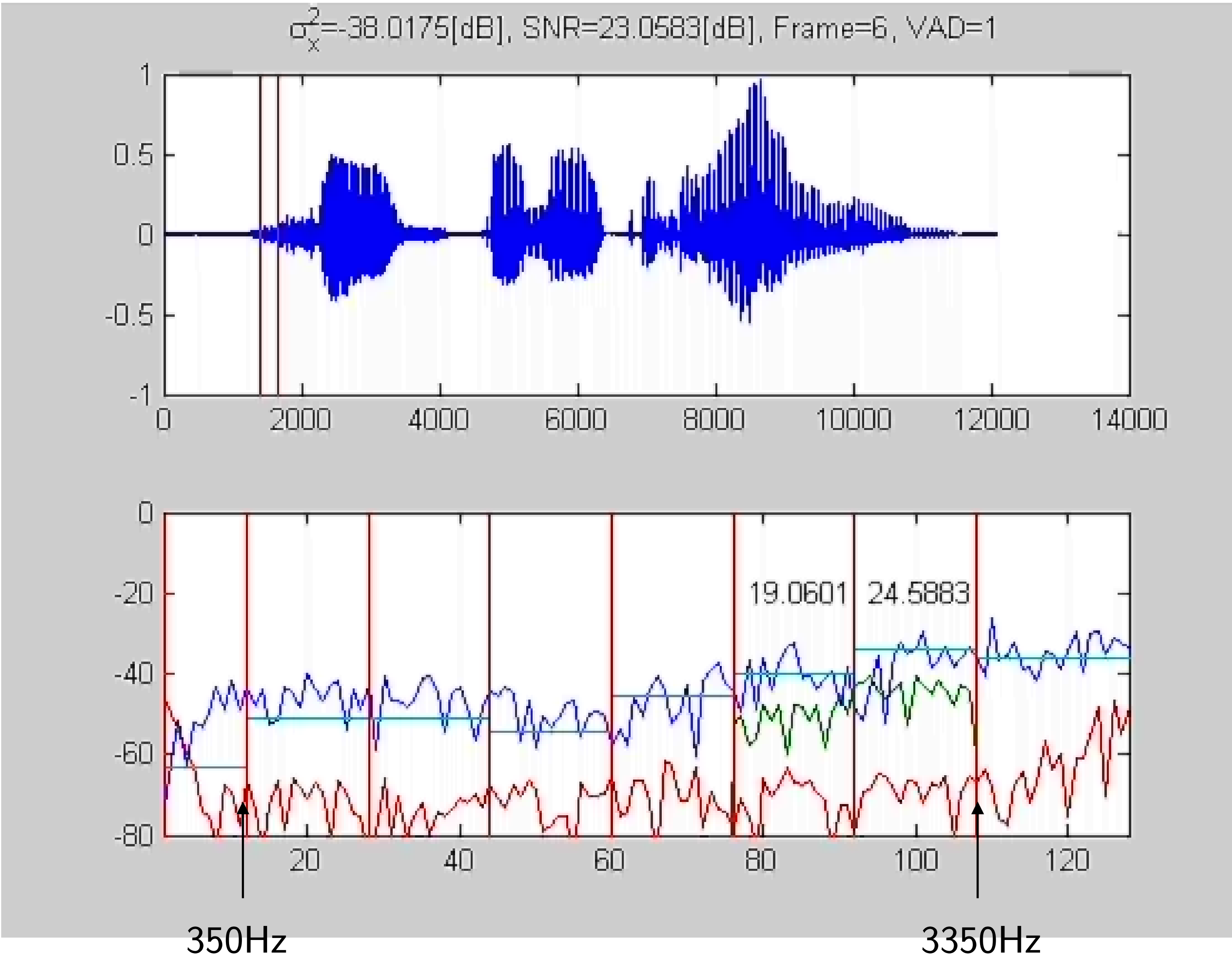
  MOS=4   WNR=20.1dB (STD=4.4dB)   #Frames=80

- Male Speaker

  Original narrowband speech   Speech with embedded-data

  MOS=3.7   WNR=19dB (STD=4.2dB)   #Frames=80

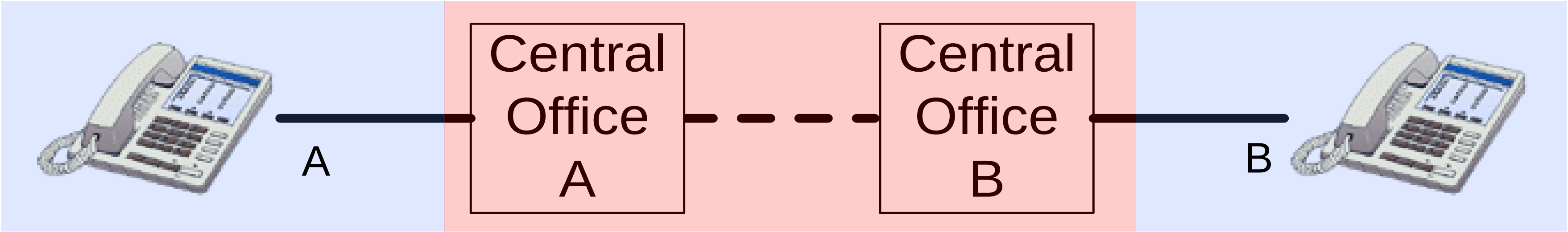
DEMONSTRATION (2/2)



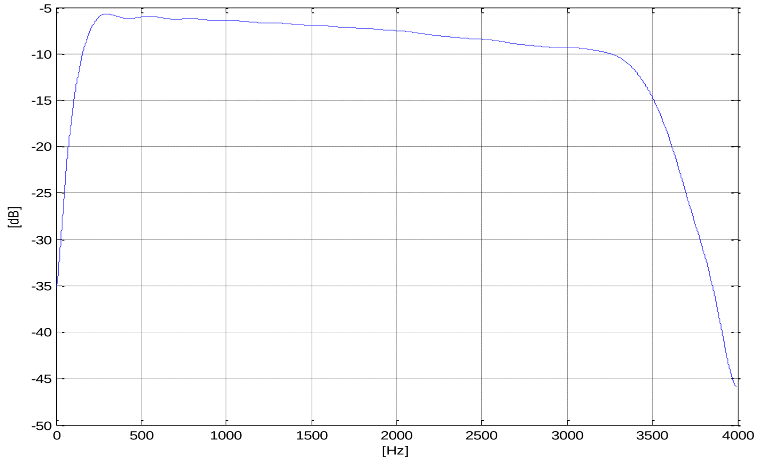
The decoder comprises of:

- **Adaptive equalizer** which reduces the channel spectral distortion
- Joint subband **embedded-data presence detection** and **quantization-step determination**
- **Embedded-data decoding**

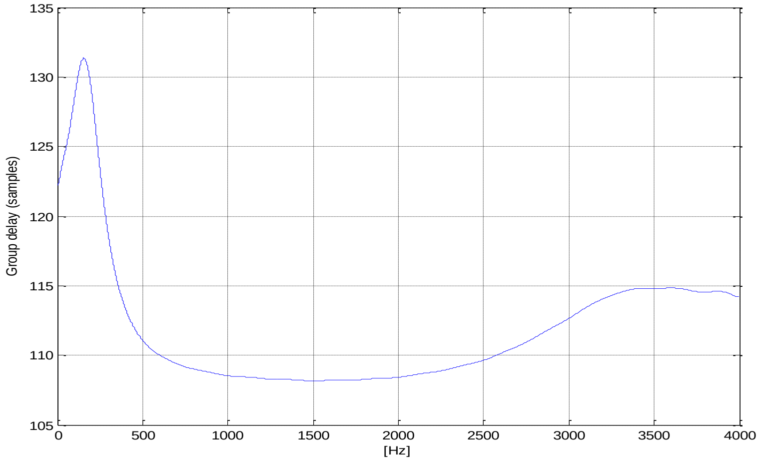
- The AWGN source is replaced with a simulation model of telephone channel
  - Amplitude and phase distortion, u-law or A-law quantization noise, Circuit (white Gaussian) noise



Point A to point B transfer function



Amplitude response



Group delay

- Common adaptive equalization algorithms
  - Time domain: NLMS, RLS
  - Frequency domain
- There is need for a training sequence for the above algorithms. In case of a telephone conversation, listening to the training sequence can be annoying
  - Solution: Select a chosen **audio/speech signal** as a **training sequence**
- Blind equalization algorithms
  - Pros: A training sequence is not needed
  - Cons: Not practical in our scenario, where data is embedded in a much stronger host signal.

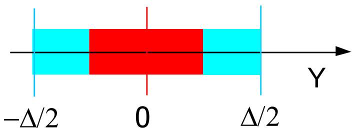
QUANTIZATION-STEP DETERMINATION (1/2)

- The estimated quantization-step will be one of $\{\Delta^0, \Delta^1, \ldots, \Delta^{J-1}\}$

- For each subband the decoder decides on the tested quantization-step values, and defines $\mathbb{G}$ as their indexes

- For each quantization-step index of $\mathbb{G}$, the decoder calculates the subband demodulated DHT coefficients

$$\boxed{Y^g_{m,k} = Q_{\Delta^g}\{R_{m,k}\} - R_{m,k}; \quad g \in \mathbb{G}}$$



where m is the subband index and k is the coefficient index

- Define two possible hypotheses
  - H_0: correct quantization-step, with PDF $p(Y|H_0)$
  - H_1: incorrect quantization-step, with PDF $p(Y|H_1)$

- The log-likelihood ratio (LLR), for each quantization-step index of $\mathbb{G}$

$$L_m^g = \log \left\{ \frac{p(\mathbf{Y}_m^g | H_0)}{p(\mathbf{Y}_m^g | H_1)} \right\}; \quad g \in \mathbb{G}$$

- The quantization-step index that maximize the LLRs, L_m^g

$$g^* = \arg\max_{g \in \mathbb{G}} L_m^g$$

- The estimated quantization step in the m'th subband is the quantization-step value that maximize the LLR

$$\boxed{\hat{\Delta}_m = \Delta^{g^*}}$$

- The maximal LLR, denoted by $L_m^{g^*}$, is used in the subband embedded-data presence detection rule

$$\boxed{\mathbb{I}_m = \begin{cases} 1; & L_m^{g^*} > T \\ 0; & L_m^{g^*} \leq T \end{cases}}$$

SUBBAND EMBEDDED-DATA STRUCTURE

- 256 coefficients/frame, 8 subbands/frame, 32 coefficients/subband

Information (12 bits)	Error correction code (11 bits)	Parameter protection code (9 bits)

- **Error correction code**
  - Golay code (23,12)

- **Parameter protection code**

  Improve robustness by using part of the subband coefficients for embedding a known sequence, denoted $\mathbf{u}$

  The hamming distance, d_u, between the hard decoded sequence, $\hat{\mathbf{u}}$, and the original sequence, $\mathbf{u}$, is calculated.

  The LLR computed from Y values and the LLR from the parameter protection code can be combined together for the quantization-step determination.

Simulation setup

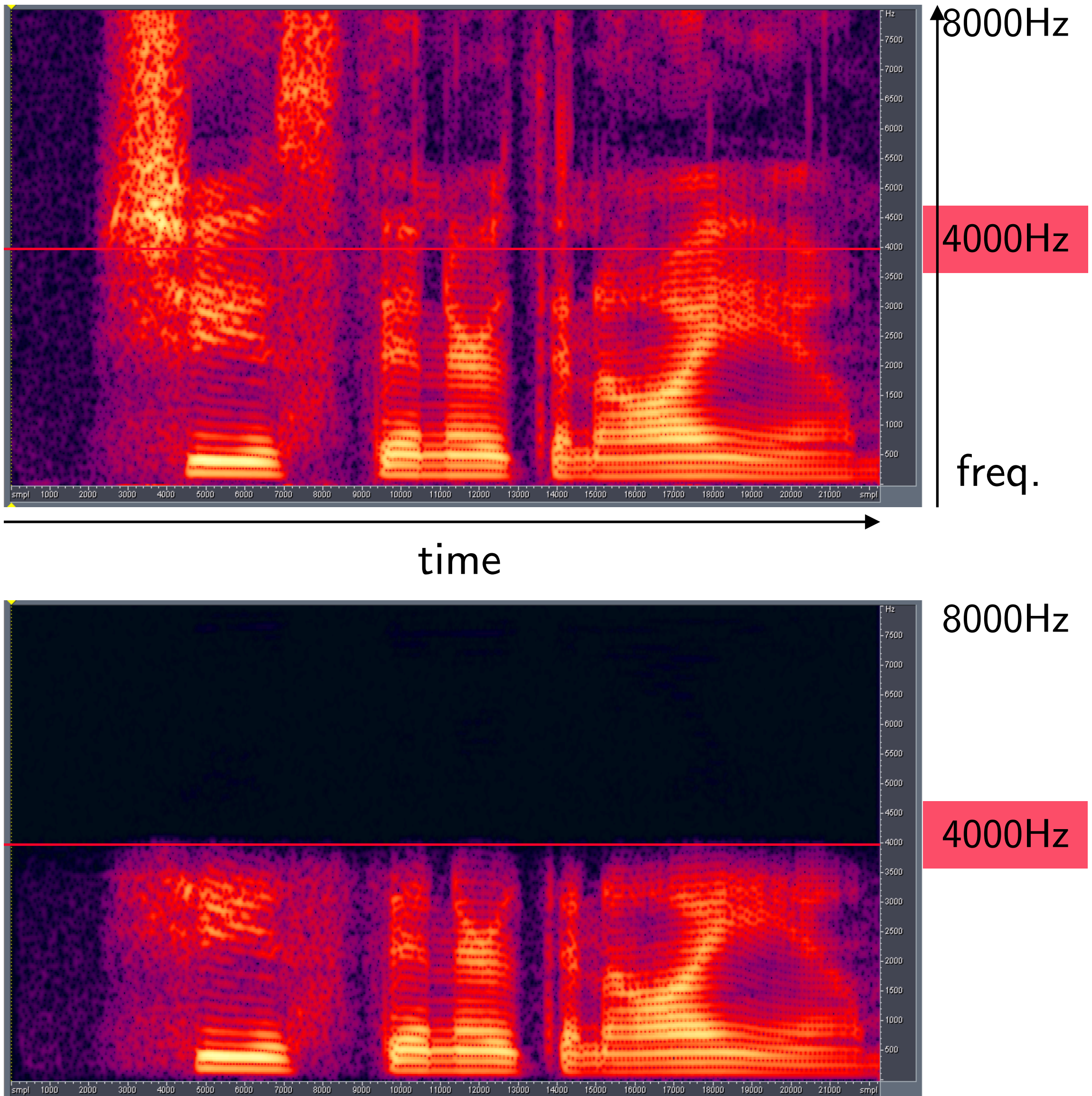
- Telephone channel model

The proposed data-embedding system performance is evaluated by the following objectives:

- Transparency MOS=3.9
- Embedding-rate RATE=(8000/256)*24*0.8=600[bits/sec.]
- Robustness BER(coded) $= \sim 3 \cdot 10^{-6}$
  BER(uncoded) $= \sim 3 \cdot 10^{-4}$

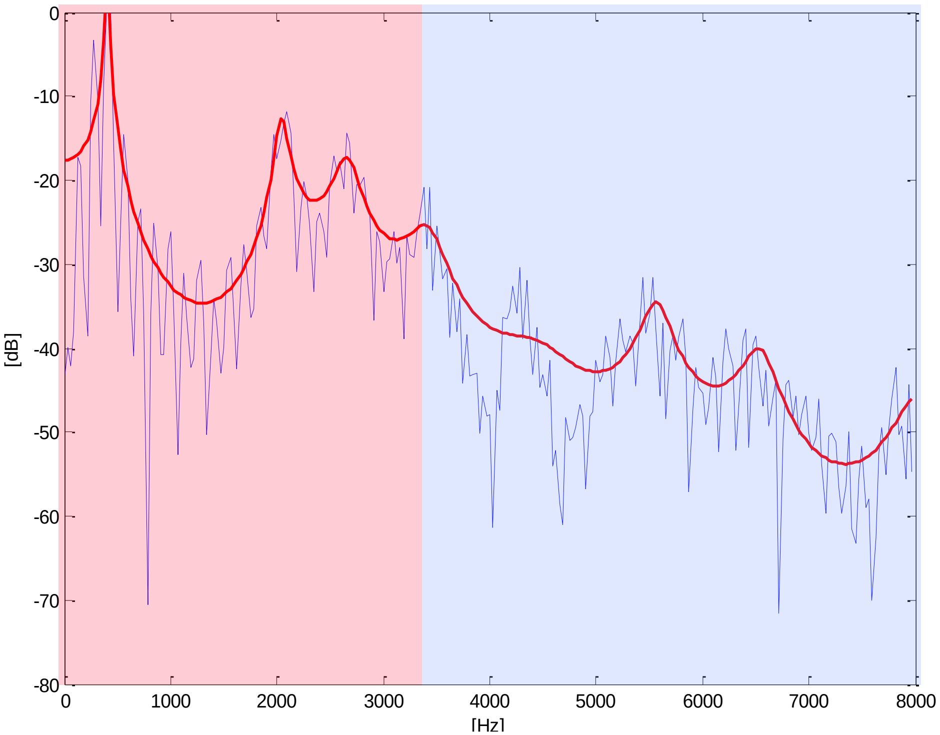
Wideband speech
bandwidth: 50-7000 Hz

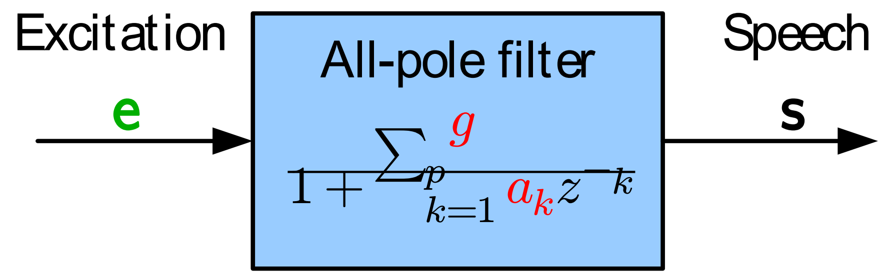
Telephone speech
bandwidth: 300-3400 Hz



speech spectrograms

HIGH-FREQUENCY SYNTHESIS

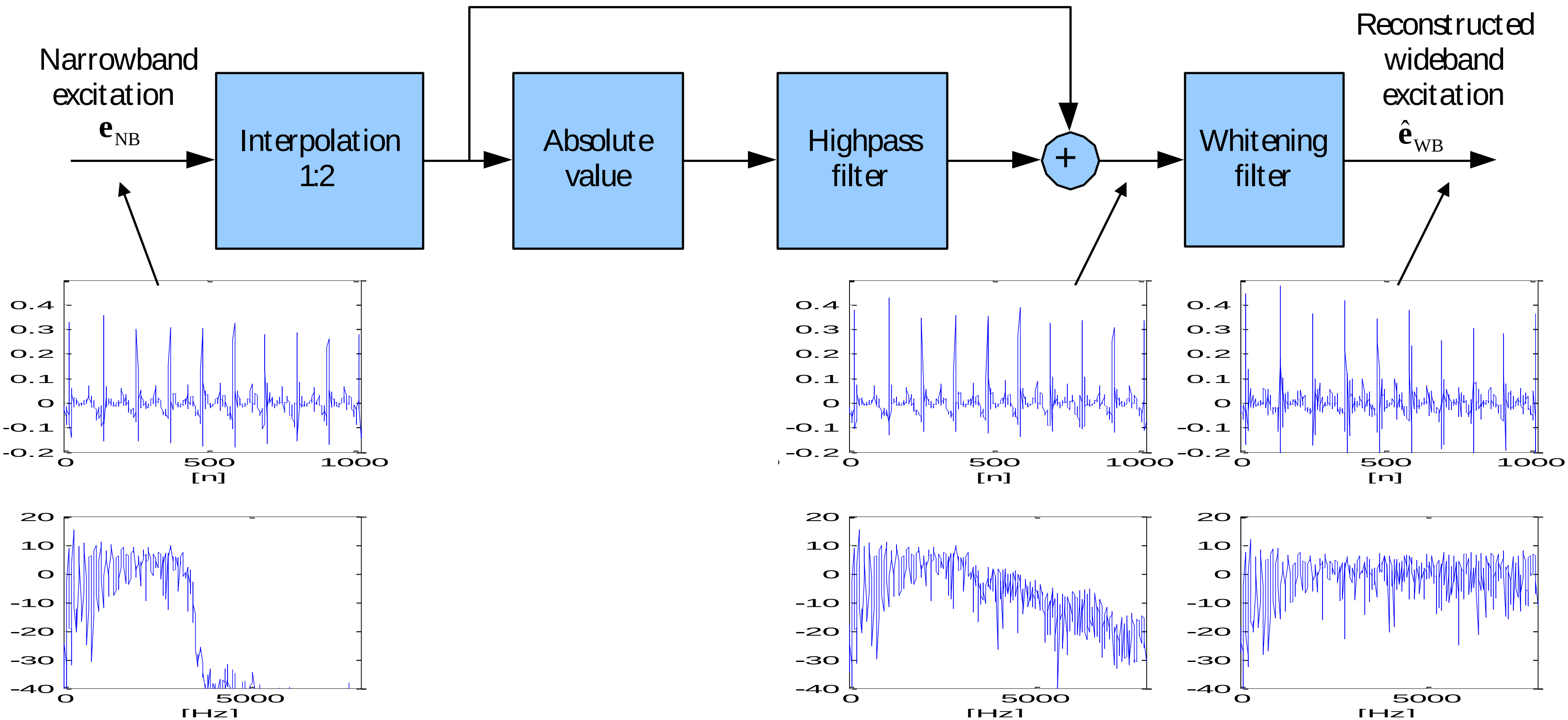




- Extract information about the high-frequency band from the narrowband speech
  - Extracted information: **High-frequency excitation**
- Use coding for the information that cannot be extracted
  - **Side information**: High-frequency **spectral envelope**, High-frequency **gain**

HIGH-FREQUENCY EXCITATION GENERATION

- A **non-linear operation**, the **absolute value**, expands the narrowband excitation bandwidth
- The **whitening filter** flattens the high-frequency **tilt** of the reconstructed wideband excitation



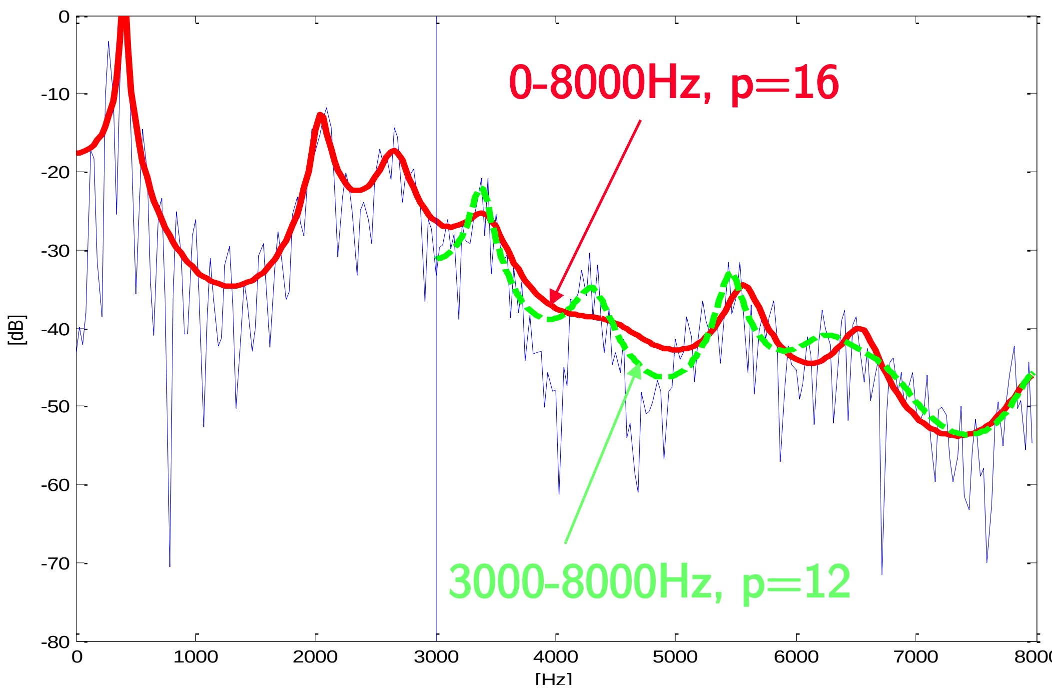
narrowband excitation

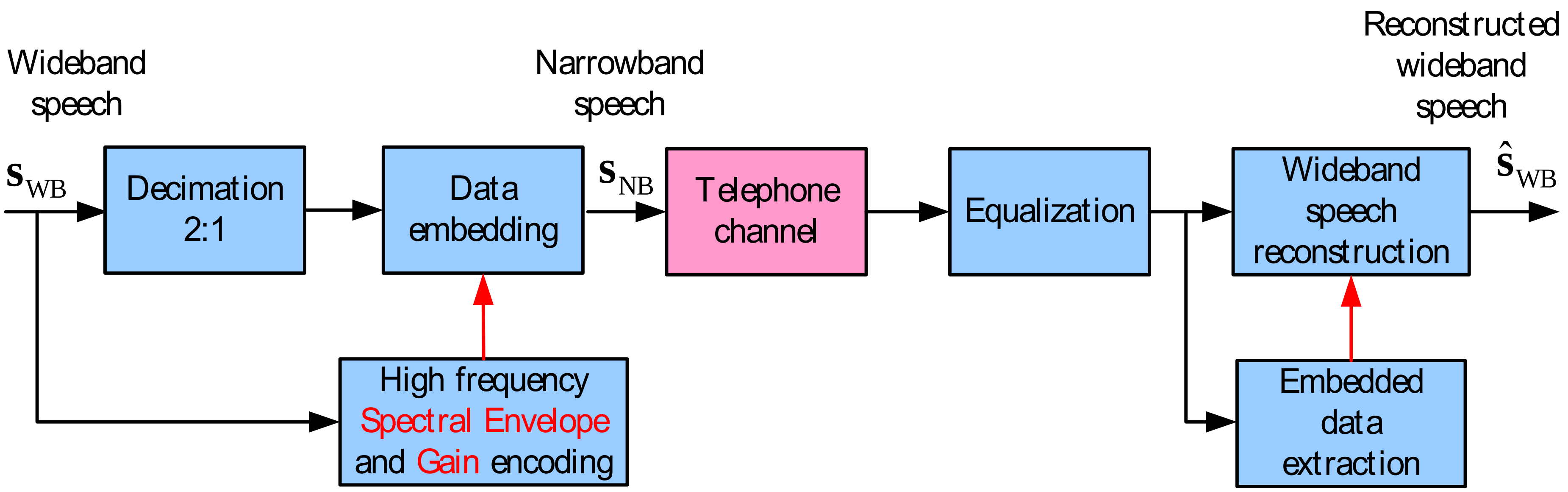
reconstructed wideband excitation

- By **spectral linear prediction** the spectrum $P(\omega)$ is modeled by an all pole spectrum $\hat{P}(\omega)$

$$\hat{P}(\omega) = \frac{G^2}{|1 + \sum_{k=1}^{p} a_k e^{-jk\omega}|^2}$$

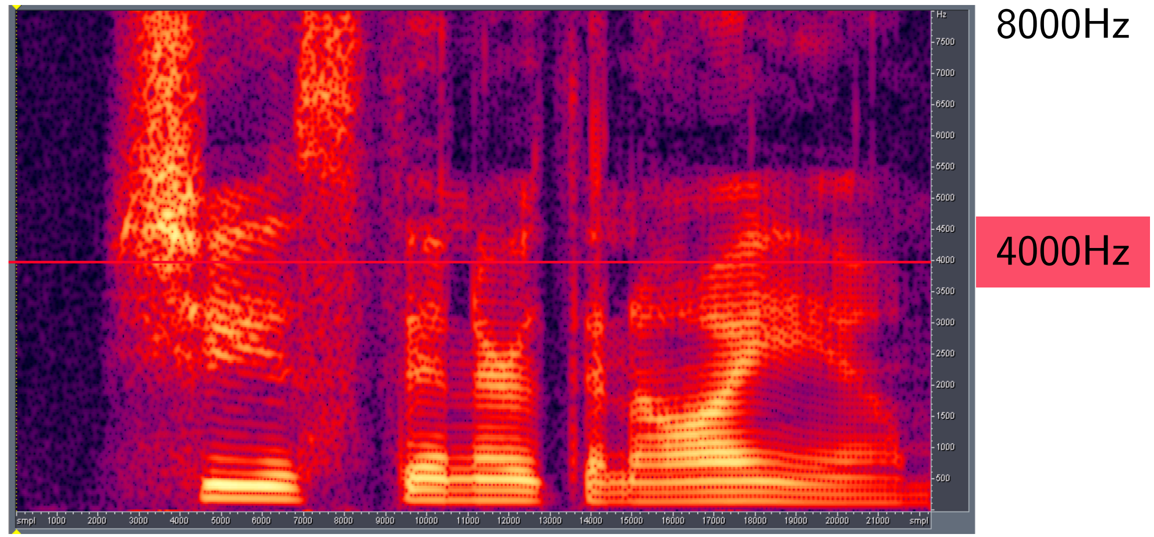
- By **selective spectral linear prediction** a specified frequency range $\omega_0 \leq \omega \leq \omega_1$ of the spectrum $P(\omega)$ is mapped to the frequency range $0 \leq \omega \leq \pi$, and the modified spectrum is analyzed with **spectral linear prediction**.





- The side information is embedded within the speech signal
  - **High frequency envelope**
    - The LSFs are coded using a 8-bit vector quantizer
  - **High frequency gain**
    - The gain, in the log domain, is coded using a 4-bit non-uniform scalar quantizer
  - A total of 12 bits per frame of 16msec

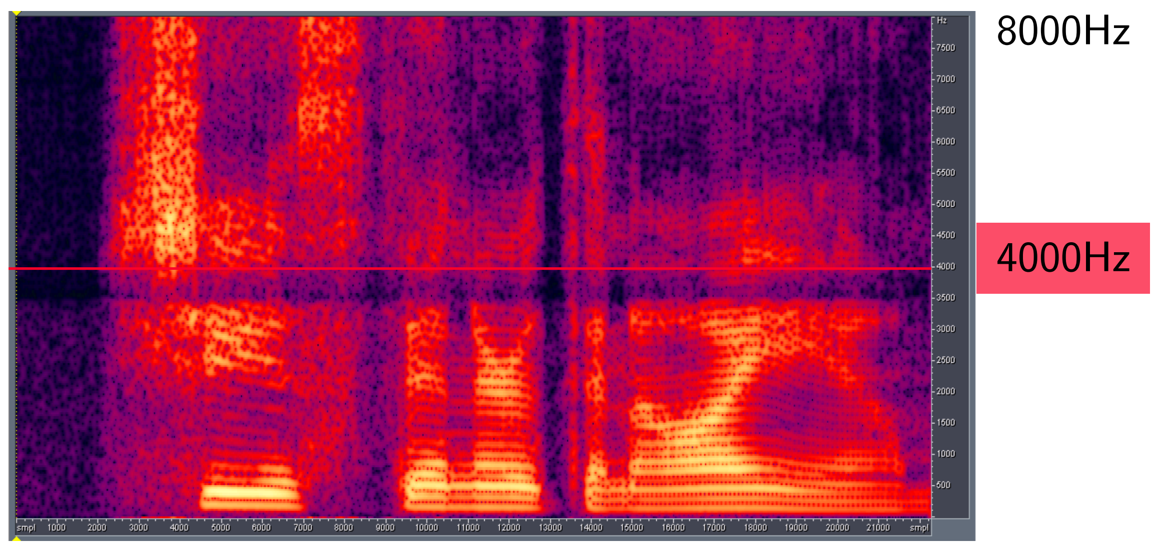
- Original wideband speech
- Telephone speech



- Reconstructed wideband speech

Additional example: Male speaker

- Original wideband speech
- Telephone speech
- Reconstructed wideband speech



- We showed how to combine **informed embedding** principles with a **perceptual model** for speech and audio signal.
- We developed methods for parameter estimation, and tested the methods under degradations caused by a **telephone channel**.
- We demonstrated a possible use of the embedded-data for **speech bandwidth extension**
- **Future research**
  - Increasing the rate of embedded-data
    - Increasing the number of subbands
    - D-ary SCS (D>2)
  - Data-embedding in Audio
    - Embedding-data in an Audio CD
  - Telephone speech recognition systems