POLITECNICO DI MILANO

MMSP 2nd Module – Lab4

Nicolò Bonettini
nicolo.bonettini@polimi.it

Transform coding

EXERCISE 1

Exercise 1 - Problem

1. Load the first 4s of the file 'gb.wav' and quantize it with PCM and R=8 bit. Compute the MSE and perceptually evaluate the result.

2. Consider groups of 8 symbols and quantize them using an optimal allocation of the 8 bits

3. Consider DCT transformation and repeat step 2 over transformed coefficients. Find the distortion and evaluate the perceived quality.

4. Consider a Karhunen-Loeve transformation and repeat step 2 over transformed coefficients. Find the distortion and evaluate the perceived quality.

Exercise 1 - Hints

1. In transform coding, each “coefficient” is quantized separately.

2. Optimal bit allocation is given by:

$$R_k = R + \frac{1}{2} \log \frac{\sigma^2_{y_k}}{\left(\prod_{i=1}^{N} \sigma^2_{y_i} \right)^{1/N}}$$

3. DCT matrix is built according to:

$$t_{kl} = \begin{cases} \sqrt{\frac{1}{N}} \cos\left(\frac{\pi}{2N}(k-1)(2l-1) \right) & k = 1 \\ \sqrt{\frac{2}{N}} \cos\left(\frac{\pi}{2N}(k-1)(2l-1) \right) & k = 2, 3, \ldots, N \end{cases}$$

4. To compute KLT, remember autocorrelation definition:

$$R_x = E[\mathbf{x}\mathbf{x}^t]$$

5. To compute eigen-values/vectors, use **eig(R)**

Additional background

LPC-RELATED PARAMETERS ESTIMATION

Pitch prediction using auto-correlation

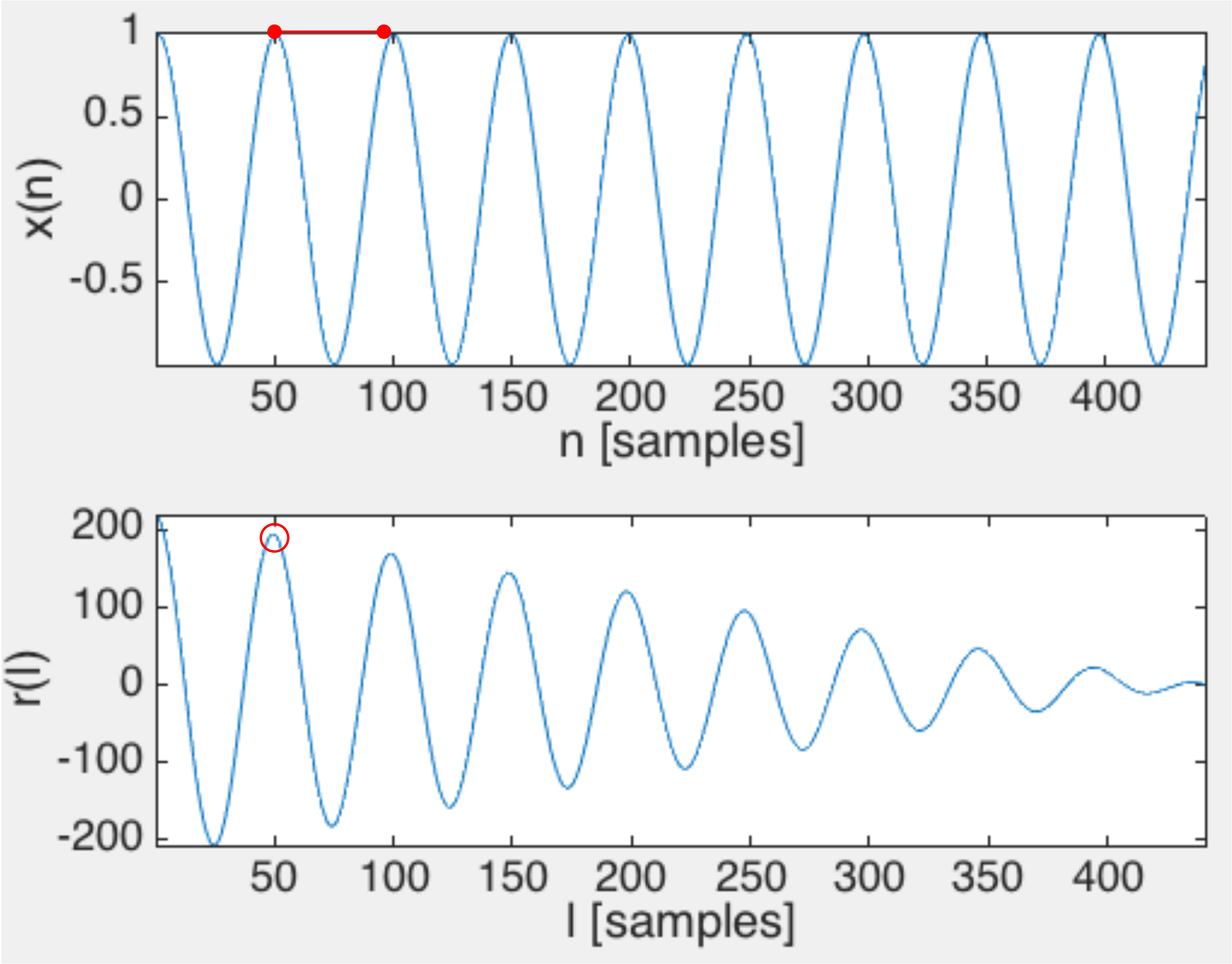
- The **autocorrelation function** (ACF) of a sequence x(n) is defined as:

$$r(l) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)x(n-l)$$

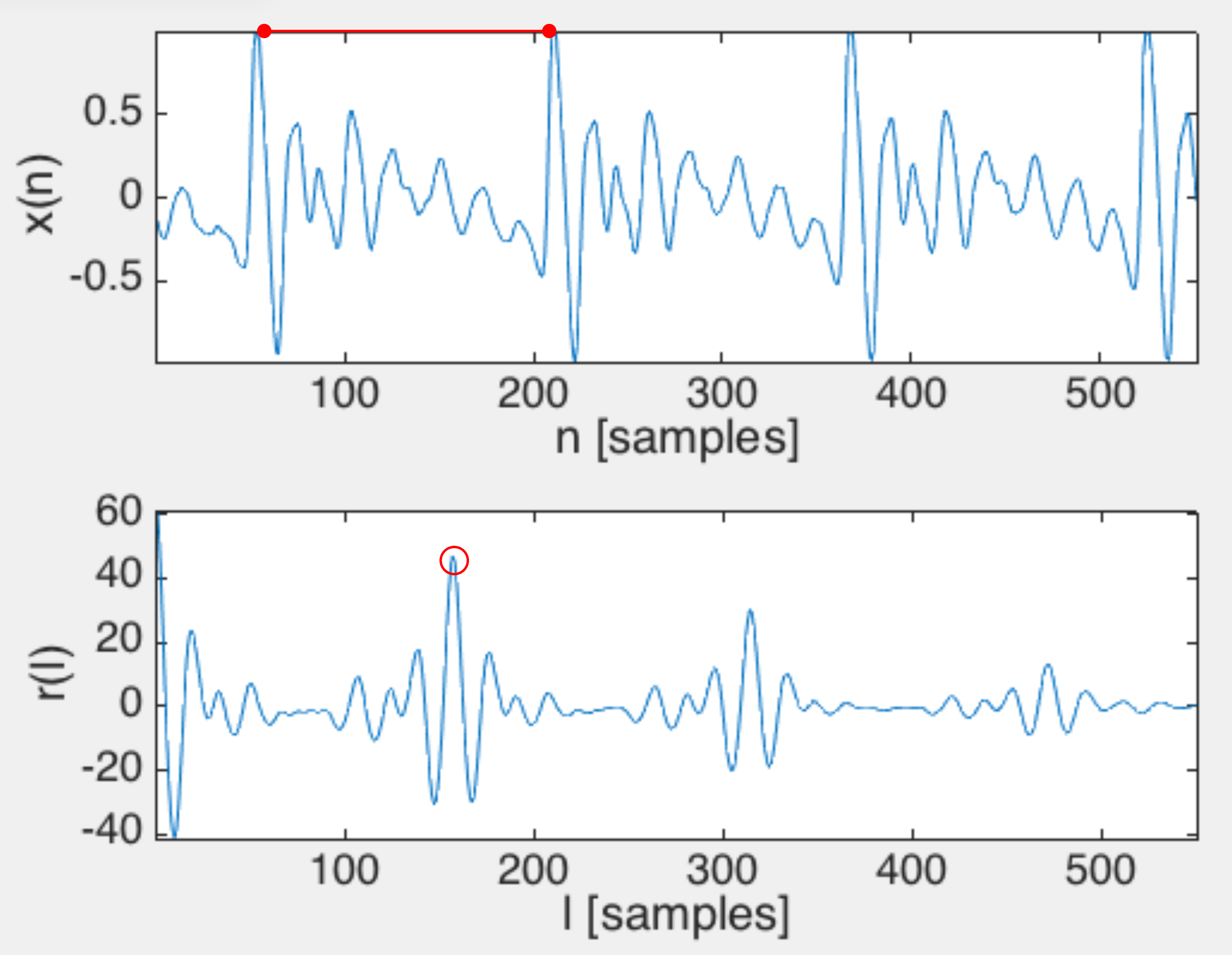
- For a **pure tone** with period L
  - ACF exhibits an ordered set of peaks at L, 2L, 3L, etc.

- For a pitched **real signal**:
  - the fundamental frequency component will behave like a pure tone, with the highest peak at lag L
  - other harmonics will produce one peak (not the highest) at lag L

Pitch prediction using auto-correlation: pure tone



Pitch prediction using auto-correlation: real signal



Voiced VS unvoiced

- **Zero-Crossing Rate (ZCR):**
  - ZCR is higher for unvoiced rather than voiced segments. The i-th segment is likely to be voiced if

$$\mathrm{zcr}_i < \tau_{\mathrm{zcr}}, \;\; \text{where } \tau_{\mathrm{zcr}} = \mathrm{median}(\mathrm{zcr}).$$

- **Short-time Energy (STE):**
  - STE is motivated by the fact that voiced segments have higher energy than unvoiced segments. The short-time energy is defined as the energy of the i-th frame, i.e.

$$\mathrm{ste}_i = \sum_{n=1}^{N} |s(n)|^2.$$

  - The i-th segment is likely to be voiced if

$$\mathrm{ste}_i > \tau_{\mathrm{ste}}, \;\; \text{where } \tau_{\mathrm{ste}} = \mathrm{median}(\mathrm{ste}).$$

Voiced vocoder

EXERCISE 2

Exercise 2 - Problem

1. Load the file 'voiced_a.wav' and consider only a 300ms frame. Plot the magnitude of the frequency response of the frame

2. Perform pitch detection using auto-correlation method. Consider only frequencies between 60 Hz and 500 Hz

3. Compute LPC coefficients of order 12

4. Plot the prediction error and its magnitude spectrum

5. Build an impulse train with the estimated pitch period

6. Consider the impulse train as excitation and build synthetic speech

7. Listen to the original and the synthetic speech

Exercise 2 - Hints

1. LPC coefficients alternative methods:
    1. Use the function lpc(). Notice that also the coefficient 1 of the filter is returned.
    2. Use autocorrelation function

$$\mathbf{a} = \mathbf{R}^{-1}\mathbf{r} \,,$$

$$\text{where} \quad \mathbf{R} = \begin{bmatrix} r(0) & r(1) & \cdots & r(p-1) \\ r(1) & r(0) & \cdots & r(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(p-1) & r(p-2) & \cdots & r(0) \end{bmatrix} \,,\ \mathbf{r} = \begin{bmatrix} r(1) \\ r(2) \\ \vdots \\ r(p) \end{bmatrix} \,,\ \mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}$$

Exercise 2 - Hints

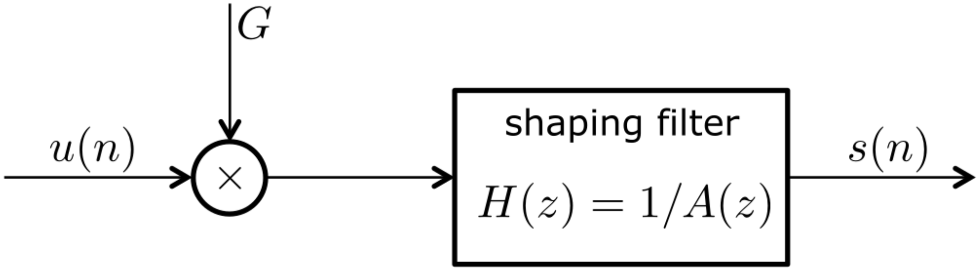
1. AR models for voiced signals

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n)$$

$$S(z) = \sum_{k=1}^{p} a_k z^{-k} S(z) + GU(z)$$

2. AR model TF

$$H(z) \triangleq \frac{S(z)}{GU(z)} = \frac{1}{1-\sum_{k=1}^{p} a_k z^{-k}} \triangleq \frac{1}{A(z)} \quad \text{with} \quad A(z) \triangleq 1 - \sum_{k=1}^{p} a_k z^{-k}$$



Voiced and unvoiced vocoder

EXERCISE 3

Exercise 3 - Problem

1. Load the files 'a.wav' and 'shh.wav' and build a single signal concatenating them

2. Implement a vocoder with the following characteristics
    1. Process frames of the signal windowed using Hamming windows of 40ms length and spaced of 10ms
    2. Voiced VS unvoiced detection
    3. LPC
        1. Voiced frames synthesized as in Ex.1
        2. Unvoiced frames synthesized using randn() as input signal