Exercises and Questions

October 26, 2018

1 Wiener and adaptive filtering

1. We want to implement an acoustic echo cancellation system based on the Wiener filter. The microphone signal is described as follows:

$$u(n) = 0.4v(n) + 0.15v(n-1) + b(n),$$

where $b(n)$ represents the speech signal and $v(n)$ is the echo component emitted by the loudspeaker. For simplicity, we assume that $b(n)$ and $v(n)$ are mutually uncorrelated within the observation window. The autocorrelation function of $v(n)$ has been estimated as $r(0) = 0.8$, $r(1) = 0.36$, $r(k) = 0$ for $k > 2$. The echo canceller has to estimate the speech signal by means of a Wiener filter with 2 samples.

Questions:

(a) Compute the autocorrelation matrix $\mathbf{R}$ of the input signal, and the cross-correlation vector $\mathbf{p}$ of the input signal and the desired response.

(b) Compute the set of the optimum (in the MSE sense) filter coefficients, namely $w(i)$ with $i = 1, 2$.

Solution:

(a) The input of the Wiener filter is the loudspeaker signal $v(n)$, while the microphone signal $u(n)$ represents the desired response.

Since the signals are real valued, the auto-correlation of the input signal $v(n)$ is a symmetric function, described by the following matrix:

$$\mathbf{R} = \begin{bmatrix} r(0) & r(1) \\ r(1) & r(0) \end{bmatrix} = \begin{bmatrix} 0.8 & 0.36 \\ 0.36 & 0.8 \end{bmatrix}$$

The cross-correlation between the input signal $v(n)$ and the desired response $u(n)$ is given by the function

$$p(k) = E[v(n)u(n-k)] = E[v(n)\{0.4v(n-k) + 0.15v(n-1-k) + b(n-k)\}]$$

Expanding the terms and considering the uncorrelatedness of $u(n)$ and $v(n)$, this function simplifies to

$$p(k) = 0.4E[v(n)v(n-k)] + 0.15E[v(n)v(n-1-k)] = 0.4r(k) + 0.15r(k+1)$$

The cross-correlation vector is computed for $k = 0$ and $k = -1$. Recalling that the autocorrelation function is symmetric (i.e., $r(k) = r(-k)$), it follows

$$\mathbf{p} = \begin{bmatrix} p(0) \\ p(-1) \end{bmatrix} = \begin{bmatrix} 0.4r(0) + 0.15r(1) \\ 0.4r(-1) + 0.15r(0) \end{bmatrix} = \begin{bmatrix} 0.4r(0) + 0.15r(1) \\ 0.4r(1) + 0.15r(0) \end{bmatrix} = \begin{bmatrix} 0.374 \\ 0.264 \end{bmatrix}$$

(b) Finally, the two coefficients of the filter are calculated by solving the Wiener-Hopf equations as follows:

$$\mathbf{w} = \mathbf{R}^{-1}\mathbf{p} = \frac{1}{0.8^2 - 0.36^2} \begin{bmatrix} 0.8 & -0.36 \\ -0.36 & 0.8 \end{bmatrix} \begin{bmatrix} 0.374 \\ 0.264 \end{bmatrix} = \begin{bmatrix} 0.4 \\ 0.15 \end{bmatrix}$$

Notice that the optimum filter $w(0) = 0.4$, $w(1) = 0.15$ exactly matches the coefficients of the impulse response linking the loudspeaker and the microphone. This means that the Wiener filter perfectly identifies the channel response. Thus the error signal is given by

$$e(n) = u(n) - v(n)^{\star}w(n) = u(n) - 0.4v(n) - 1.5v(n-1) = b(n)$$

i.e., $e(n)$ coincides with the speech signal $b(n)$.

2. Using the Linear Predictive Coding we want to estimate the coefficients of an auto regressive model of order 2:

$$s(n) = \sum_{i=1}^{2} a_i s(n-i) + u(n)$$

that approximates the audio signal $s(n)$. The estimated values of the auto-correlation function of the signal $s(n)$ are: $r(0) = 0.8$, $r(1) = 0.2$, $r(2) = 0.1$

Questions:

(a) Write the Wiener-Hopf equations relative to the analysed problem. The estimated values of the auto- correlation function of the signal $s(n)$ are: $r(0) = 0.8$, $r(1) = 0.2$, $r(2) = 0.1$

(b) Estimate the optimum (in the MSE sense) coefficient set a_i

Solution:

(a) Wiener-Hopf equations:

$$r(i) = \sum_{k=1}^{p} a_k r(i-k) = \sum_{k=1}^{2} a_k r(i-k) \quad i = 1, 2$$

Recalling that for real-valued signals $r(k) = r(-k)$, we have that

$$r(1) = \sum_{k=1}^{2} a_k r(1-k) = a_1 r(0) + a_2 r(1)$$

$$r(2) = \sum_{k=1}^{2} a_k r(2-k) = a_1 r(1) + a_2 r(0)$$

(b) The matrix form of the Wiener-Hopf equations is:

$$\begin{bmatrix} r(0) & r(1) \\ r(1) & r(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} r(-1) \\ r(-2) \end{bmatrix} = \begin{bmatrix} r(1) \\ r(2) \end{bmatrix}$$

And therefore we obtain the two coefficients:

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} r(0) & r(1) \\ r(1) & r(0) \end{bmatrix}^{-1} \begin{bmatrix} r(1) \\ r(2) \end{bmatrix} = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}^{-1} \begin{bmatrix} 0.2 \\ 0.1 \end{bmatrix} = \begin{bmatrix} 0.233 \\ 0.067 \end{bmatrix}$$

3. Consider a digital audio signal $x(n)$ sampled at 44100 Hz. The signal is the digitalized version of an analogical recording of a speech signal on a magnetic tape. In order to attenuate the additive noise of the recording, we adopt a frequency domain Wiener filtering. The Wiener filtering is based on the knowledge of the power spectral densities $V(f)$ of the noise and $S(f)$ of the signal. Assume that, in a specific time interval, the two quantities are estimated as follows

$$\hat{S}(f) = \begin{cases} af^2, & [W/\mathrm{Hz}] \quad 0 < f < 500\,\mathrm{Hz} \\ \frac{b}{f-400}, & [W/\mathrm{Hz}] \quad f \geq 500\,\mathrm{Hz} \end{cases}$$

$$\hat{V}(f) = 0.1, \quad [W/\mathrm{Hz}]$$

Questions:

(a) Determine the magnitude of the optimum filter (in the MSE sense) for this specific time interval.

(b) Assume now that the signal $x(n)$ has a duration of $5\,\text{s}$, and it is not stationary. In order to perform the filtering, the Overlap and Add technique is adopted. Assuming also that the signal can be considered as stationary on intervals of 20ms, calculate the maximum length (in samples) of the analysis window $w(t)$.

(c) We decide to employ the Hamming window. Which is the minimum overlap factor to honor the COLA (Constant Overlap and Add) condition?

(d) Using the results obtained in the previous steps, calculate the number of multiplications needed for processing a single frame. The filtering is performed in the frequency domain. For simplicity, neglect the number of multiplications needed for computing the Discrete Fourier Transform and its inverse.

(e) How many multiplications are needed for filtering the entire signal $x(n)$?

Solution:

(a)

$$H(f) = \frac{\hat{S}(f)}{\hat{V}(f) + \hat{S}(f)} = \begin{cases} \frac{af^2}{0.1+af^2}, & [W/\text{Hz}] \quad 0 < f < 500\,\text{Hz} \\ \frac{b}{0.1(f-400)+b}, & [W/\text{Hz}] \quad f \geq 500\,\text{Hz} \end{cases}$$

(b)

$$M \leq 0.02\,\text{s} \times 44100\,\text{Hz} = 882\,[\text{samples}]$$

(c) Minimum overlap for Hanning window: 50% ⇒ Hopsize: $H = 441\,[\text{samples}]$

(d) Each frame of the signal has a length of 882 samples. For an efficient calculation of the Discrete Fourier Trasform (DFT), the frame is zero-padded in order to reach a final length of 1024, corresponding to the power of 2 immediately next to 882. The filtering operation is therefore performed by multiplying the DFT of the signal with the DFT of the filter, for a total of 1024 multiplications.

(e) Being L the signal length in samples, the signal x(n) will be windowed with a total number

$$N = \lfloor \frac{L-M}{H} \rfloor + 1$$

frames. The signal has a length of 5 seconds, corresponding to $L = 220500\,[\text{samples}]$ at $44100\,\text{Hz}$. The total number of frames is therefore 499, and the total number of multiplications is:

$$N \times 1024 = 499 \times 1024 = 510976$$

4. Consider the following tele-conferencing system, where a monophonic signal $x(n)$ is reproduced by a pair of loudspeakers. The signal $y(n)$ acquired at the microphone contains: the clean signal $s(n)$ generated by a human speaker; a disturb signal (noise) $v(n)$; the signal coming from the loudspeakers (i.e., the echoed signal), after being filtered by the room impulse responses $H_1(z)$ and $H_2(z)$ (see Figure 1).

We want to design an echo-cancellation system, based on the Wiener filtering. In particular, we aim at attenuating the echo components captured by the microphone, in order to obtain (and transmit) an estimate of the clean signal $s(n)$. For the sake of simplicity, we assume that $x(n)$, $v(n)$, $s(n)$ are uncorrelated zero-mean white noises with variances

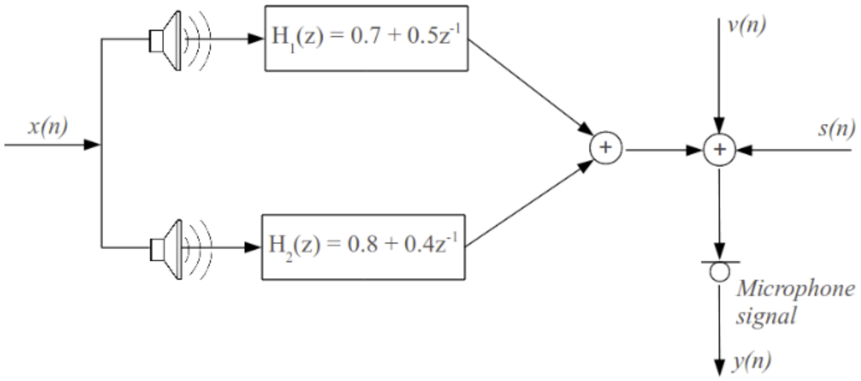

Figure 1: Tele-conferencing system

$\sigma_x^2 = 0.7$, $\sigma_v^2 = 0.01$ and $\sigma_s^2 = 0.8$, respectively. All the signals are real-valued. Moreover, the microphone transfer function is $H_{\text{MIC}(z)} = 1$.

Questions:

(a) Complete the above block-diagram including the echo-cancellation block, and discuss its working principle

(b) Compute the optimal (in the MSE-sense) Wiener filter with $M = 2$ taps

(c) Consider the implementation of the system using the iterative Steepest-Descent algorithm. Write the update equation for the filter coefficients. Which is the maximum value of the step-size μ such that the algorithm is guaranteed to be stable?

Solution:

(a) The input to the Wiener filter (with impulse response $w(n)$) is given by the signal on the loudspeakers $x(n)$, while the microphone signal $y(n)$ represents the desired response. If $q(n) = x(n)*w(n)$ is the output of the filter, the goal of the Wiener filter is to minimize the MSE, i.e. to minimize $J = E[e(n)e(n)]$, where $e(n) = y(n)–q(n)$. With reference to the orthogonality principle, minimizing J means guaranteeing that $e(n)$ will be uncorrelated with respect to the input signal $x(n)$. As a consequence, the signal $e(n) = y(n) - q(n)$ will contain an estimate of the clean signal $s(n)$, i.e. the portion of $y(n)$ that is not correlated with the signal $x(n)$ on the speakers. The complete block diagram is shown in Figure 2.

(b) We first compute the auto-correlation matrix of the input signal to the Wiener filter. As $x(n)$ is a white noise, this matrix will be diagonal, and its diagonal elements will correspond to the noise variance:

$$\mathbf{R} = \begin{bmatrix} r(0) & r(1) \\ r(1) & r(0) \end{bmatrix} = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_x^2 \end{bmatrix}$$

Let us compute the cross-correlation vector between the input $x(n)$ and the desired response $y(n)$. In particular, we need to compute

$$p(k) = E[x(n)y(n-k)] \quad k = 0,\, k = -1$$

In order to do so, from the block diagram we can derive:

$$\begin{aligned} y(n) &= 0.7x(n) + 0.5x(n-1) + 0.8x(n) + 0.4x(n-1) + v(n) + s(n) \\ &= 1.5x(n) + 0.9x(n-1) + v(n) + s(n) \end{aligned}$$

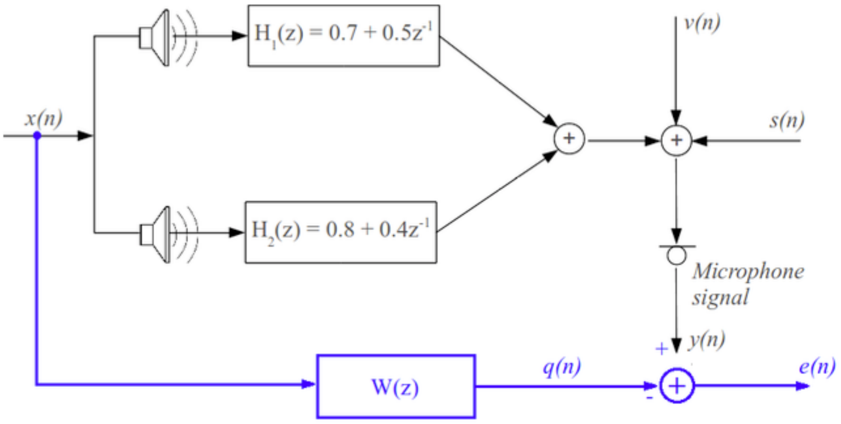

Figure 2: Tele-conferencing system complete block diagram

After replacing this expression in that of the cross-correlation, we obtain

$p(k) = E[x(n)y(n-k)] = E[x(n)\{1.5x(n-k)+0.9x(n-k-1)+v(n-k)+s(n-k)\}]$

After some simple math, considering that $x(n)$, $v(n)$ and $s(n)$ are mutually uncorrelated, we obtain

$$p(k) = 1.5E[x(n)x(n-k)] + 0.9E[x(n)x(n-k-1)] = 1.5r(k) + 0.9r(k+1)$$

After computing the values of $p(k)$ for $k = 0$ and $k = -1$, the cross-correlation vector becomes

$$\mathbf{p} = \begin{bmatrix} p(0) \\ p(1) \end{bmatrix} = \begin{bmatrix} 1.5r(0) + 0.9r(1) \\ 1.5r(1) + 0.9r(0) \end{bmatrix} = \begin{bmatrix} 1.5\sigma_x^2 \\ 0.9\sigma_x^2 \end{bmatrix}$$

We can finally compute the optimal two-taps filter (M=2), i.e. the samples $w(0)$ and $w(1)$, by solving the Wiener-Hopf equation

$$\mathbf{w}_0 = \begin{bmatrix} w(0) \\ w(1) \end{bmatrix} = \mathbf{R}^{-1}\mathbf{p} = \begin{bmatrix} 1/\sigma_x^2 & 0 \\ 0 & 1/\sigma_x^2 \end{bmatrix} \begin{bmatrix} 1.5\sigma_x^2 \\ 0.9\sigma_x^2 \end{bmatrix} = \begin{bmatrix} 1.5 \\ 0.9 \end{bmatrix}$$

(c) The update equation of the coefficient, with reference to the Steepest-Descent algorithm, is as follows:

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu[\mathbf{p} - \mathbf{R}\mathbf{w}(n)],$$

where the vector $\mathbf{w}(n)$ contains the two filter coefficient at the iteration n. An alternate and equivalent expression of this equation is as follows:

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu E[\mathbf{x}(n)e(n)],$$

where $\mathbf{x}(n) = [x(n), x(n-1)]^T$.

Finally, the step-size μ must satisfy the following relationship in order to guarantee stability:

$$0 < \mu < \frac{2}{\lambda_{\max}},$$

where $\lambda_{\max}$ is the maximum eigenvalue of the autocorrelation matrix $\mathbf{R}$. As $\mathbf{R}$ is diagonal, the eigenvalues can be readily found on its diagonal, therefore $\lambda_{\max} = \sigma_x^2$. This means that:

$$0 < \mu < \frac{2}{\sigma_x^2} \Rightarrow 0 < \mu < 2.85$$

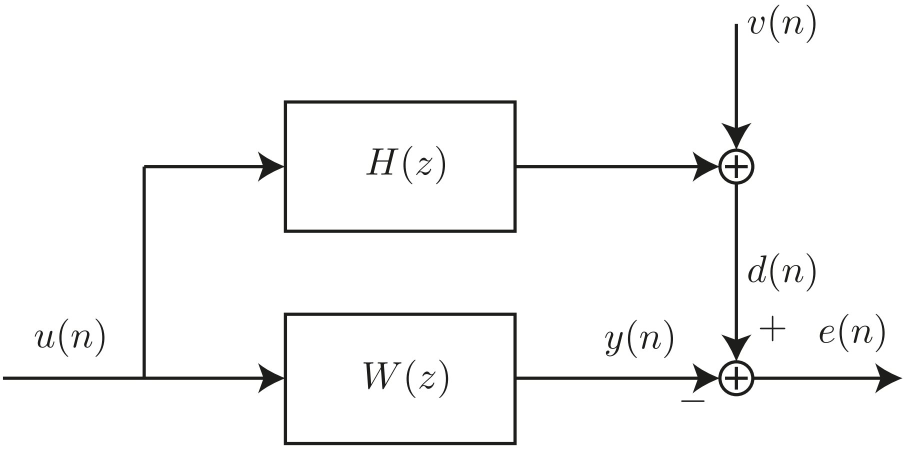

Figure 3: Block diagram

5. Consider the system in Figure 3, we know that $H(z) = 2 + 3z^{-1}$, $u \sim N(0, \sigma_u^2 = 1)$, $v \sim N(0, \sigma_v^2 = 0.1)$, $u \perp tv$ and $u, v \in \mathbb{R}$.

Questions:

(a) Calculate the optimum filter $W_0(z)$ with $M = 2$ taps.

Solution:

(a) First of all lets calculate the autocorrelation matrix, that will be symmetric since the signals are real

$$\mathbf{R} = \begin{bmatrix} r(0) & r(1) \\ r(1) & r(0) \end{bmatrix}$$

with

$$r(0) = E[u(n)u(n)] = E[u(n)^2] = \sigma_u^2 = 1$$

$$r(1) = E[u(n)u(n-1)] = 0 \quad \text{(remember the } u \text{ is white noise)}$$

Hence,

$$\mathbf{R} = \begin{bmatrix} r(0) & r(1) \\ r(1) & r(0) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Lets now calculate the cross-correlation vector

$$\mathbf{p} = \begin{bmatrix} p(0) \\ p(-1) \end{bmatrix}$$

with

$$\begin{aligned} p(0) &= E[u(n)d(n)] = E[u(n)\{2u(n) + 3u(n-1) + v(n)\}] \\ &= E[2u(n)^2] + E[3u(n)u(n-1)] + E[u(n)v(n)] = 2\sigma_u^2 + 0 + 0 = 2 \end{aligned}$$

$$\begin{aligned} p(-1) &= E[u(n-1)d(n)] = E[u(n-1)\{2u(n) + 3u(n-1) + v(n)\}] \\ &= E[2u(n)u(n-1)] + E[3u(n-1)^2] + E[u(n-1)v(n)] = 0 + 3\sigma_u^2 + 0 = 3 \end{aligned}$$

Finally, we obtain the optimum filter coefficients as follows

$$\mathbf{w}_0 = \mathbf{R}^{-1}\mathbf{p} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

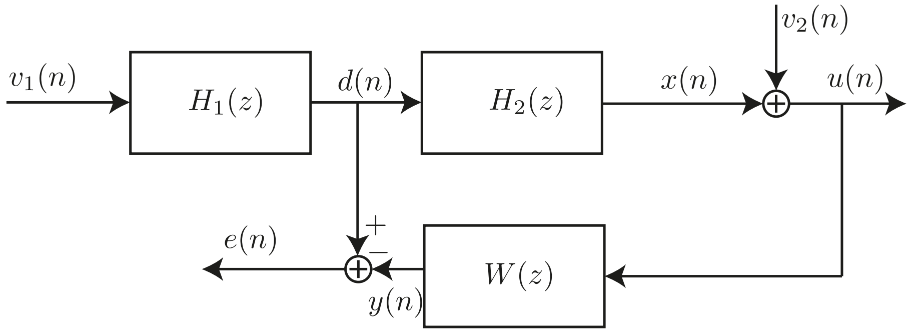

Figure 4: Block diagram

6. Consider the system in Figure 4, we know that $v_1 \sim N(0, \sigma_1^2 = 0.27)$, $v_2 \sim N(0, \sigma_2^2 = 0.1)$, $v_1(n), v_2(n) \in \mathbb{R}$ and $v_1 \perp v_2$, A consequence of the last hypothesis is that $x \perp v_2$. The two filters are given by

$$H_1(z) = \frac{1}{1 + a_1 z^{-1}}, \quad a_1 = 0.8458$$

$$H_2(z) = \frac{1}{1 + a_2 z^{-1}}, \quad a_2 = -0.9458$$

Questions:

(a) Calculate the optimum filter $W_0(z)$ with $M = 2$ taps.

(b) Calculate the minimum error that we achieve using $W_0(z)$

Solution:

(a) Lets try to solve the problem step by step

A) Calculate the autocorrelation of the input signal $u(n)$

$$r(0) = E[u(n)u(n)]$$

$$r(1) = E[u(n)u(n-1)]$$

First of all, since $v_2 \perp x$ we can write

$$r(k) = r_x(k) + r_{v_2}(k) \quad \forall k,$$

where $r_x(k)$ is the autocorrelation of x and $r_{v_2}(k)$ is the autocorrelation of v_2.

A.1) Calculate the autocorrelation of x

Looking at the portion of the complete block diagram shown in Figure 5, we can find the transfer function from v_1 to x as

$$H = \frac{1}{1 + a_1 z^{-1}} \frac{1}{1 + a_2 z^{-1}}$$
$$= \frac{1}{1 + (a_1 + a_2) z^{-1} + a_1 a_2 z^{-2}} = \frac{1}{1 + \alpha z^{-1} + \beta z^{-2}}$$

with $\alpha = (a_1 + a_2) = -0.1$ and $\beta = a_1 a_2 = -0.8$. From this expression we can easily see that x is an autoregressive process of order two. In order to calculate

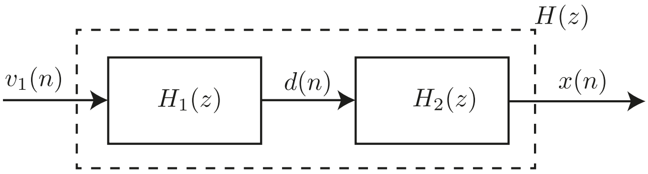

Figure 5: Portion of block diagram

the autocorrelation of x ($r_x(k)$), we have to solve the Yule-Walker equation (see Appendix. A)

$$\begin{cases} \sigma_1^2 = r_x(0) + \alpha r_x(1) + \beta r_x(2) \\ r_x(1) = -\alpha r_x(0) - \beta r_x(1) \\ r_x(2) = -\alpha r_x(1) - \beta r_x(2) \end{cases} .$$

Solving the above system we obtain

$$\begin{cases} r_x(0) = \frac{1+\beta}{(\beta-1)(\alpha+1+\beta)(\alpha-1-\beta)}\sigma_1^2 = 1 \\ r_x(1) = \frac{-\alpha}{(\beta-1)(\alpha+1+\beta)(\alpha-1-\beta)}\sigma_1^2 = 0.5 \\ r_x(2) = \frac{\alpha^2-\beta(1+\beta)}{(\beta-1)(\alpha+1+\beta)(\alpha-1-\beta)}\sigma_1^2 = 0.85 \end{cases}$$

As a consequence the autocorrelation matrix of x is

$$\mathbf{R}_x = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

A.2) Calculate the autocorrelation of v_2 Remember the v_2 is a white noise hence

$$\begin{cases} r_{v_2}(0) = \sigma_2^2 = 0.1 \\ r_{v_2}(1) = 0 \end{cases}$$

The autocorrelation matrix of v_2 is

$$\mathbf{R}_{v_2} = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}$$

Finally, the autocorrelation matrix of the input u is

$$\mathbf{R} = \mathbf{R}_x + \mathbf{R}_{v_2} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} + \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix} = \begin{bmatrix} 1.1 & 0.5 \\ 0.5 & 1.1 \end{bmatrix}$$

B) Calculate the cross-correlation between u and d

$$p(-k) = E[u(n-k)d(n)]$$

Looking at the block diagram in Figure 4 we can derive the following relations

$$\begin{cases} d(n) = x(n) + a_2 x(n-1) \\ u(n) = x(n) + v_2(n) \end{cases}$$

In particular, the first relation derives from the fact that $X(z) = H_2(z)D(z)$ and as a consequence $D(z) = 1/H_2(z)X(z)$. Using these relations we can calculate the cross-correlation function as follows

$$\begin{aligned} p(-k) =& E[\{x(n-k) + v_2(n-k)\}\{x(n) + a_2 x(n-1)\}] \\ =& E[x(n-k)x(n)] + a_2 E[x(n-k)x(n-1)] + \\ & E[v_2(n-k)x(n)] + a_2 E[v_2(n-k)x(n-1)] \\ =& r_x(k) + r_x(k-1) + 0 + 0. \end{aligned}$$

As a consequence we have

$$p(0) = r_x(0) + a_2 r_x(1) = 1 - 0.9458 \times 0.5 = 0.5272$$

$$p(-1) = r_x(1) + a_2 r_x(0) = 0.5 - 0.9458 \times 1 = -0.4458$$

Finally, we obtain the optimum filter coefficients as follows

$$\mathbf{w}_0 = \mathbf{R}^{-1}\mathbf{p} = \begin{bmatrix} 1.1 & 0.5 \\ 0.5 & 1.1 \end{bmatrix}^{-1} \begin{bmatrix} 0.5272 \\ -0.4458 \end{bmatrix} = \begin{bmatrix} 0.8360 \\ -0.7853 \end{bmatrix}$$

(b) The minimum error that we achieve is given by the value of the cost function J when the filter is optimum, i.e

$$J_{\min} = \sigma_d^2 - \mathbf{p}^H \mathbf{R}^{-1} \mathbf{p}.$$

In order to find the value σ_d^2 we can use the Yule-Walker equations (Appendix A). Recalling that the autocorrelation of d in 0 is equal to the variance ($r_d(0) = \sigma_d^2$) and that the autocorrelation is an even function, we have

$$\begin{cases} \sigma_1^2 = r_d(0) + a_1 r_d(-1) \\ r_d(1) = -a_1 r_d(0) \end{cases}$$

that gives

$$\sigma_d^2 = r_d(0) = \frac{\sigma_1^2}{1 - a_1^2} = \frac{0.27}{1 - 0.8458^2} = 0.9486.$$

Hence, the minimum error is equal to

$$J_{\min} = 0.1579$$

2 Microphone arrays

1. Consider a uniform linear microphone array, composed of M microphones. The spacing of the microphones is d.

   Questions:

   (a) Describe the concept of spatial filtering, and outline the working principle of the beamforming algorithm.

   (b) A single acoustic source, located in the far-field of the array, emits a narrow-band signal with center frequency $f_c = 2kHz$. Which is the maximum spacing of the microphones in order to avoid spatial aliasing? We recall that the speed of sound is approximately $c = 340m/s$.

   (c) Assume now the presence of a second acoustic source, which emits from the far-field another narrow-band signal, at the same center frequency $f_c = 2kHz$. The directions of arrival of the sources are supposed to be close to $\theta = 0°$ (i.e., the sources are almost frontal with respect to the array). Using the spacing d obtained before, how many microphones are required for an accurate localization of the two sources, knowing that their spatial separation is $\Delta_\theta = 15°$?

Solution:

(a) *Spatial filtering*: linear combination (with complex coefficients) of the microphone signals. The coefficients of the linear combination (=spatial filter) can be designed in such a way to emphasize the signal coming from some directions, and dim down those coming from other directions.

Beamforming: spatial filter design aimed at obtaining a filtered signal y_F such that:

- The global energy of y_F is minimized,
- The energy associated to sources in a given direction θ remains unaltered

(b) The maximum distance between sensors is given by $d_{\text{MAX}} = \lambda/2$, where $\lambda = c/f_c$. In our case $d_{\text{MAX}} = 8.5cm$.

(c) For sources that are directly in front of the array the beamforming algorithm is able to resolve the directions of arrival of two sources placed at an angular distance Δ_θ from each other, according to the formula:

$$\Delta_\theta > \frac{1}{L}$$

where L is the length of the array, expressed in number of wavelengths. We can therefore rewrite this condition as:

$$\Delta_\theta > \frac{\lambda}{(M-1)d}$$

which implies:

$$M > \frac{\lambda}{d\Delta_\theta} + 1$$

In order to achieve a resolution of $\Delta_\theta = 15° = 0.262rad$, using $d = d_{\text{MAX}} = 8.5cm$, we need at least $M = 9$ microphones.

2. We want to design an audio-localization system based on a uniform linear microphone array. More specifically, we want to localize a speech signal having components in the frequency range from 300 Hz to 3400 Hz, using the Beamforming algorithm. We recall that the speed of sound in air is $340m/s$.

Questions:

(a) Determine the maximum possible distance between the sensors in order to avoid spatial aliasing.

(b) Consider a scenario in which two people are speaking at the same time, and their angular separation is 10° in normal conditions (assume that their DOAs are around 0°). How many microphone are needed for localizing both the speakers? (For the calculation, use the maximum distance calculated in the previous step).

(c) The beamformer algorithm is not very efficient in terms of angular resolution. How can we increase the resolution?

Solution:

(a) The wavelength relative to the lower and upper bound of the frequency range are given by

$$\lambda_{\text{max}} = c/f_{\text{min}} = 340/300 = 1.133m$$

$$\lambda_{\text{min}} = c/f_{\text{max}} = 340/3400 = 0.1m$$

In order to avoid spatial aliasing, the distance between the sensors can not be higher than half of the wavelength. In our case, the maximum distance between the sensors is determined by the minimum wavelength, that corresponds to the highest frequency:

$$d_{\text{max}} = \lambda_{\text{min}}/2 = 0.05m$$

(b) Angular resolution of beamforming

$$\Delta_\theta \geq \frac{\lambda_{\text{max}}}{(M-1)d}$$

Minimum number of microphones:

$$M \geq \frac{\lambda_{\text{max}}}{\Delta_\theta d} + 1 \Rightarrow M \geq 131$$

(c) A first choice for increasing the angular resolution could be the CAPON beamforming. In our scenario, since we know the exact number of sources to be localized (two speakers), we can also adopt a parametric method (MUSIC or ESPRIT), which present higher performances. Notice that, however, in this case we need further assumptions:

- The source signals have to be independent
- The error on the microphones is a zero-mean white noise with the same variance for all the sensors.
- The number of microphones M is greater than the number of sources N.
- Sensor noise is uncorrelated to source signals
- All the DOAs are different, thus they lead to different spatial frequencies

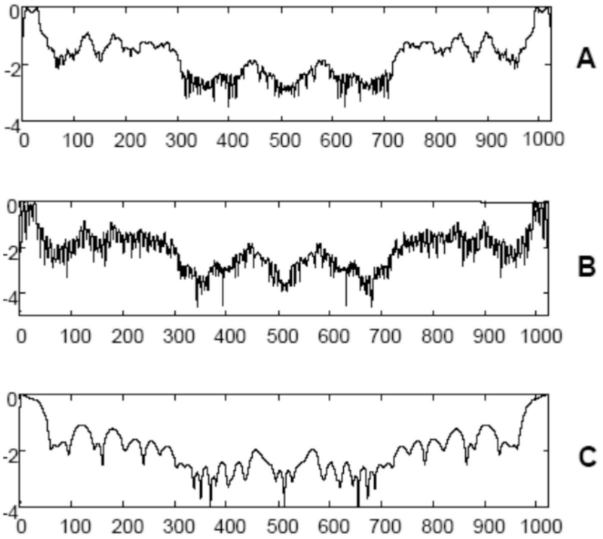

Figure 6: STFT

3 Sound analysis tools

1. Short-Time Fourier Transform (STFT) Analysis

   Questions:

   (a) Briefly describe how the window length and shape affects the STFT analysis. In particular, given a certain window length, describe the differences between a Hamming and a rectangular window

   (b) Consider a 1024-sample FFT computed on a speech segment. You can use Hamming windows of 64, 256, and 512 samples. Which one of these window lengths were used for which of the three cases (see Figure 6 A, B, and C)? Why?

   (c) Assuming that the sampling frequency is 16KHz, identify the highest-frequency formant and determine its approximate central frequency in Hz.

   Solution:

   (a) The main lobe of the Fourier transform of a Hamming window is twice as wide as that of a rectangular window, therefore its spectral resolution is half. The side lobes, on the other hand, are significantly smaller (23-24 dB vs. the 13 dB of the rectangular window). This makes the Hamming- based estimation accuracy of the frequency peaks much better than in the rectangular case.

   (b) Each frequenct peak, after windowing, will be shaped like the main lobe of the window. In the case of a 64-sample window, the main lobe will be $L \times 1024/64 = 64$ samples wide, L being the shape factor of the window ($L = 4$ for Hamming). Similarly 256-sample window will have a 16-sample wide main lobe, the 512-sample window will have a main lobe of 8 samples. Quite clearly, these windows correspond to C, A and B, respectively.

   (c) The highest formant frequency is centered around to the 440th sample (approximately), which corresponds to a frequency of $440 \times 16000/1024 = 6.9$ KHz

2. Let us consider a speech signal associated to a male voice, whose fundamental is never below 100 Hz, and whose bandwidth is limited to 5 KHz. We want to perform sinusoidal

analysis of this signal with the aim of deriving the control signals (magnitude and frequency) for an oscillator bank in additive synthesis (phase vocoder). In order to perform this task, we can use an A/D converter operating at 3 possible frequencies: 11 KHz, 22 KHz, 44 KHz. The available analysis tool is a Short-Time Fourier Transform (STFT) with 3 possible windows:

- Rectangular, 1024 samples, no overlap
- Bartlett (triangular), 2048 samples, 50% overlap
- Blackman, 2048 samples, 50% overlap

In all such cases the frame rate is the same (1 FFT computation every 1024 samples).

Questions:

(a) What spectral resolution (minimum resolvable spacing between spectral peaks) can be achieved with each combination of frequency and window?

(b) Which combinations of sampling frequency and window shape are suitable for resolving all the spectral peaks of the considered signal?

(c) If you want to minimize mutual interference between spectral peaks, which combination of window and sampling frequency it would be advisable to select, and why?

(d) What localization accuracy can be achieved when using a FFT of the same length of the window in the various considered cases? Which of the choices will cause problems and why?

(e) Briefly describe a suitable overall system for the re-synthesis of the considered harmonic signal.

Solution:

(a) The spectral resolution corresponds to about the width B_w of the window's main lobe, which can be computed as LF_s/M, L being the shape factor, F_s the sampling frequency, and M the window length.

		F_s1	F_s2	F_s3
L	M	11000	22000	44000
2	1024	21	43	86
4	2048	21	43	86
6	2048	32	64	129

(b) As B_w must be smaller than the minimum spacing between peaks, all but the Blackman window at 44KHz will be suitable for resolving the peaks.

(c) In order to minimize the interference between peaks we need to keep the side lobes to a minimum, which is possible using a Blackman Window (better if with a higher sampling frequency, e.g. 22KHz)

(d) Localization accuracy is given by half the size of the frequency bin of the FFT, which is F_s/M.

		F_s1	F_s2	F_s3
L	M	11000	22000	44000
2	1024	11	21	43
4	2048	5	11	21
6	2048	5	11	21

As the Just Noticeable Difference (JND) at 100Hz is 3Hz, only the two windows in yellow will generate localization errors that we cannot hear. Those in orange will need a 2:1 oversampling. The others will need a larger oversampling and/or quadratic interpolation.

(e) Describe the system for spectral continuation etc.

4 Theoretical questions

1. Define the concept of Time-Scaling for a periodic signal.

2. Describe the Generalized Cross Correlation method for the localization of wideband sources. In particular, explain why we should adopt a pre-whitening filter for the signals.

3. Describe the similarity between spatial filtering and temporal filtering. Based on this similarity, derive the condition on the minimum distance between adjacent microphones.

4. We need to restore an audio signal corrupted by clicks, stationary background noise and template pulse noises. Describe, with the help of a block diagram, the whole restoration processing, and make sure you specify the order in which the individual restoration operations are to be perform. Properly justify the answer.

5. Starting from the orthogonality principle, obtain the Wiener-Hopf equations. Which assumptions must be valid to write the Wiener-Hopf equations in a matrix form?

6. Provide a general definition of time scaling and pitch scaling based on sinusoidal modeling. Show how time scaling can be implemented using pitch scaling and vice-versa (duality principle).

7. Describe how disturbances on audio signals such as scratches or crackles are modeled for the purpose of restoration. In particular, define the switching process and how to estimate it using linear prediction coding.

8. Briefly describe the reverberation scheme based on image sources, while describing pros and cons. In particular, explain what particular aspect is mostly responsible for the computational cost.

9. Define the Energy Decay Curve (EDC) and the Energy Decay Relief (EDR), and emphasize the differences between them. Define the reverberation time T60 as a function of the EDC.

10. Define and describe the main reverberation parameters such as the EDC, the T60, the clarity index and the parameters that are a function of the statistical modeling of reverberation.

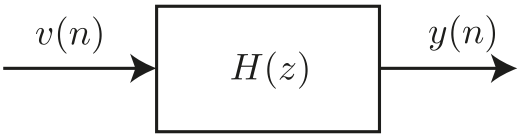

Figure 7: Autoregressive process

A Appendix: Yule-Walker equations

Yule-Walker equations for an autoregressive process of order K (see Figure 7) with

$$H = \frac{1}{1 + \sum_{k=1}^{K} a_k z^{-k}}$$

$$\begin{cases} \sigma_v^2 = r(0) + \sum_{k=1}^{K} a_k r(-k) \\ r(n) = -\sum_{k=1}^{K} a_k r(n-k), \end{cases}$$

where $r(n)$ is the autocorrelation function at time n.