Speech Enhancement

Single and Dual Channel Noise Reduction

- Noisy speech is everywhere. For example, in a bus/train.
- Objective : ① Remove the noise
  or ② Estimate the good speech from noisy speech.
- We consider additive model.

$$y(k) = x(k) + n(k)$$

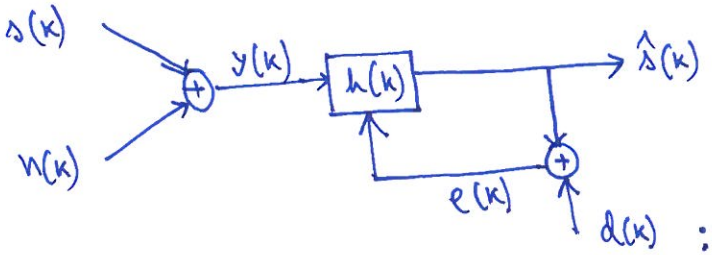
- Assumption : $s(k)$ and $n(k)$ are statistically independent.
  So, $\mathcal{E}\{s(k)\, n(i)\} = 0$, $\forall\, k, i$.

Linear MMSE Estimators

- We do some linear operation on $y(k)$ and get an $\hat{s}(k)$.
  However MMSE means minimum $\mathcal{E}\{|s(k) - \hat{s}(k)|^2\}$.

Non-causal IIR Wiener Filter

- Linear Filter



$$\hat{s}(k) = \sum_{l=-\alpha}^{+\alpha} h(l)\, y(k-l) = y(k) * h(k)$$

- Assume target signal $d(k) = s(k)$. and error $e(k) = \hat{s}(k) - s(k)$.
- minimize $E\{e^2(k)\} = E\{(\hat{s}(k) - d(k))^2\}$

$$= E\{(\hat{s}(k) - s(k))^2\}.$$

$$= E\left\{\left(\sum_{l=-}^{-\alpha} h(l)\, y(k-l)\right) - s(k)\right\}^2$$

- $\forall i, \quad \dfrac{\partial E\{e^2(k)\}}{\partial h(i)} = 0$

or, $\quad E\left\{2e(k) \cdot \dfrac{\partial e(k)}{\partial h(i)}\right\} = 0$

or, $\quad E\left\{e(k) \cdot \dfrac{2\hat{s}(k)}{2h(i)}\right\} = 0$

or, $\quad E\{e(k) \cdot y(k-i)\} = 0$

or, $\quad E\{(\hat{s}(k) - s(k))\, y(k-i)\} = 0$

or, $\quad E\{\hat{s}(k)\, y(k-i)\} = E\{s(k)\, y(k-i)\}$

or, $\quad E\left\{\left(\sum_{l=-\alpha}^{+\alpha} h(l)\, y(k-l)\right) y(k-i)\right\} = E\{s(k)\, y(k-i)\}$

or, $\quad \boxed{\sum_{l=-\alpha}^{+\alpha} h(l)\, \varphi_{yy}(i-l) = \varphi_{ys}(i)}$

[Assumption : all signals are wide-sense stationary (second-order stationary)

$$\sum_{\lambda=-\alpha}^{+\alpha} h(\lambda)\, \phi_{yy}(i-\lambda) = \phi_{ys}(i)$$

or, $\quad h(i) * \phi_{yy}(i) = \phi_{ys}(i) \qquad$ [convolution relation]

• Note: Convolution of two signals can be represented by multiplication in the power spectral domain.

• $$H(e^{j\Omega}) \cdot \Phi_{yy}(e^{j\Omega}) = \Phi_{ys}(e^{j\Omega})$$

or, $$H(e^{j\Omega}) = \frac{\Phi_{ys}(e^{j\Omega})}{\Phi_{yy}(e^{j\Omega})}$$

$$\left[h(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} H(e^{j\Omega})\, e^{j\Omega k}\, d\Omega \right.$$

$$\left. \text{and also, assuming } \Phi_{yy}(e^{j\Omega}) \neq 0 \right]$$

• Now, $\quad y(k) = s(k) + n(k)$

$$\therefore \quad \mathcal{E}\{ y(k)\, y(k-i) \} = \mathcal{E}\{ (s(k)+n(k))\,(s(k-i)+n(k-i)) \}$$

$$= \mathcal{E}\{ s(k)\, s(k-i) \} + \mathcal{E}\{ n(k)\, s(k-i) \} + \mathcal{E}\{ s(k)\, n(k-i) \} + \mathcal{E}\{ n(k)\, n(k-i) \}$$

or, $$\phi_{yy}(i) = \phi_{ss}(i) + \phi_{nn}(i)$$

[as $s(k)$ and $n(k)$ are statistically independent]

or, $$\sum_{\forall i} \phi_{yy}(i)\, e^{-j\Omega i} = \sum_{\forall i} \phi_{ss}(i) e^{-j\Omega i} + \sum_{\forall i} \phi_{nn}(i) e^{-j\Omega i}$$

or, $$\Phi_{yy}(e^{j\Omega}) = \Phi_{ss}(e^{j\Omega}) + \Phi_{nn}(e^{j\Omega})$$

$$H(e^{j\Omega}) = \frac{\Phi_{ys}(e^{j\Omega})}{\Phi_{yy}(e^{j\Omega})}$$

$$\varphi_{ys}(i) = E\{s(k)\, y(k-i)\}$$

$$= E\{ s(k) \left(s(k-i) + n(k-i) \right) \}$$

$$= E\{ s(k)s(k-i) \} + E\{ s(k)\, n(k-i) \}$$

($s(k)$ and $n(k-i)$ are statistically independent, so $E\{s(k)\, n(k-i)\} \to 0$)

$$= \varphi_{ss}(i)$$

$$\therefore \quad \phi_{yy}(e^{j\Omega}) = \sum_{\forall i} \varphi_{ys}(i) e^{-j\Omega i}$$

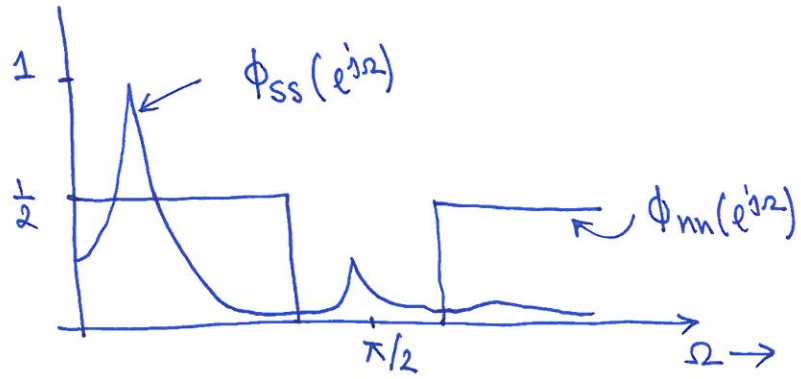
$$= \sum_{\forall i} \varphi_{ss}(i) e^{-j\Omega i} = \phi_{ss}(e^{j\Omega}).$$

$$\boxed{H(e^{j\Omega}) = \frac{\phi_{ys}(e^{j\Omega})}{\phi_{yy}(e^{j\Omega})} = \frac{\phi_{ss}(e^{j\Omega})}{\phi_{ss}(e^{j\Omega}) + \phi_{nn}(e^{j\Omega})}} \quad \text{Filter PSD.}$$

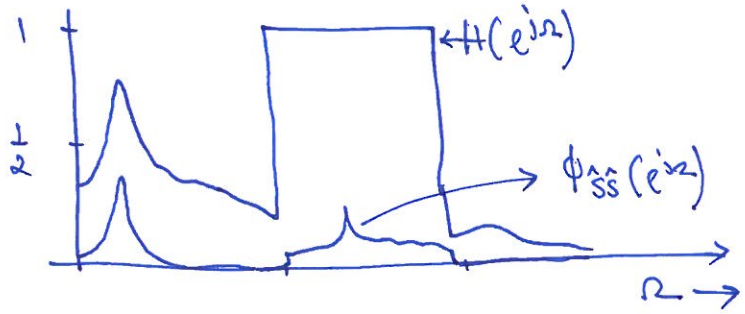
$$= \frac{\frac{\phi_{ss}(e^{j\Omega})}{\phi_{nn}(e^{j\Omega})}}{1 + \frac{\phi_{ss}(e^{j\Omega})}{\phi_{nn}(e^{j\Omega})}}.$$

$$\hat{S}(e^{j\Omega}) = H(e^{j\Omega}) \cdot Y(e^{j\Omega}) \qquad \text{Output PSD.}$$

The non-causal IIR Wiener Filter evaluates signal-to-noise (SNR) ratio $\frac{\phi_{ss}(e^{j\Omega})}{\phi_{nn}(e^{j\Omega})}$ at a given frequency Ω.



- When SNR is large,
  $H(e^{j\Omega}) \rightarrow 1$.
  The corresponding frequency component comes out without attenuation.



- When SNR is low,
  $\phi_{ss}(e^{j\Omega}) \ll \phi_{nn}(e^{j\Omega})$
  and $H(e^{j\Omega}) \approx 0$.
  So, corresponding frequency component is attenuated.

- Important Note : The noise reduction task will be efficient if the speech signal and noise do not occupy the same frequency bands.

FIR WIENER FILTER

- We now restrict the impulse response $h(k)$ of the optimal filter to be of finite and even order N, and to be causal, i.e.,

$$h(k) = \begin{cases} \text{arbitrary} & 0 \le k \le N \\ 0 & \text{otherwise} \end{cases}$$

- So, $\hat{s}(k) = h(k) * y(k) = \sum_{l=0}^{N} h(l)\, y(k-l)$

- We again try to minimize

$$E\{e^2(k)\} = E\{(\hat{s}(k) - d(k))^2\}$$

- Assuming $d(k) = s(k)$, we will again can get

$$\sum_{l=0}^{N} h(l)\, \varphi_{yy}(i-l) = \varphi_{ys}(i) \quad , \forall i$$

also, due to statistical independence $\varphi_{ys}(i) = \varphi_{ss}(i)$.

$$\therefore \sum_{l=0}^{N} h(l)\, \varphi_{yy}(i-l) = \varphi_{ss}(i) \quad , \forall i$$

$$\begin{bmatrix} \varphi_{yy}(0) & \varphi_{yy}(1) & \cdots & \varphi_{yy}(N) \\ \varphi_{yy}(1) & \varphi_{yy}(0) & \cdots & \varphi_{yy}(N-1) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{yy}(N) & \cdots & \cdots & \varphi_{yy}(0) \end{bmatrix} \begin{bmatrix} h(0) \\ h(1) \\ \vdots \\ h(N) \end{bmatrix} = \begin{bmatrix} \varphi_{ss}(0) \\ \varphi_{ss}(1) \\ \vdots \\ \varphi_{ss}(N) \end{bmatrix}$$

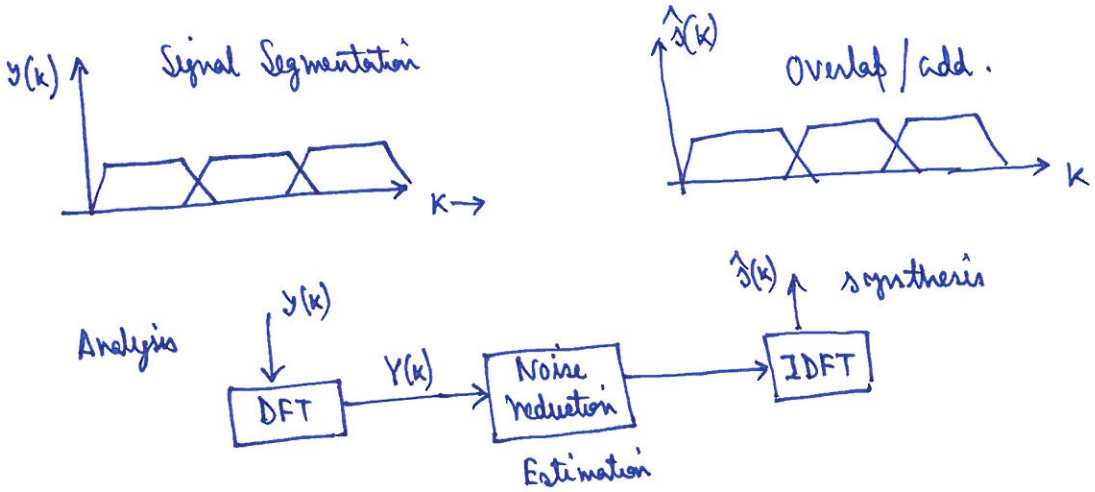
R_{yy} $\quad \underline{h}$ $\quad \underline{\varphi}_{ss}$

$$\therefore R_{yy}\, \underline{h} = \underline{\varphi}_{ss}$$

$$\underline{h} = R_{yy}^{-1}\, \underline{\varphi}_{ss} .$$

Speech Enhancement in the DFT domain

- IIR wiener filter is a time domain filter, but it gives us clear understanding in the frequency domain.
- For a time domain signal, the frequency domain behavior and techniques can be realized by using either Fourier Transform or filter banks.
- Let us treat the estimation in the DFT domain.



- M-size segment (or frame)

$$\underline{y}(k) = \left(y(k-M+1) \quad y(k-M+2) \quad \cdots \quad y(k) \right)^T$$

$$\underline{s}(k) = \left(s(k-M+1) \quad s(k-M+2) \quad \cdots \quad s(k) \right)^T$$

$$\underline{n}(k) = \left(n(k-M+1) \quad n(k-M+2) \quad \cdots \quad n(k) \right)^T$$

- An analysis window $\underline{w} = \left(w(0) \quad w(1) \quad \cdots \quad w(M-1) \right)^T$
- $\underline{Y}(k) = \text{DFT} \left\{ \underline{w} \otimes \underline{y}(k) \right\}$
  $\underline{S}(k) = \text{DFT} \left\{ \underline{w} \otimes \underline{s}(k) \right\}$
  $\underline{N}(k) = \text{DFT} \left\{ \underline{w} \otimes \underline{n}(k) \right\}$

[Here $\otimes$ denotes multiplication element-by-element]

- $\underline{Y}(k) = \left(Y_0(k) \ \cdots \ Y_\mu(k) \ \cdots \ , Y_{M-1}(k) \right)^T$

  same for $\underline{S}(k)$ and $\underline{N}(k)$.

Now, we come back to Wiener Filter

- Time domain convolution $\longleftrightarrow$ Frequency (Fourier) domain multiplication
- Enhanced signal $\hat{\underline{s}}(k)$ has the DFT $\hat{\underline{S}}(k)$
- $$\hat{\underline{S}}(k) = \underline{H}(k) \otimes \underline{Y}(k)$$

  where $\underline{H}(k) = \text{DFT}\left\{ \underline{h}(k) \right\}$

- We want to minimize

  $$E\left\{ \| \underline{s}(k) - \hat{\underline{s}}(k) \|^2 \right\}$$

- But for windowed signal, we have to minimize

  $$E\left\{ \| (\underline{w} \otimes \underline{s}(k)) - (\underline{w} \otimes \hat{\underline{s}}(k)) \|^2 \right\}$$

  $$= E\left\{ \| \underline{S}(k) - \hat{\underline{S}}(k) \|^2 \right\}$$ [As $\underline{S}(k) = \text{DFT}\left\{ \underline{w} \otimes \underline{s}(k) \right\}$ and DFT is an orthonormal transform]

- assuming statistical indepence between the DFT coefficients, we can minimize

  $$E\left\{ | S_\mu(k) - \hat{S}_\mu(k) |^2 \right\}, \ \forall \mu .$$

- $E\left\{|S_\mu(k)-\hat{S}_\mu(k)|^2\right\} = E\left\{\left(S_\mu(k)-H_\mu(k)Y_\mu(k)\right)\left(S_\mu(k)-H_\mu(k)Y_\mu(k)\right)^*\right\}$

- (a) $\dfrac{2E\left\{|S_\mu(k)-\hat{S}_\mu(k)|^2\right\}}{2\,Re\left\{H_\mu(k)\right\}} = 0$

$$\Rightarrow \quad Re\left\{H_\mu(k)\right\} = \frac{E\left\{|S_\mu(k)|^2\right\}}{E\left\{|Y_\mu(k)|^2\right\}} = \frac{E\left\{|S_\mu(k)|^2\right\}}{E\left\{|S_\mu(k)|^2\right\}+E\left\{|N_\mu(k)|^2\right\}}$$

(b) $\dfrac{2E\left\{|S_\mu(k)-\hat{S}_\mu(k)|^2\right\}}{2\,Im\left\{H_\mu(k)\right\}} = 0$

$$\Rightarrow \quad Im\left\{H_\mu(k)\right\} = 0.$$

Therefore, $$H_\mu(k) = \frac{E\left\{|S_\mu(k)|^2\right\}}{E\left\{|S_\mu(k)|^2\right\}+E\left\{|N_\mu(k)|^2\right\}}.$$

- The DFT based solution is straight-forward. But, suffers in a very difficult practical problem.

  - The clean speech frequency domain information is not available.

- To get a practical solution, we discuss "SPECTRAL SUBTRACTION".

Spectral Subtraction

- The basic idea is to subtract an estimate of the noise floor from an estimate of the spectrum of the noisy signal.

- $\Omega_\mu = \frac{2\pi}{M} \cdot \mu \quad , \mu = 0, 1, \cdots$

- $$E\left\{|S_\mu(k)|^2\right\} = E\left\{|Y_\mu(k)|^2\right\} - E\left\{|N_\mu(k)|^2\right\}$$
$$= E\left\{|Y_\mu(k)|^2\right\}\left[1 - \frac{E\left\{|N_\mu(k)|^2\right\}}{E\left\{|Y_\mu(k)|^2\right\}}\right]$$
$$= E\left\{|Y_\mu(k)|^2\right\} \cdot \left|\tilde{H}_\mu(k)\right|^2$$

- A DFT may be interpreted as an analysis filter bank, $E\left\{|S_\mu(k)|^2\right\}$ represents the power of a complex-valued subband signal $S_\mu(k)$ for any fixed μ.

- The spectral subtraction method may be interpreted as a filter with frequency response

$$\tilde{H}_\mu(k) = \sqrt{1 - \frac{E\left\{|N_\mu(k)|^2\right\}}{E\left\{|Y_\mu(k)|^2\right\}}} = \sqrt{\frac{E\left\{|S_\mu(k)|^2\right\}}{E\left\{|Y_\mu(k)|^2\right\}}} = \sqrt{H_\mu(k)}.$$

- ~~This~~ Since we are subtracting in the PSD domain, this approach is called "power subtraction".

- We can use other variations of the spectral gain function, such as "magnitude subtraction"

$$\sqrt{E\{|Y_\mu(k)|^2\}} - \sqrt{E\{|N_\mu(k)|^2\}} = \sqrt{E\{|Y_\mu(k)|^2\}} \left[1 - \frac{\sqrt{E\{|N_\mu(k)|^2\}}}{\sqrt{E\{|Y_\mu(k)|^2\}}} \right].$$

These variations can be written in a generalized version

$$|\tilde{H}_\mu(k)|^2 = \left[1 - \left(\frac{E\{|N_\mu(k)|^2\}}{E\{|Y_\mu(k)|^2\}} \right)^\beta \right]^\alpha$$

- Using magnitude squared DFT spectra and short-time estimates $|\hat{N}_\mu(k)|^2$ and $|\hat{Y}_\mu(k)|^2$, a practical final approach can be

$$|\hat{S}_\mu(k)|^2 = |Y_\mu(k)|^2 \left[1 - \left(\frac{|\hat{N}_\mu(k)|^2}{|\hat{Y}_\mu(k)|^2} \right)^\beta \right]^\alpha.$$

  - One point: In practice we have to make sure $|\hat{S}_\mu(k)|^2$ is non-negative and real valued.