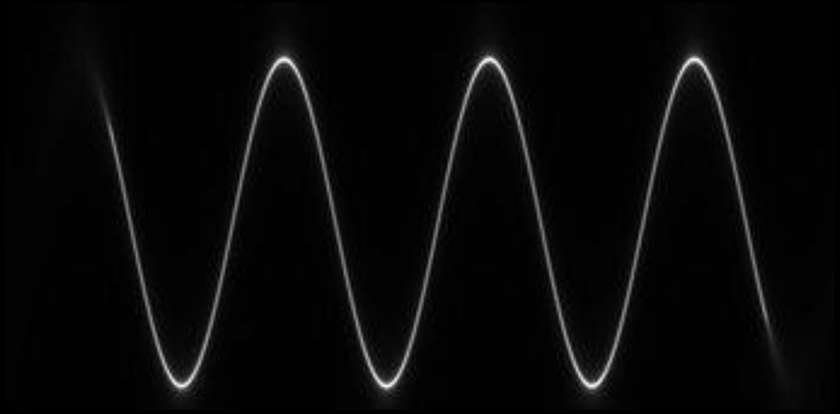
Speech Enhancement in TTS

Ananyapam De

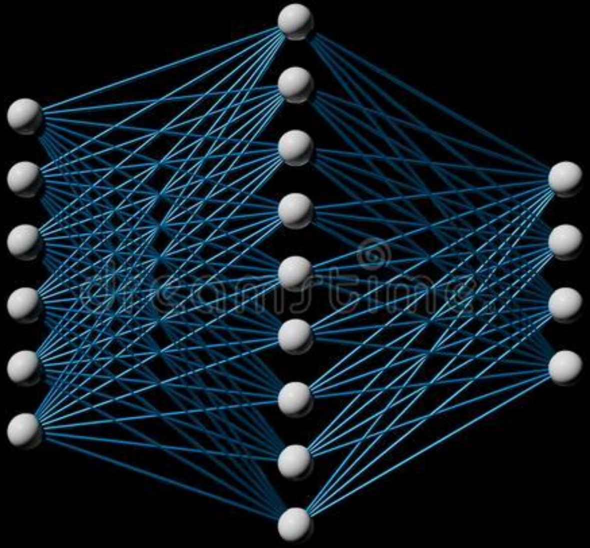
Flow:

- Categories of Speech Enhancement methods
- Datasets
- Brief description, math and intuition of the implemented filters and examples
- Hypothesis
  - Which method works best for which noise class?
  - Which method works best for which SNR level?
- Intelligibility metrics and Observations
- Ongoing work

Speech Enhancement Techniques



Signal Processing



Deep Learning

A. Spectral-Subtractive Algorithms

- *Assuming additive noise to the signal:* `y(n) = z(n) + s(n)`

A. Statistical Model Based Algorithms

- *Estimating the spectrum of clean signal using the statistical techniques*

A. Subspace Algorithms

- *Assuming the noisy signal can be decomposed as a direct sum of the subspaces containing the clean signal and the pure noise*

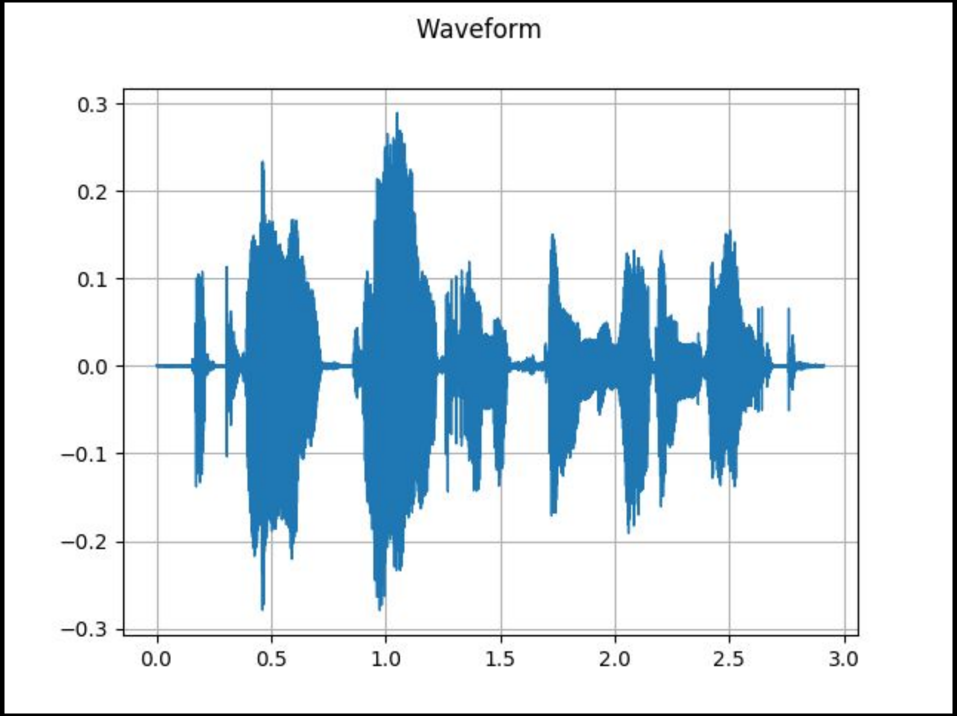
Real world dataset: NOIZEUS

30 audio samples + distorted to different degrees and different noise classes

SNR Ratios: A) 0 dB B) 5 dB C) 10 dB D) 15dB

Exhibition	**Car**	**Airport**	**Train**
Babble	**Restaurant**	**Station**	**Street**

Clean Audio Sample

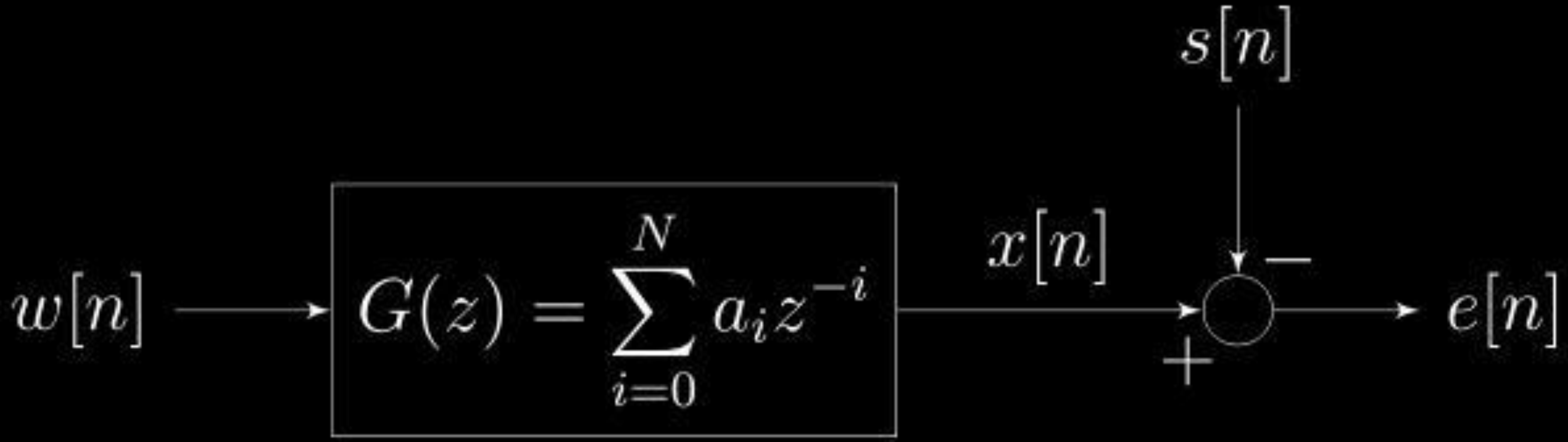


Implemented Filters-

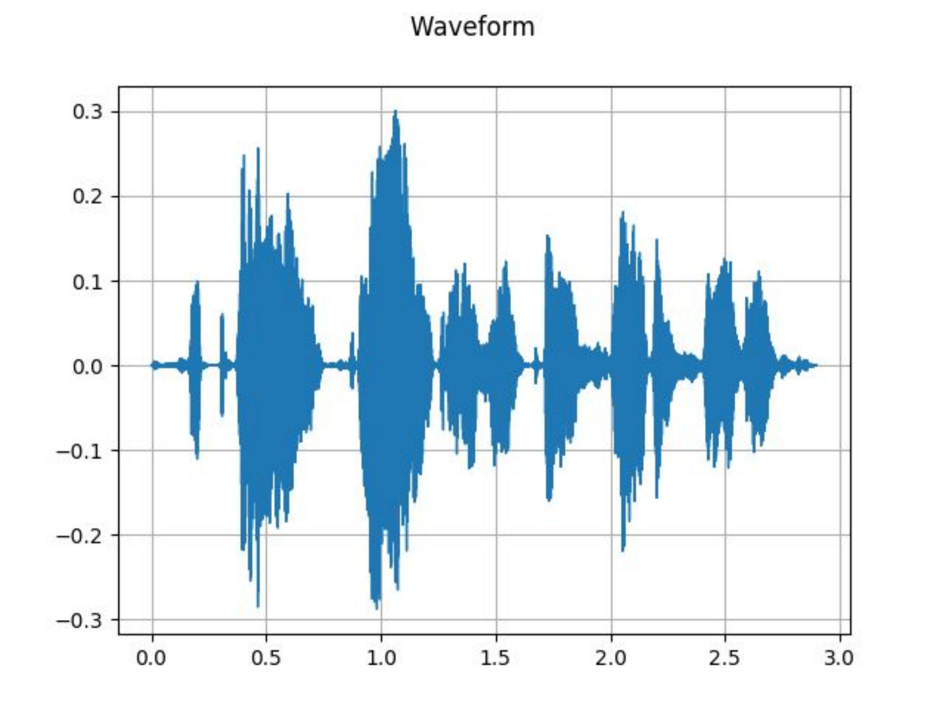
1. *Wiener Filter*
2. *Kalman Filter*
3. *Spectral Subtraction (Over subtraction)*
4. *Bayesian MMSE Filter*
5. *Bayesian MMSE Log Filter*

Wiener Filter

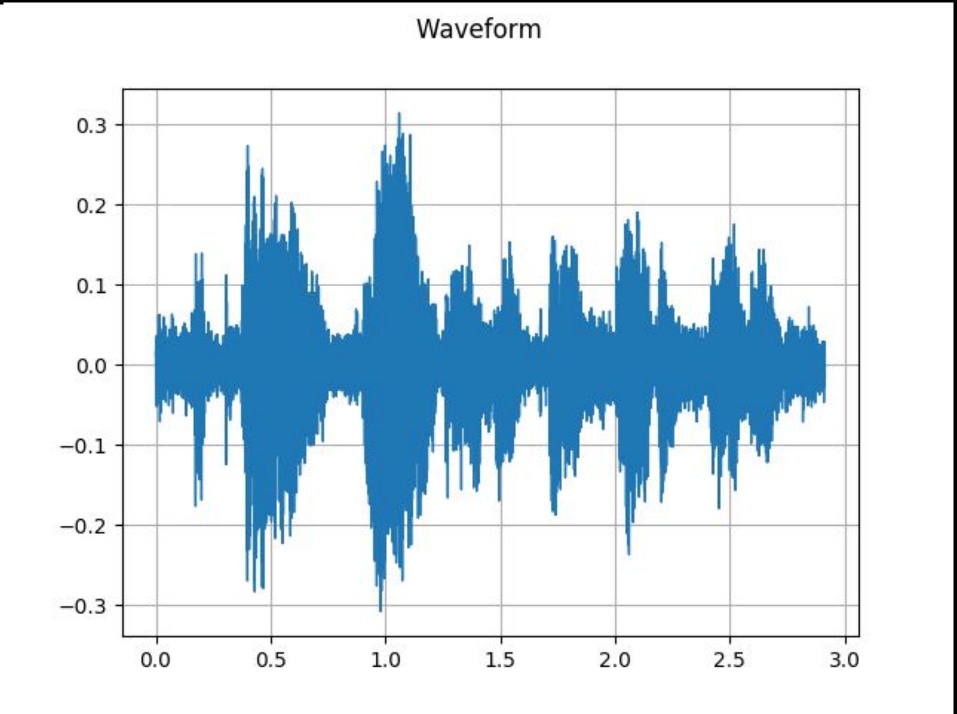
- The optimal **linear** complex spectral estimator which minimizes the expected mean squared error
- Constraint: L**inear + time invariant**



Wiener Filter



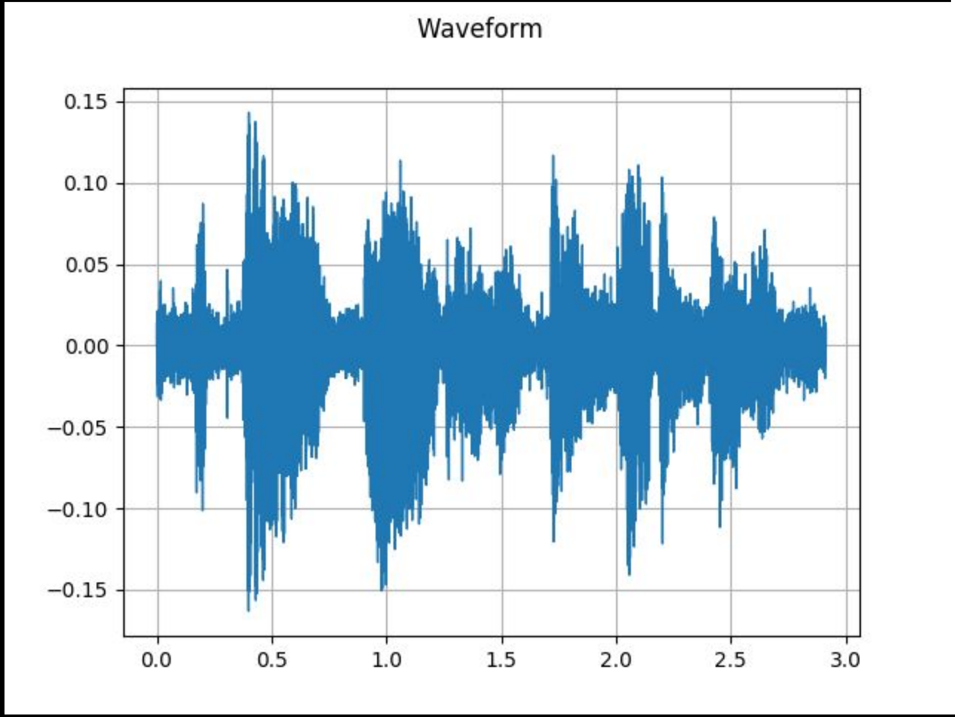
Noisy Sample



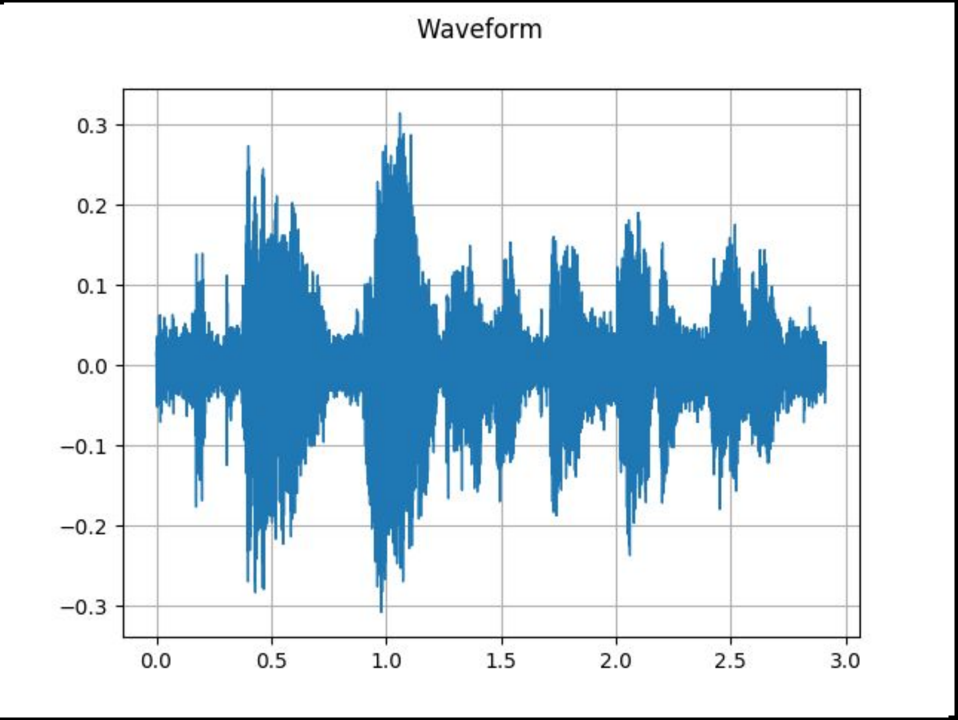
Intuitions and Properties

- When SNR ratio is high, the filter does not provide noise reduction, that is, the noisy signal passes unaltered (hence no speech distortion).
- When SNR ratio is extremely low, the output of the Wiener filter is heavily attenuated, which provides undesirable distortion in the speech.
- Notice a tradeoff between no speech enhancement and undesirable speech distortion.

Kalman Filter

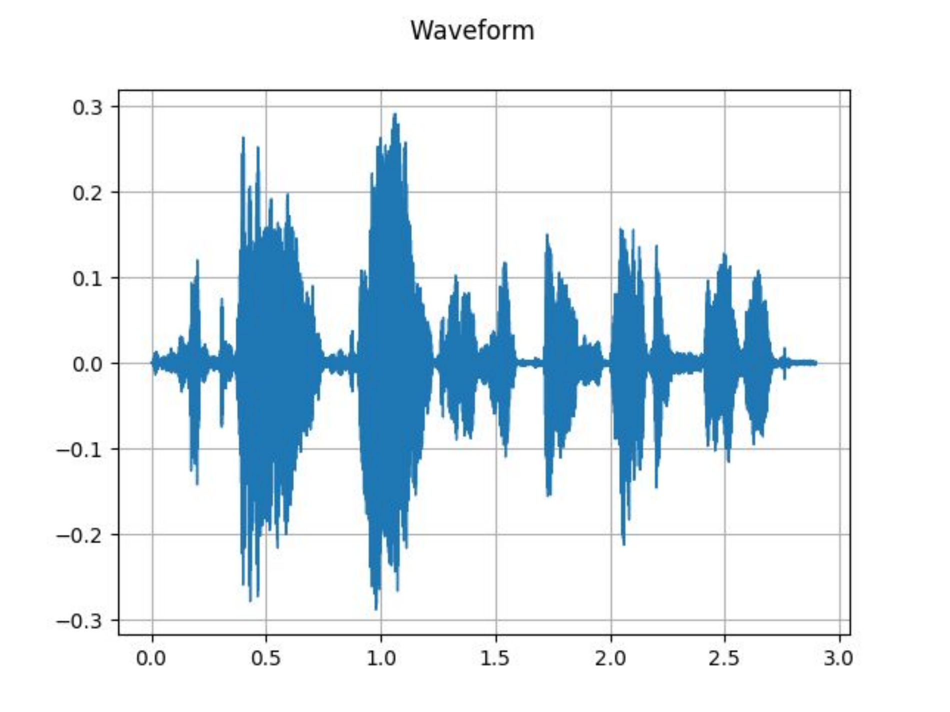


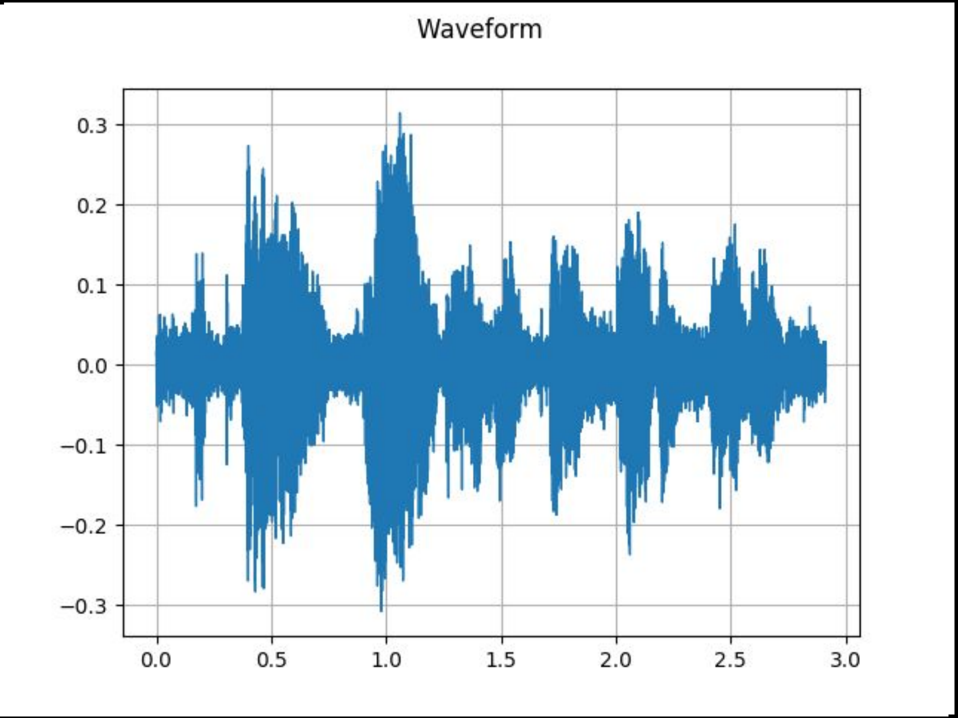
Noisy Sample



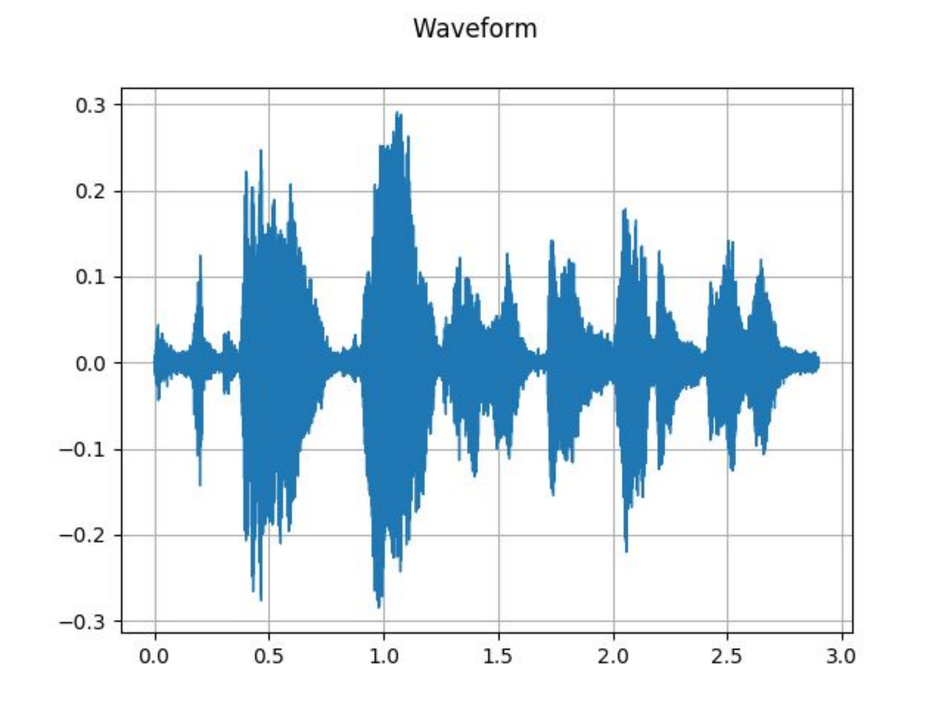
OverSubtraction

Noisy Sample

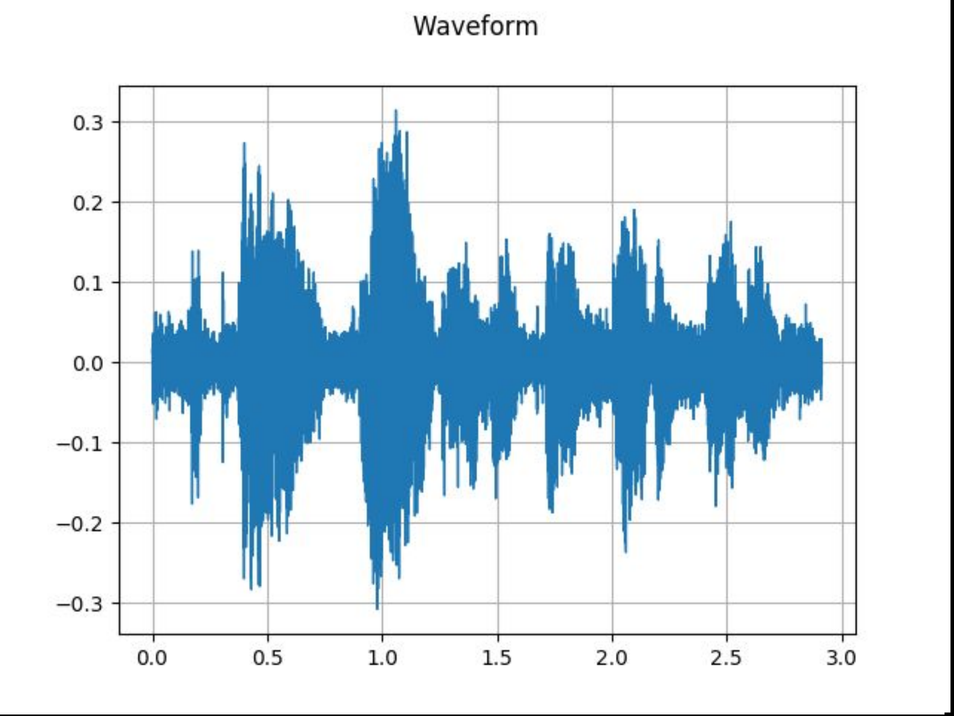




Bayesian MMSE Filter

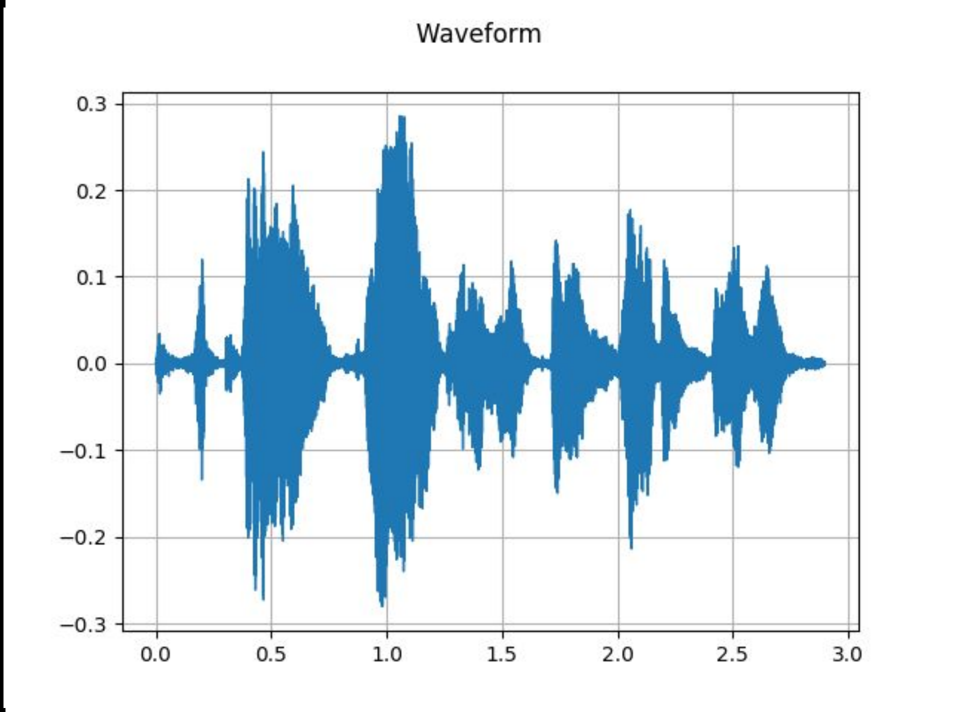


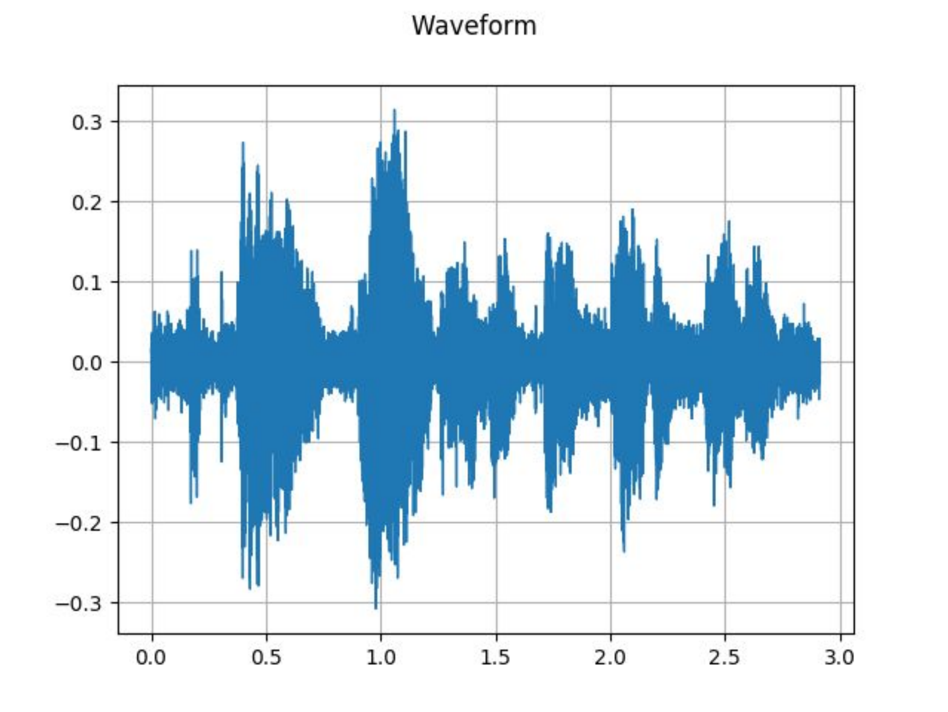
Noisy Sample



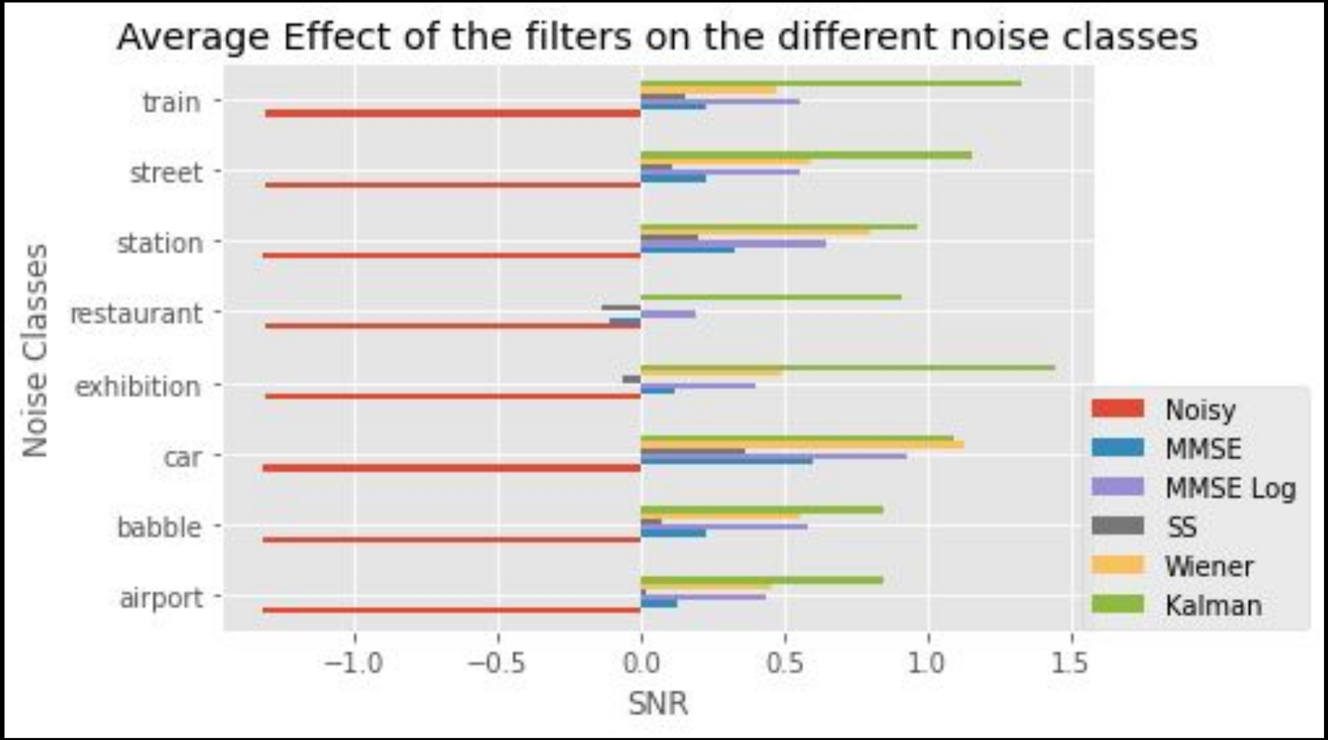
Bayesian MMSE Log Filter

Noisy Sample

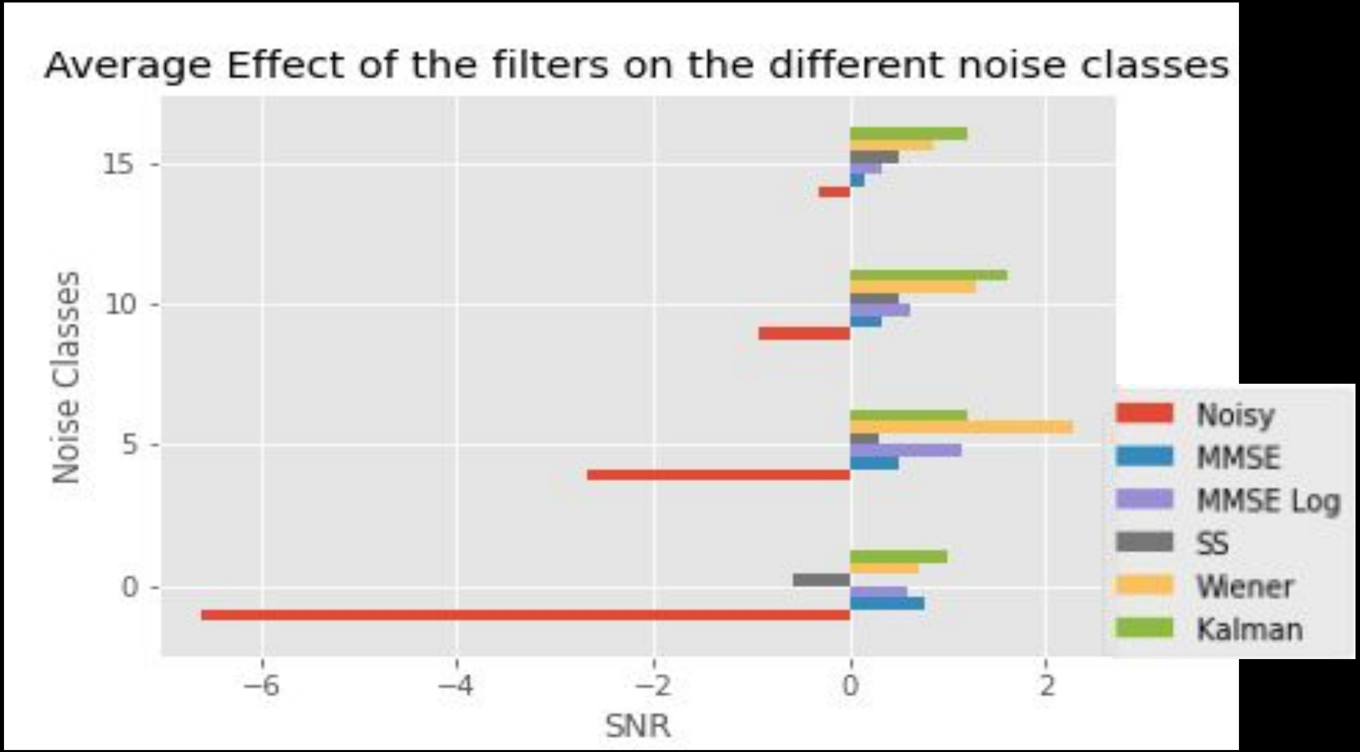




Filter performance (SNR) on noise classes-



Filter performance (SNR) on different SNR levels-



Observations on NOIZEUS

Notice their overall behaviours-

Kalman >~ Wiener > MMSE Log > MMSE > Spectral Subtraction

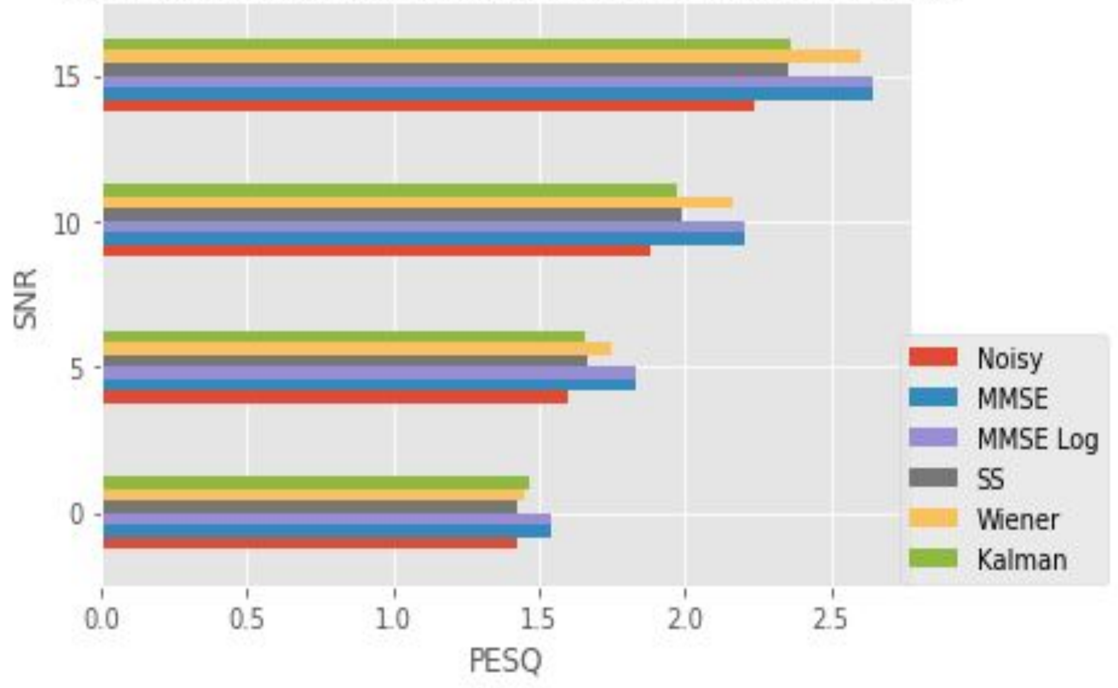
- Kalman filters works really well because the assumptions under which it is based are very realistic.
- Wiener Filter performs well on mid ranged SNR audios ie from range 5 SNR to 10 SNR
- MMSE Log almost always performs better than MMSE as it's an improvement on the previous one
- Over Subtraction method does not perform well because of it's extremely sensitive hyperparameters

Intelligibility Metrics/ Voice Quality Test Algorithms

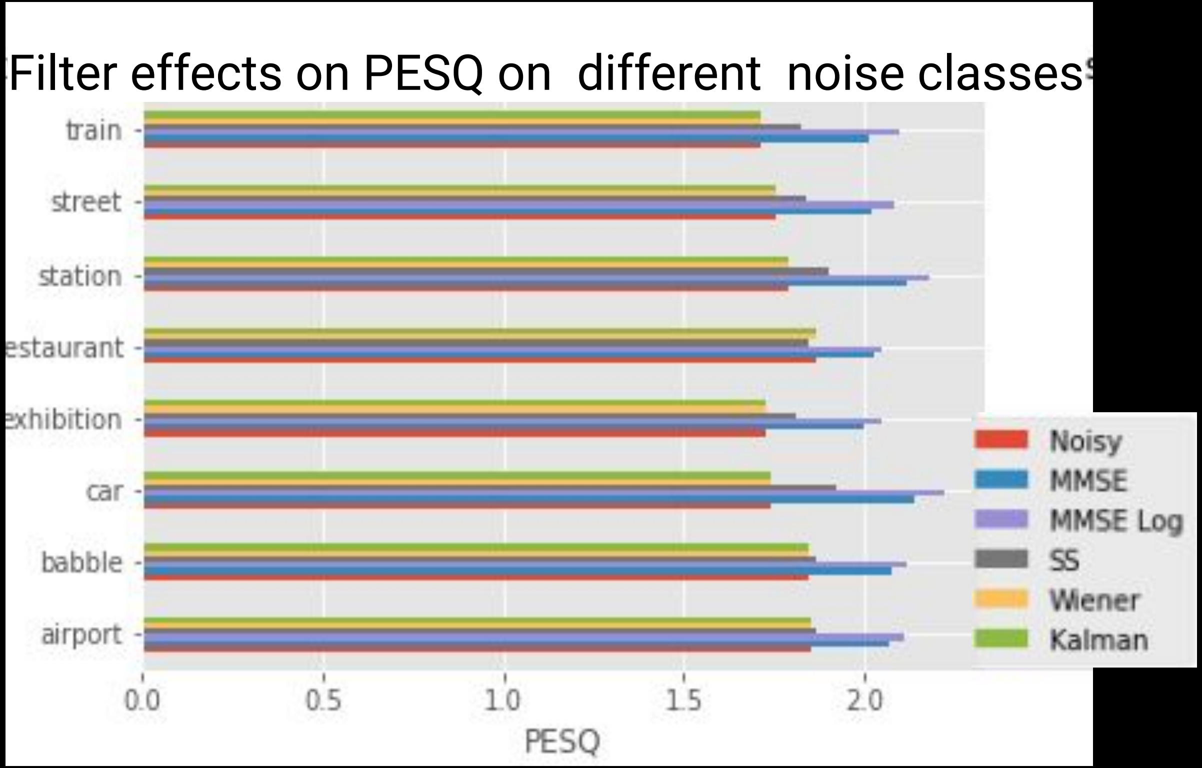
- PESQ (Perceptual Evaluation of Speech Quality)
- STOI (Short Term Objective Intelligibility)
- MCD (mel-cepstral distortion)
- GPE (Gross Pitch Error)
- FFE (F0- Frame Error)

PESQ Metric

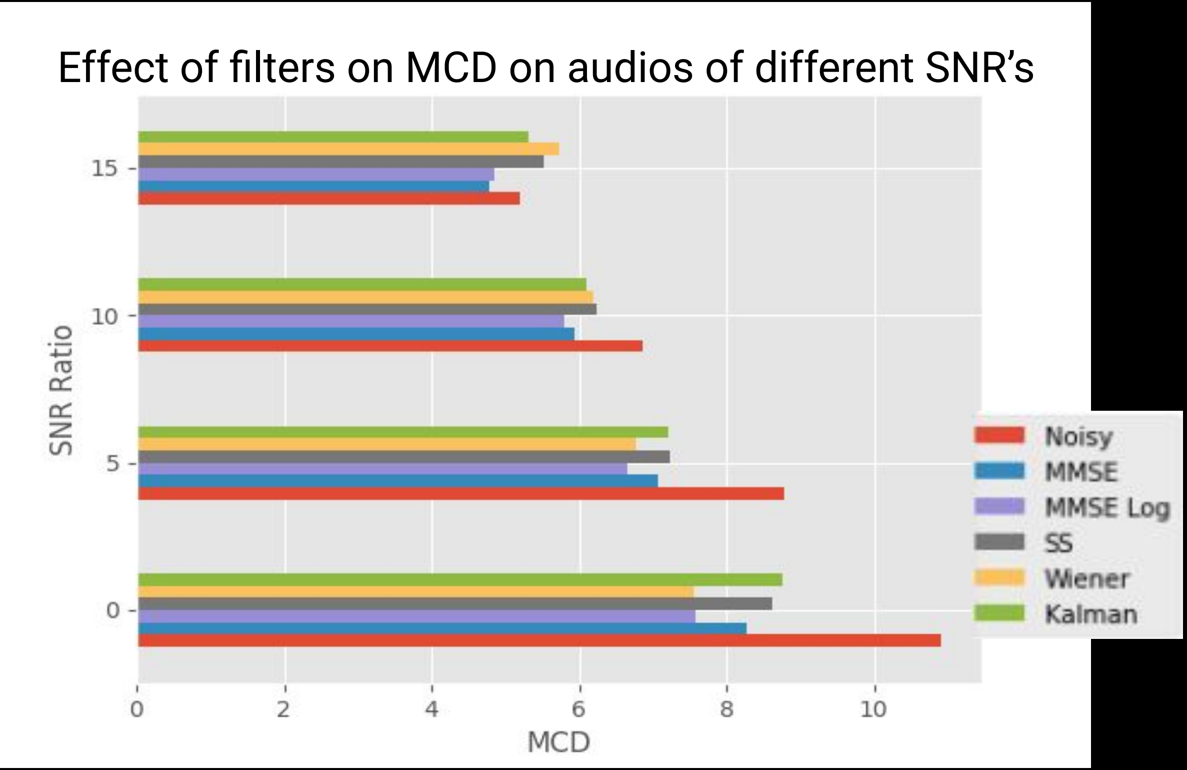
Effect of filters on PESQ on audios of different SNR's



PESQ Metric


Filter effects on PESQ on different noise classes
train
street
station
estaurant
exhibition
car
babble
airport
0.0
0.5
1.0
1.5
2.0
PESQ
Noisy
MMSE
MMSE Log
SS
Wiener
Kalman

MCD Metric



Effect of filters on MCD on audios of different SNR's
SNR Ratio
15
10
5
0
0
2
4
6
8
10
MCD
Noisy
MMSE
MMSE Log
SS
Wiener
Kalman

Observations on the TTS Dataset

Clean Audio

Filtered Audio

- Pretty bad :(
- Signal Processing Filters do not perform well on the TTS dataset unlike real world noise datasets
- Possible reasons: TTS dataset does not satisfy the assumptions of the filters
  - Noise is not additive as NOIZEUS is created
  - Can’t expect subspace decomposition
  - Very subtle noise instead of coarse noise

Ongoing Work!!

- Testing out some more metrics!
- Deep Learning Methods on the TTS Dataset
    - Facebook Denoiser
    - RNN Noise
- How do our methods perform on the different types of noise on the TTS Dataset


nk you

Over subtraction [Berouti et al]

- Subtract an overestimate of the noise power spectrum, while preventing the spectral components from going below a preset minimum value
- (α: Oversubtraction factor and β:Spectral floor parameter)

$$|\hat{X}(\omega)|^2 = \begin{cases} |Y(\omega)|^2 - \alpha |\hat{D}(\omega)|^2 & \text{if } |Y(\omega)|^2 > (\alpha+\beta)|\hat{D}(\omega)|^2 \\ \beta |\hat{D}(\omega)|^2 & \text{else} \end{cases}$$

Appendix-

Kalman Filter (Linear Quadratic Estimation) [Orchisman]

- Recursive State Estimation technique
- Most optimal filter under these assumptions of normality
- Assume that noise is gaussian

$$\mathbf{x}_k = \boldsymbol{F}_k \mathbf{x}_{k-1} + \boldsymbol{B}_k \mathbf{u}_k + \boldsymbol{w}_k \text{ where } \boldsymbol{w}_k \sim N(\boldsymbol{0}, \boldsymbol{Q}_k)$$

$$\mathbf{z}_k = \boldsymbol{H}_k \mathbf{x}_k + \boldsymbol{v}_k \text{ where } \boldsymbol{v}_k \sim N(\boldsymbol{0}, \boldsymbol{R}_k)$$

Bayesian MMSE Filter

- Optimal estimators that minimized the mean-square error between the estimated and true magnitude spectra-

$$e = E\left\{\left(\hat{X}_k - X_k\right)^2\right\}$$

$$BMSE(\hat{X}_k) \;\; = \int\int (X_k - \hat{X}_k)^2 \;\; p(\mathbf{Y}, X_k)\; d\mathbf{Y} dX_k$$

Bayesian MMSE Log Filter

- Optimal estimators that minimized the mean-square error between the estimated and true **log** magnitude spectra-

$$E\{(\log X_k - \log \hat{X}_k)^2\}$$