Digital Audio and Speech Processing (IN4182)

Noise PSD Estimation

Richard C. Hendriks
Overview of single-channel NR algorithm

- Wiener gain: \( \hat{s}_k(l) = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_N^2} y_k(l) \)
- \( \hat{s}_k(l) = E[S|y] = g(\sigma_N^2, \sigma_S^2, y, \nu, \gamma) y_k(l) \)
- power spectral subtraction
Lecture 2:

Conclusions MMSE Estimation for Noise Reduction:

- Wiener gain is the optimal LINEAR MSE estimator.

- The global optimal MSE estimator is \( E[S|y] = g(\sigma_N^2, \sigma_S^2, y, \nu, \gamma) \cdot y \) and is generally non-linear with respect to \( y \).

- for Rayleigh \( p_A(a) \) (and thus a Complex-Gaussian pdf for the complex-DFTs) Wiener gain results.

\[
g(\sigma_N^2, \sigma_S^2, y, \nu, \gamma) = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_N^2}
\]

- \( g(\sigma_N^2, \sigma_S^2, y, \nu, \gamma) \) is always real and positive. Thus, noisy phase is always used (and is optimal.)
DFT Based Noise Suppression

Today:

Noise PSD estimation

- VAD
- Minimum Statistics
- MMSE based with speech presence uncertainty.
Why PSD Estimation?

- Recall signal model:
  \[
  Y_k(l) = S_k(l) + N_k(l).
  \]

- All estimators that we discussed are function of the speech and/or noise PSD
  \[
  P_{SS,k} = \frac{1}{L} E \left[ |S_k(l)|^2 \right] \quad \text{and} \quad P_{NN,k} = \frac{1}{L} E \left[ |N_k(l)|^2 \right].
  \]

- Examples:
  
  - Power spectral subtraction
    \[
    \hat{s}_k(l) = \left( \max \left\{ 1 - \frac{E[|N_k(l)|^2]}{|y_k(l)|^2}, 0.2 \right\} \right)^{1/2} y_k(l).
    \]
Why PSD Estimation?

- Examples (continued):
  - Wiener Smoother

\[
\hat{s}_k(l) = \left(1 - \frac{\sigma^2_{N,k}(l)}{\sigma^2_{Y,k}(l)}\right) y_k(l)
\]

\[
= \frac{\sigma^2_{S,k}(l)}{\sigma^2_{S,k}(l) + \sigma^2_{N,k}(l)} y_k(l).
\]

- MMSE (Super-Gaussian Speech, Gaussian Noise)

\[
\hat{s}_k(l) = E \left[S_k(l) | y_k(l)\right] = g \left(\sigma^2_{N,k}(l), \sigma^2_{S,k}(l), y_k(l), \nu, \gamma\right) y_k(l)
\]
Why PSD Estimation?

Practically all speech enhancement estimators depend on the speech and noise PSD.

To estimate $P_{SS,k}(l)$ we already discussed the

- Maximum likelihood estimator
- Decision directed approach.

However, in practice we also have to make estimates $\hat{P}_{NN,k}(l)$. 
Voice Activity Detection

Voice activity detection (VAD) is a method that can be used to estimate $P_{NN,k}(l)$.

- Use the VAD to determine whether there is speech activity.
- Estimate noise psd prior to speech activity.
- Use this noise psd estimate during speech precense assuming the noise psd estimate remains valid.
Voice Activity Detection

A VAD can be implemented using a Bayesian hypothesis test:

\[ H_0 : \ Y_k(l) = N_k(l) \] (speech absence)

\[ H_1 : \ Y_k(l) = S_k(l) + N_k(l) \] (speech presence).

Based on statistical models for \( Y, S \) and \( N \) and a hypothesis criterium we automatically decide whether speech is absent or present.

The estimate \( \hat{\sigma}^2_{N,k}(\ell) \) can then be obtained by recursive smoothing, i.e.,

\[
\hat{\sigma}^2_{N,k}(\ell) = \begin{cases} 
\alpha \hat{\sigma}^2_{N,k}(\ell - 1) + (1 - \alpha) |y_k(\ell)|^2 & \text{when } H_0(\ell) \\
\hat{\sigma}^2_{N,k}(\ell - 1) & \text{when } H_1(\ell),
\end{cases}
\]

where \( 0 \leq \alpha \leq 1 \) is a smoothing constant.
Bayes Hypothesis test

How to decide between $H_0$ and $H_1$?

Decide $H_0$ when

$$P(H_0|y_k(l)) > P(H_1|y_k(l))$$

and $H_1$ when

$$P(H_0|y_k(l)) < P(H_1|y_k(l)).$$

This can be written more compact as

$$\frac{P(H_1|y_k(l))}{P(H_0|y_k(l))} \stackrel{H_1}{\gg} \frac{1}{H_0}$$
Bayes Hypothesis test

Using Bayes criterium:

\[ P(H_0 | y_k(l)) = \frac{p_Y(y_k(l) | H_0) P(H_0)}{p_Y(y_k(l))} \]

and

\[ P(H_1 | y_k(l)) = \frac{p_Y(y_k(l) | H_1) P(H_1)}{p_Y(y_k(l))} \]

we get

\[ \frac{p_Y(y_k(l) | H_1) P(H_1)}{p_Y(y_k(l) | H_0) P(H_0)} \underset{H_0}{\geq} 1. \]
Bayes Hypothesis test

This leads to

$$\Lambda_k(l) = \frac{p_Y(y_k(l)|H_1)}{p_Y(y_k(l)|H_0)} \frac{P(H_0)}{P(H_1)} = \lambda,$$

$$\Lambda_k(l) = \frac{p_Y(y_k(l)|H_1)}{p_Y(y_k(l)|H_0)}$$ is often called the likelihood ratio and $\lambda = \frac{P(H_0)}{P(H_1)}$ the threshold.

The Hypothesis test is thus based on a comparison between the likelihood ratio $\Lambda_k(l)$ and the threshold $\lambda$.

We need:

- A priori speech absence and speech presence probabilities (in bin $k$ and frame $l$),
- probability density functions.
Combining the Likelihoods

The hypothesis test \( \Lambda_k(l) = \frac{p_Y(y_k(l)|H_1)}{p_Y(y_k(l)|H_0)} \) can now be evaluated for each frequency bin and for each time frame.

To be more certain about the VAD we can combine the likelihood ratios from all frequencies into one decision by taking the geometric mean of the log likelihoods:

\[
T(l) = \frac{1}{L} \sum_{k=1}^{k=L} \log (\Lambda_k(l))
\]
VAD Example

![VAD Example Diagram]
Disadvantages of VAD

- The VAD itself is dependent on $\sigma_N^2$ via the distributions $p_Y(y_k(l)|H_1)$ and $p_Y(y_k(l)|H_0)$. The performance of the detector will degrade when the noise becomes more non-stationary and estimation errors leak into the hypothesis decision.

- The noise spectrum can only be updated when speech is absent.
Noise PSD Tracking During Speech Presence

Problem: Many natural noise sources are *not* stationary across duration of speech sentence.

Example:

Car noise:

Train noise:
Noise PSD Tracking During Speech Presence

**Problem:** Performance degrades if instantaneous noise psd drifts away from our snapshot taken in speech absence region.

\[ \sigma^2_{N,K}(l) \]

![Waveform graphs showing noise PSD tracking during speech presence.](image)
Minimum Statistics Based Noise
PSD Estimation

- Method to estimate power spectral density of non-stationary noise, given a single realization of a noisy speech signal,

- No explicit voice activity detection,

- Makes use of the fact that the power in a frequency band seen across time, regularly falls back to the level of the noise. By taking the minimum of the power across time, estimates of the noise level are obtained.
Recap: Speech Production

Recall simple speech production model:

Voiced speech (modeled):

- Periodic in time domain
- Line spectrum in freq. domain.
Recap: Speech Production

Noisy Speech:

- Additive noise: $Y_k(l) = S_k(l) + N_k(l)$.
- Uncorrelated speech and noise $P_{YY,k}(l) = P_{SS,k}(l) + P_{NN,k}(l)$.

Noisy Voiced Speech (modeled):

- There are frequency regions between harmonics with noise only.
- Voiced regions occupy up to 75% of speech activity time.
Unvoiced noisy speech (modeled):

- Signal is a mix of speech and noise at all frequencies.
- Unvoiced regions occupy much less time than voiced.
Recap: Speech Production

Time-frequency plane with noisy signal:

Observation:
At a given frequency bin $k$, the value of $P_{YY,k}(l)$ reaches the noise level $P_{NN,k}(l)$ in silence and (sometimes) in voiced regions!
Principles of MS Noise PSD Estimation

Observation:

• "Even during speech activity a short term power spectral density estimate of the noisy signal frequently decays to values which are representative of the noise power level".

• "By tracking the minimum power within a finite window...the noise floor can be estimated".
Principles of MS Noise PSD Estimation

How to estimate $\sigma^2_{N,k}(l)$?

Collect the last values of $P_{YY,k}(\ell)$ in a vector using a sliding window,

$$Q = \{P_{YY,k}(\ell - M + 1), \ldots, P_{YY,k}(\ell)\},$$

and take the minimum of $Q$, say $Q_{\min}$, to obtain estimate of $\sigma^2_{N,k}(l)$.

Taking $M$ sufficiently large it is guaranteed that $Q_{\min}$ originates from a time-frequency point without speech presence.
Principles of MS Noise PSD Estimation

Unfortunately, in practice we do not have access to $P_{YY,k}(l)$, but we can only observe *realizations* of underlying random variables:

Two problems:

1. High variance of $\frac{1}{L} |Y_k(l)|^2$

2. Choosing the minimum value of $Q$ within sliding window leads to negative bias.
Principles of MS Noise PSD Estimation

Real example:
Reducing the variance

**Problem 1:** Reducing the variance of $\frac{1}{L} |Y_k(l)|^2$.

Recap lec. 5 $\text{Var} \left\{ \frac{1}{L} |Y_k(l)|^2 \right\} = P^2_{Y_{Y,k}}(l)$ A method to reduce the variance is to compute a Bartlett estimate:

$$\hat{P}^{B}_{Y_{Y,k}}(l) = \frac{1}{M} \sum_{m=l-M+1}^{l} \frac{1}{L} |y_k(m)|^2,$$

Another method to reduce the variance is a so-called exponential smoother:

$$\hat{P}_{Y_{Y,k}}(l) = \alpha \hat{P}_{Y_{Y,k}}(l - 1) + (1 - \alpha) \frac{1}{L} |y_k(l)|^2.$$
Reducing the variance

**Problem 1:** Subsequently, store the smoothed values $\hat{P}_{YY,k}(l)$ in a vector $Q$ using a sliding window, i.e.,

$$Q = \left\{ \hat{P}_{YY,k}(l - M + 1), \ldots, \hat{P}_{YY,k}(l) \right\},$$

and take the minimum of $Q$, say $Q_{\text{min}}$, to obtain estimate of $\sigma_{N,k}^2(l)$.

Fig. 1. Periodogram $|Y(\lambda,k)|^2$, smoothed periodogram $P(\lambda,k)$ (3), $\alpha = 0.85$, and noise estimate $\tilde{\sigma}_N^2(\lambda,k)$ for a noisy speech signal and a single frequency bin $k = 25$. 
Bias Compensation

Problem 2: Choosing the minimum value of sliding vector $Q$ leads to negative bias.

Why? Consider for simplicity a sequence $\hat{P}_{Y,Y,k}(m)$, $m = l - M + 1, \ldots, l$ in stationary noise-only region.

- The samples $\hat{P}_{Y,Y,k}(l)$ are drawn from some distribution (e.g. the one shown to the right).

- In taking the minimum value of vector $Q$ we are generally sampling the density to the left of the mean $E\{\hat{P}_{Y,Y,k}(l)\} = \sigma_{N,k}^2(l)$. 

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{plot.png}
\end{figure}
Bias Compensation

Solution 2: Estimate bias from data and compensate.

- Define minimum of sliding window

\[ Q_{\text{min}} = \min \left\{ \hat{P}_{YY,k}(l - M + 1), \ldots, \hat{P}_{YY,k}(l) \right\}, \]

- Result:

\[ \hat{\sigma}_{N,k}^2(l) = Q_{\text{min}} \times B_{\text{min}} \]

Biased estimate    Bias compensation
Problem 3: Tracking delay.

Assume that $\sigma_{N,k}^2(l)$ is estimated using a sliding window with $M - 1$ past values of $\hat{P}_{YY,k}(l)$.

![Graph](image)

**Fig. 7.** Periodogram, smoothed periodogram, and noise estimate for a speech signal averaged over all frequency bins. The noise is switched on after about 780 frames.
MMSE based Noise PSD Estimation using Speech Presence Probability

Remember that the VAD based noise PSD estimator makes a hard decision between noise-only and noisy speech.

Instead of a hard decision, we could also make a soft decision in combination with an MMSE estimator.
MMSE based Noise PSD Estimation using Speech Presence Probability

\[ E[|N_k(l)|^2 | y_k(l)] = P(H_{0,k}(l) | y_k(l)) E[|N_k(l)|^2 | y_k(l), H_{0,k}] \]
\[ + P(H_{1,k}(l) | y_k(l)) E[|N_k(l)|^2 | y_k(l), H_{1,k}] \]

Assuming ergodicity we can approximate the outer expectation over \( Y \) by smoothing or averaging over time:

\[ \hat{\sigma}_{N,k}^2(l) = \alpha \hat{\sigma}_{N,k}^2(l - 1) + (1 - \alpha) E[|N_k(l)|^2 | y_k(l)] \]
MMSE based Noise PSD Estimation using Speech Presence Probability

\[
E[|N_k(l)|^2 | y_k(l)] = P(H_{0,k}(l) | y_k(l))E[|N_k(l)|^2 | y_k(l), H_{0,k}] \\
+ P(H_{1,k}(l) | y_k(l))E[|N_k(l)|^2 | y_k(l), H_{1,k}]
\]

Unknowns:

- \( P(H_{0,k}(l) | y_k(l)) \)
- \( P(H_{1,k}(l) | y_k(l)) \)
- \( E[|N_k(l)|^2 | y_k(l), H_{0,k}] \)
- \( E[|N_k(l)|^2 | y_k(l), H_{1,k}] \)
MMSE based Noise PSD Estimation using Speech Presence Probability

How to determine $E[|N_k(l)|^2|y_k(l), H_{0,k}]$ and $E[|N_k(l)|^2|y_k(l), H_{1,k}]$?

$H_0 : Y_k(l) = N_k(l) \rightarrow E[|N_k(l)|^2|y_k(l), H_{0,k}] = |y_k(l)|^2$

$H_1 : Y_k(l) = S_k(l) + N_k(l) \rightarrow E[|N_k(l)|^2|y_k(l), H_{1,k}]$ =?

How to determine $E[|N_k(l)|^2|y_k(l), H_{1,k}]$?

Make assumptions on the densities of $N$ and $S$ and perform Bayesian estimation.
MMSE based Noise PSD Estimation using Speech Presence Probability

How to determine $E[|N_k(l)|^2 |y_k(l), H_{1,k}]$?

Assuming Gaussian pdfs for $N$ and $S$ it follows that

$$E[|N_k(l)|^2 |y_k(l), H_{1,k}] = \left(\frac{1}{1 + \xi_k(l)}\right)^2 |y_k(l)|^2 + \frac{\xi_k(l)}{1 + \xi_k(l)} \sigma_{N,k}^2(l)$$

Notice that $\xi_k(l)$ is unknown. Using the maximum likelihood estimator $\hat{\xi}_k(l) = \frac{|Y|^2}{\sigma_{N,k}^2(l-1)} - 1$ we get

$$E[|N_k(l)|^2 |y_k(l), \hat{\xi}_k(l), H_{1,k}] = \hat{\sigma}_{N,k}^2(l-1)$$
MMSE based Noise PSD Estimation using Speech Presence Probability

Altogether:

$$E[|N_k(l)|^2 y_k(l)] = P(H_{0,k}(l)|y_k(l))|y_k(l)|^2 + P(H_{1,k}(l)|y)\hat{\sigma}_{N,k}^2(l - 1)$$

unknown
MMSE based Noise PSD Estimation using Speech Presence Probability

\[ P(H_{0,k}(l)|y_k(l)) = 1 - P(H_{1,k}(l)|y_k(l)) \]

\[ P(H_{1,k}(l)|y_k(l)) = \frac{P(H_{1,k}(l))p_{Y|H_1}(y_k(l)|H_{1,k}(l))}{P(H_{1,k}(l))p_{Y|H_1}(y_k(l)|H_{1,k}(l)) + P(H_{0,k}(l))p_{Y|H_0}(y_k(l)|H_{0,k}(l))} \]

\[ p_{Y|H_0} = \frac{1}{\sigma_N^2 \pi} \exp \left( -\frac{|y^2|}{\sigma_N^2} \right) \]

\[ p_{Y|H_1} = \frac{1}{\sigma_N^2 (1 + \xi_{H_1}) \pi} \exp \left( -\frac{|y^2|}{\sigma_N^2 (1 + \xi_{H_1})} \right) \]

Fixed (typical) a priori SNR when speech is present.
Tracking Delay - Increasing Noise levels

- True noise
- SPP
- MS
- VAD

Time (s)

noisy
MS
SPP
Overview of single-channel NR algorithm
Overview of single-channel NR algorithm

Target Estimate

- Wiener gain: $\hat{s}_k(l) = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_N^2} y_k(l)$
- $\hat{s}_k(l) = E[S|y] = g\left(\sigma_N^2, \sigma_S^2, y, \nu, \gamma\right) y_k(l)$
- power spectral subtraction
Overview of single-channel NR algorithm

Target (speech) PSD Estimator:
- Maximum likelihood (based on Bartlett estimate)
- Decision-directed approach
Overview of single-channel NR algorithm

Noise PSD estimator:

- Voice activity detector
- Minimum statistics
- MMSE based with speech presence uncertainty.