SPEECH PROCESSING
Speech Enhancement

Patrick Naylor
Spring Term 2018-19

Imperial College London

Aims of this module

- Provide an overview of single-microphone noise reduction methods.
- Formulate the noise reduction problem.
- Define and investigate performance measures.
- Derive an optimal filter in the time-domain.
- Investigate spectral enhancement methods:
  - Spectral subtraction
  - Statistical models

Imperial College London

Applications

There is a variety of applications for speech enhancement algorithms:

Automotive	Hands-free kits, sound reinforcement system.
Health	Hearing aids, home-care.
Home/Office	Speech recognition, speaker identification, internet telephony, teleconferencing, set-top boxes, home automation.
Mobile	Mobile phones, smartphones, personal digital assistants, mobile multimedia devices.

Hands-free solutions…

?

Problem Formulation (time-domain)

The microphone signal $y(n)$ can be expressed as:

$$y(n) = f[s(n)] + v(n)$$

where $s(n)$ is the desired signal and $v(n)$ is the undesired signal. It is assumed that the desired and undesired signals are mutually independent.

Two types of distortions:

1. Linear or non-linear distortions of the desired signal $s(n)$.
2. Additive distortions such as $v(n)$.

Here we focus on additive distortions:

$$y(n) = s(n) + v(n)$$

Problem Formulation (time-domain)

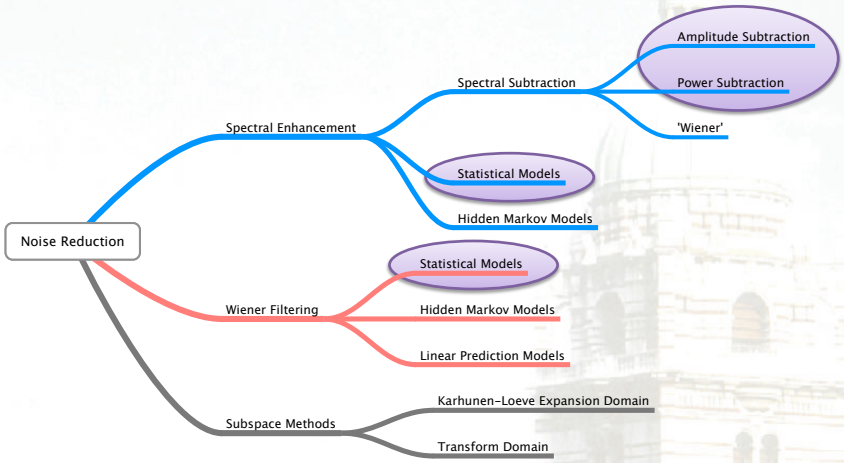
Our aim is to estimate the desired signal $s(n)$. Alternatively, we can suppress the undesired signal $v(n)$ while minimizing the distortion on $s(n)$.

Consider filtering $y(n)$ by the impulse response of a filter $H(z)$. We denote the enhanced signal by $z(n)$.

Sample-based:
$$\begin{aligned} z(n) &= \mathbf{h}^T\mathbf{y}(n) \\ &= \mathbf{h}^T[\mathbf{s}(n) + \mathbf{v}(n)] \\ &= s_{\mathrm{F}}(n) + v_{\mathrm{F}}(n) \end{aligned}$$

Frame-based:
$$\begin{aligned} \mathbf{z}(n) &= \mathbf{H}\mathbf{y}(n) \\ &= \mathbf{H}[\mathbf{s}(n) + \mathbf{v}(n)] \\ &= \mathbf{s}_{\mathrm{F}}(n) + \mathbf{v}_{\mathrm{F}}(n) \end{aligned}$$

Overview of Noise Reduction Methods



Single-Microphone Speech Enhancement

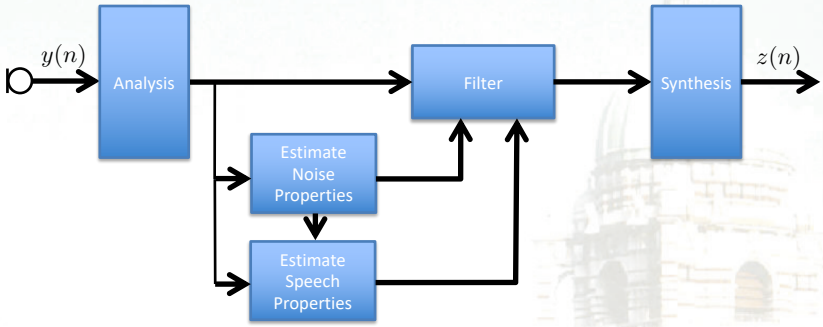


Figure: Block diagram of a single-microphone noise reduction system.

Performance Measures (time-domain)

Input signal-to-noise ratio (iSNR):

$$\mathbf{R}_s = E\left[\mathbf{s}(n)\mathbf{s}^T(n)\right]$$

$$\text{iSNR} = \frac{E\left[\mathbf{s}^T(n)\mathbf{s}(n)\right]}{E\left[\mathbf{v}^T(n)\mathbf{v}(n)\right]} = \frac{\sigma_s^2}{\sigma_v^2} = \frac{\text{tr}(\mathbf{R}_s)}{\text{tr}(\mathbf{R}_v)}$$

Output signal-to-noise ratio (oSNR):

$$\text{oSNR}(\mathbf{H}) = \frac{E\left[\mathbf{s}_{\text{F}}^T(n)\mathbf{s}_{\text{F}}(n)\right]}{E\left[\mathbf{v}_{\text{F}}^T(n)\mathbf{v}_{\text{F}}(n)\right]} = \frac{\text{tr}\left(\mathbf{H}\mathbf{R}_s\mathbf{H}^T\right)}{\text{tr}\left(\mathbf{H}\mathbf{R}_v\mathbf{H}^T\right)}$$

Noise reduction factor:

$$\xi_{\text{nr}}(\mathbf{H}) = \frac{E\left[\mathbf{v}^T(n)\mathbf{v}(n)\right]}{E\left[\mathbf{v}_{\text{F}}^T(n)\mathbf{v}_{\text{F}}(n)\right]} = \frac{\text{tr}\left(\mathbf{R}_v\right)}{\text{tr}\left(\mathbf{H}\mathbf{R}_v\mathbf{H}^T\right)}$$

Speech distortion index:

$$\nu_{\text{sd}}(\mathbf{H}) = \frac{E\left\{\left[\mathbf{s}_{\text{F}}(n) - \mathbf{s}(n)\right]^T\left[\mathbf{s}_{\text{F}}(n) - \mathbf{s}(n)\right]\right\}}{E\left[\mathbf{s}^T(n)\mathbf{s}(n)\right]}$$

Optimal Filtering in the Time-Domain

The error signal vector is defined by

$$\begin{aligned}\mathbf{e}(n) &= \mathbf{z}(n) - \mathbf{s}(n) \\ &= \mathbf{H}(n)[\mathbf{s}(n) + \mathbf{v}(n)] - \mathbf{s}(n)\end{aligned}$$

Now we can define the mean-squared error (MSE) criterion:

$$J(\mathbf{H}) = \text{tr}\left\{E\left[\mathbf{e}(n)\mathbf{e}^T(n)\right]\right\}$$

Derivation of the Wiener Filter

$$
\begin{aligned}
J(\mathbf{H}) &= \mathrm{tr}\left\{E\left[\mathbf{e}(n)\mathbf{e}^T(n)\right]\right\} \\
&= \mathrm{tr}\left\{E\left[(\mathbf{z}-\mathbf{s})(\mathbf{z}-\mathbf{s})^T\right]\right\} \\
&= \mathrm{tr}\left\{E\left[\mathbf{z}\mathbf{z}^T - \mathbf{z}\mathbf{s}^T - \mathbf{s}\mathbf{z}^T + \mathbf{s}\mathbf{s}^T\right]\right\} \\
&= \mathrm{tr}\left\{E[\mathbf{z}\mathbf{z}^T] - E[\mathbf{z}\mathbf{s}^T] - E[\mathbf{s}\mathbf{z}^T] + E[\mathbf{s}\mathbf{s}^T]\right\}
\end{aligned}
$$

$$E[\mathbf{s}\mathbf{s}^T] = \mathbf{R}_s$$

$$
\begin{aligned}
E[\mathbf{z}\mathbf{s}^T] &= E[(\mathbf{H}\mathbf{y})\mathbf{s}^T] \\
&= E[(\mathbf{H}(\mathbf{s}+\mathbf{v}))\mathbf{s}^T] \\
&= E[\mathbf{H}\mathbf{s}\mathbf{s}^T] + E[\mathbf{H}\mathbf{v}\mathbf{s}^T] \\
&= \mathbf{H}E[\mathbf{s}\mathbf{s}^T] + \mathbf{H}E[\mathbf{v}\mathbf{s}^T] \\
&= \mathbf{H}\mathbf{R}_s
\end{aligned}
$$

$$
\begin{aligned}
E[\mathbf{z}\mathbf{z}^T] &= E[(\mathbf{H}\mathbf{y})(\mathbf{H}\mathbf{y})^T] \\
&= E[\mathbf{H}\mathbf{y}\mathbf{y}^T\mathbf{H}^T] \\
&= \mathbf{H}E[\mathbf{y}\mathbf{y}^T]\mathbf{H}^T \\
&= \mathbf{H}E[(\mathbf{s}+\mathbf{v})(\mathbf{s}+\mathbf{v})^T]\mathbf{H}^T \\
&= \mathbf{H}E[\mathbf{s}\mathbf{s}^T]\mathbf{H}^T + \mathbf{H}E[\mathbf{s}\mathbf{v}^T]\mathbf{H}^T \\
&+ \mathbf{H}E[\mathbf{v}\mathbf{s}^T]\mathbf{H}^T + \mathbf{H}E[\mathbf{v}\mathbf{v}^T]\mathbf{H}^T \\
&= \mathbf{H}\mathbf{R}_s\mathbf{H}^T + \mathbf{H}\mathbf{R}_v\mathbf{H}^T
\end{aligned}
$$

$$J(\mathbf{H}) = \mathrm{tr}\left\{\mathbf{H}\mathbf{R}_s\mathbf{H}^T + \mathbf{H}\mathbf{R}_v\mathbf{H}^T - 2\mathbf{H}\mathbf{R}_s + \mathbf{R}_s\right\}$$

Min J is found by setting $\dfrac{\partial J}{\partial \mathbf{H}} = 0$

$$\mathrm{tr}(-2\mathbf{R}_s + 2\mathbf{H}(\mathbf{R}_s + \mathbf{R}_v) = 0$$

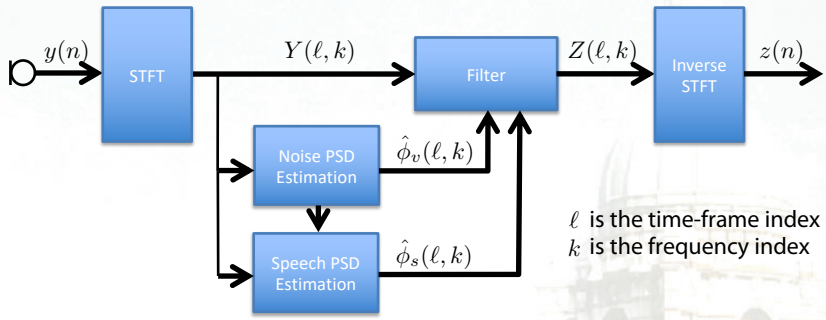
So the Wiener filter $\mathbf{H}_w$ is found from

$$\mathbf{R}_s = \mathbf{H}_w(\mathbf{R}_s + \mathbf{R}_v)$$

giving

$$\mathbf{H}_{\mathrm{W}}^T = \mathbf{R}_y^{-1}\mathbf{R}_s = \mathbf{I} - \mathbf{R}_y^{-1}\mathbf{R}_v$$

Spectral Enhancement Methods



Microphone signal: $Y(\ell,k) = S(\ell,k) + V(\ell,k)$

Spectrum: $B(\ell,k) = A_b(\ell,k)\,\mathrm{e}^{j\varphi_b(\ell,k)}$

Spectral variance:

$$\begin{aligned}\phi_b(\ell,k) &= E\left[|B(\ell,k)|^2\right] \\ &= E\left[A_b^2(\ell,k)\right]\end{aligned}$$

Short-Time Fourier Transform (recap)

Analysis:

$$A(\ell,k) = \sum_{n=0}^{K-1} a(n+\ell R)w(n)\mathrm{e}^{-j\frac{2\pi}{K}nk},$$

where $w(n)$ is the analysis window (e.g., Hamming window).

Synthesis:

$$a(n) = \sum_{\ell}\sum_{k=0}^{K-1} A(\ell,k)\tilde{w}(n-\ell R)\mathrm{e}^{j\frac{2\pi}{K}k(n-\ell R)},$$

where $\tilde{w}(n)$ is the synthesis window.

The synthesis can be implemented efficiently using the *overlap and add method*.

Spectral Subtraction Methods

Amplitude Subtraction:

$$A_z(\ell,k) = A_y(\ell,k) - \sqrt{\hat{\phi}_v(\ell,k)}$$

If we use the noisy phase spectrum, we obtain:

$$Z(\ell,k) = H(\ell,k)Y(\ell,k) \text{ with } H(\ell,k) = 1 - \sqrt{\frac{\hat{\phi}_v(\ell,k)}{|Y(\ell,k)|^2}}$$

Power Subtraction:

$$A_z^2(\ell,k) = A_y^2(\ell,k) - \hat{\phi}_v(\ell,k)$$

If we use the noisy phase spectrum, we obtain:

$$Z(\ell,k) = H(\ell,k)Y(\ell,k) \text{ with } H(\ell,k) = \sqrt{1 - \frac{\hat{\phi}_v(\ell,k)}{|Y(\ell,k)|^2}}$$

Example

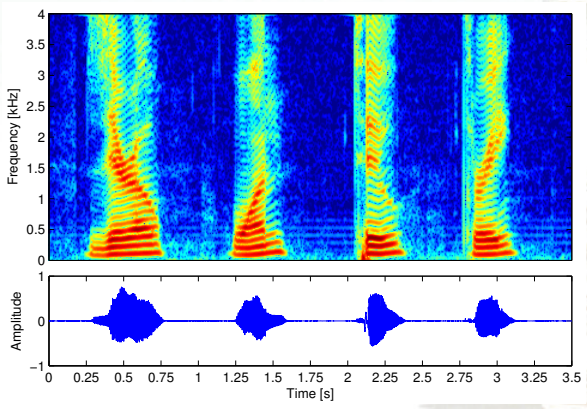


Figure: Spectrogram and waveform of a desired signal.

Example

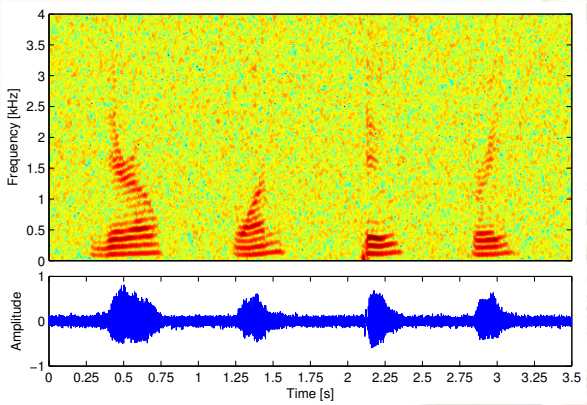


Figure: Spectrogram and waveform of the microphone signal (iSNR = 5 dB).

Example – Amplitude Subtraction

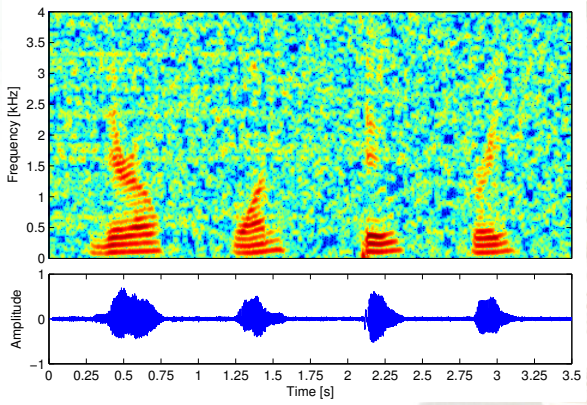


Figure: Spectrogram and waveform of the processed signal.

Statistical-Model based Methods

Based on different (i) fidelity criteria, (ii) distortion measures and (iii) statistical models.

$$Z(\ell,k) = \min_{Z(\ell,k)} E\left\{ d[S(\ell,k), Z(\ell,k)] | \hat{\phi}_s(\ell,k), \hat{\phi}_v(\ell,k), Y(\ell,k) \right\}$$

Most commonly used distortion measure is the Mean Square Error

$$d[S(\ell,k), Z(\ell,k)] = |g[Z(\ell,k)] - g[S(\ell,k)]|^2$$

Subtraction is performed in a chosen 'domain': complex, abs, log, etc

where $g(X)$ is a function of X such as $X, |X|, \log(|X|)$

$$E\{X\} = \int_{-\infty}^{\infty} x\, p(x)\, dx$$

The estimator for $Z(\ell,k)$ is then given by:

$$g[Z(\ell,k)] = E\left\{ g[S(\ell,k)] | \hat{\phi}_s(\ell,k), \hat{\phi}_v(\ell,k), Y(\ell,k) \right\}$$

- meaning that the estimator operates in the chosen 'domain'.

Statistical-Model based Methods

The desired and undesired spectral coefficients are assumed zero-mean mutually independent Gaussian random variables.

Short-term spectral amplitude (SA): $g(X) = |X|$

$$E\{A_s(\ell,k) | \phi_s(\ell,k), \phi_v(\ell,k), Y(\ell,k)\} = H_{\mathrm{SA}}(\ell,k) A_y(\ell,k)$$

$$H_{\mathrm{SA}}(\ell,k) = \frac{1}{A_y(\ell,k)} \int A_s(\ell,k)\, p[S(\ell,k) | \phi_s(\ell,k), \phi_v(\ell,k)]\, dS$$

$$= \frac{\sqrt{\pi}}{2} \frac{\sqrt{\zeta(\ell,k)}\phi_v(\ell,k)}{|Y(\ell,k)|^2} \exp\left[-\frac{\zeta(\ell,k)}{2} \right]$$

$$\times \left\{ [1+\zeta(\ell,k)]\, I_0\left[\frac{\zeta(\ell,k)}{2} \right] + \zeta(\ell,k)\, I_1\left[\frac{\zeta(\ell,k)}{2} \right] \right\}$$

Modified Bessel functions

iSNR is also known as the a priori SNR

$$\zeta(\ell,k) = \frac{\mathrm{iSNR}(\ell,k)}{1+\mathrm{iSNR}(\ell,k)} \frac{|Y(\ell,k)|^2}{\phi_v(\ell,k)}$$

A posteriori SNR

Statistical-Model based Methods

Log-spectral amplitude (LSA) [Ephraim and Malah,1985]:

$$g(X) = \log(|X|)$$

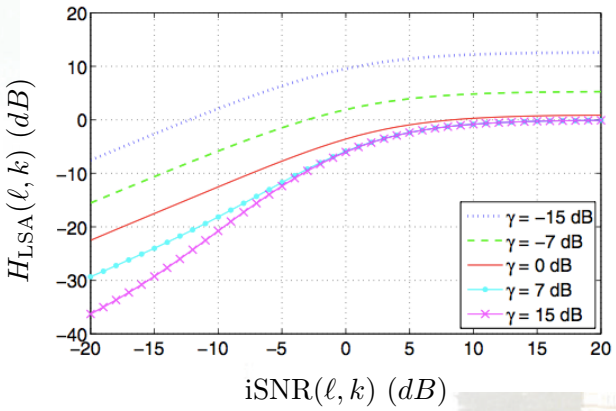
$$E\left\{\log[A_s(\ell,k)]|\phi_s(\ell,k),\phi_v(\ell,k),Y(\ell,k)\right\} = \log[H_{\mathrm{LSA}}(\ell,k)A_y(\ell,k)]$$

$$\begin{aligned} H_{\mathrm{LSA}}(\ell,k) &= \frac{1}{A_y(\ell,k)}\exp\left\{\int \log[A_s(\ell,k)]\, p[S(\ell,k)|\phi_s(\ell,k),\phi_v(\ell,k)]\, dS\right\} \\ &= \frac{\mathrm{iSNR}(\ell,k)}{1+\mathrm{iSNR}(\ell,k)}\exp\left(\frac{1}{2}\int_{\zeta(\ell,k)}^{\infty}\frac{e^{-x}}{x}\,\mathrm{d}x\right) \end{aligned}$$

$$\zeta(\ell,k) = \frac{\mathrm{iSNR}(\ell,k)}{1+\mathrm{iSNR}(\ell,k)}\,\frac{|Y(\ell,k)|^2}{\phi_v(\ell,k)}$$

See Eq. 20 of the above reference.

Log-Spectral Amplitude function



$$\gamma(\ell,k) = \frac{|Y(\ell,k)|^2}{\phi_v(\ell,k)}$$

(a posteriori SNR)

Speech Spectral Variance Estimation

To estimate the iSNR we require an estimate of the speech and noise spectral variances.

Decision-directed approach [Ephraim and Malah 1984]:

$$\hat{\phi}_s(\ell,k) = \alpha\hat{\phi}_s(\ell-1,k) + (1-\alpha)\max\left\{A_y^2(\ell,k) - \phi_v(\ell,k), \beta\phi_v(\ell,k)\right\}$$

The parameter $(0 \leq \alpha < 1)$ controls the tradeoff between the noise reduction and transient distortion introduced into the desired speech signal. Note: This also reduces musical tones!

Noise Spectral Variance Estimation

Estimating the spectral variance of the undesired signal component:

1. Voice activity detection (VAD)
2. Minimum statistics approach

Voice activity detection:

$$\hat{\phi}_v(\ell,k) = \begin{cases} \hat{\phi}_v(\ell-1,k) & \text{if speech active} \\ \alpha\,\hat{\phi}_v(\ell-1,k) + (1-\alpha)\,|Y(\ell,k)|^2 & \text{otherwise} \end{cases}$$

Questions:

- What happens when the noise spectral variance is non-stationary?
- What happens when the VAD makes a wrong decision?

Minimum Statistics Approach [Martin, 2001]

This technique is based on the assumption that during a speech pause, or within brief periods between words and even syllables, the speech energy is close to zero. As a result, a short-term power spectrum estimate of the noisy signal, even during speech activity, decays frequently to the noise power. Thus, by tracking the temporal spectral minimum without distinguishing between speech presence and speech absence, the noise power in a specific frequency band can be estimated.

"Naïve" approach:

$$\hat{\phi}_y(\ell,k) = \alpha\,\hat{\phi}_y(\ell-1,k) + (1-\alpha)\,|Y(\ell,k)|^2$$

$$\hat{\phi}_v(\ell,k) = \min\left\{\hat{\phi}_y(\ell,k), \hat{\phi}_y(\ell-1,k), \ldots, \hat{\phi}_y(\ell-D+1,k)\right\}$$

Minimum Statistics Approach [Martin, 2001]

Problems with the aforementioned naïve approach:

- Heavy smoothing reduces the variance but widens peaks during speech activity. This will lead to inaccurate noise estimates as the sliding window for the minimum search might slip into broad peaks.
- In case of increasing noise power, the minimum tracking lags behind.
- The noise estimate is biased toward lower values.

Improved approach:

$$\hat{\phi}_y(\ell,k) = \alpha(\ell,k)\,\hat{\phi}_y(\ell-1,k) + [1-\alpha(\ell,k)]\,|Y(\ell,k)|^2$$

$$P_{\min}(\ell,k) = \min\left\{\hat{\phi}_y(\ell,k), \hat{\phi}_y(\ell-1,k), \ldots, \hat{\phi}_y(\ell-D+1,k)\right\}$$

$$\hat{\phi}_v(\ell,k) = \frac{P_{\min}(\ell,k)}{E[P_{\min}(\ell,k)]_{\hat{\phi}_v(\ell,k)=1}}$$

Example – MMSE Log-Spectral Amplitude

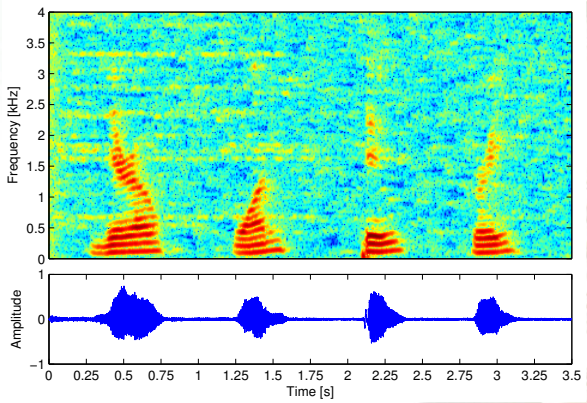


Figure: Spectrogram and waveform of the processed signal.

Control the noise suppression…

Random narrowband fluctuations in the residual noise, which are known as musical tones, are annoying and disturbing to the perception of the enhanced signal.

We can put a lower-bound on the filter to minimize the noise suppression and decrease speech distortion:

$$H(\ell, k) = \max(H(\ell, k), H_{\min})$$

Example – Amplitude Subtraction

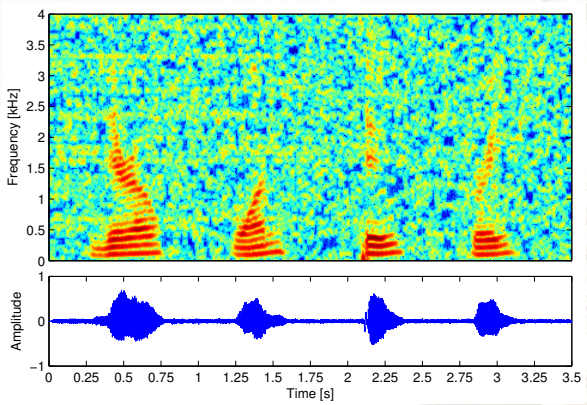


Figure: Spectrogram and waveform of the processed signal.

Example – Modified Amplitude Subtraction

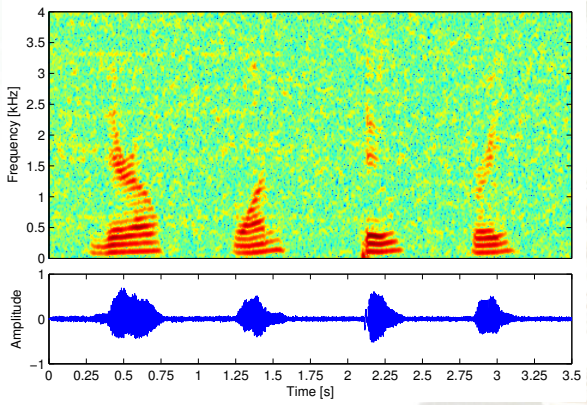


Figure: Spectrogram and waveform of the processed signal using a lower-bound on the filter (-12 dB).

Summary

- We formulated the single-microphone noise reduction problem.
- An overview of different methods was provided.
- We derived the Wiener filter in the time-domain.
- Spectral subtraction approaches have a very low computational complexity and only require an estimate of the noise PSD.
- A huge variety of spectral enhancement approaches exists. They are based on different fidelity criteria, distortion measures, and statistical models.
- In general, for single-microphone noise reduction methods there is a tradeoff between speech-distortion and noise reduction.
  Note: This is not true for multi-microphone noise reduction methods.