Speech Modeling and Enhancement in Nonstationary Noise Environments Part I

Prof. Israel Cohen and Prof. Sharon Gannot

Elect. Eng. Dept., Technion - Israel Inst. of Tech., Israel
School of Engineering, Bar-Ilan University, Israel

SIPA 2011

Outline

Hands-free communication systems

Enhancement of speech signals is of great interest in many hands-free communication systems:

- Hearing-aids devices.
- Cell phones and hands-free accessories for wireless communication systems.
- Conference and telephone speakerphones.
- Etc.

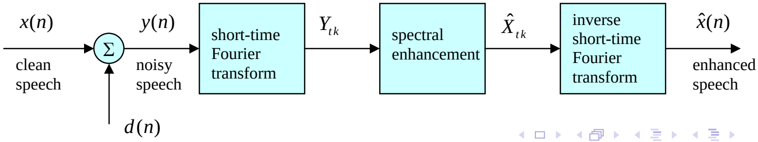
Spectral Enhancement

The observed signal $y(n) = x(n) + d(n)$ is transformed into the time-frequency domain:

$$Y_{tk} = \sum_{n=0}^{N-1} y(n + tM)\, h(n)\, e^{-j\frac{2\pi}{N}nk} \,.$$

$\hat{X}_{tk}$ is computed from $\hat{Y}_{tk}$.
$\hat{x}(n)$ is the inverse STFT of $\hat{X}_{tk}$

$$\hat{x}(n) = \sum_{t} \sum_{k=0}^{N-1} \hat{X}_{tk}\, \tilde{h}(n - tM)\, e^{j\frac{2\pi}{N}k(n-tM)} \,.$$



Spectral Subtraction

Boll, 1979; Berouti, Schwartz and Makhoul, 1978

Let the observed signal be:

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

where $x(n)$ is the clean speech signal and $d(n)$ is the noise signal. The noisy signal in the STFT domain is therefore:

$$Y_{tk} = X_{tk} + D_{tk}.$$

The short-term power spectrum is given by:

$$|Y_{tk}|^2 = |X_{tk}|^2 + |D_{tk}|^2 + 2\Re\{X_{tk} D_{tk}^*\}.$$

Spectral Subtraction (cont.)

- Cross-term is approaching zero.
- Estimated noise power $\widehat{\sigma_k^2} \approx \mathrm{mean}\{|D_{tk}|^2\}$ in noise-only segments.
- Spectral subtraction

$$|\hat{X}_{tk}|^2 \approx \begin{cases} |Y_{tk}|^2 - \widehat{\sigma_k^2} & \text{if } |Y_{tk}|^2 > \widehat{\sigma_k^2} \\ 0 & \text{otherwise} \end{cases} .$$

- Use noisy phase to obtain

$$\hat{X}_{tk} = |\hat{X}_{tk}| e^{\angle Y_{tk}}$$

- Since the STFT phase is not estimated, the theoretical limit in estimating the original STFT by this approach is

$$\hat{X}_{tk} = |X_{tk}| e^{\angle Y_{tk}}$$

- STFT phase estimation is a more difficult problem than STFT magnitude estimation.
- This is in part due to the difficulty in characterizing phase in low-energy regions of the spectrum, and in part due to the use of only second-order statistical averages.
- Generally, speech degradation is not perceived in the theoretical limit for

$$\mathrm{SegSNR} > 6\mathrm{dB}$$

- However, for SegSNR considerably below 6 dB, a roughness of the reconstruction is perceived.

Musical noise

The half-wave rectification and the difference between the estimated noise level and the current noise spectrum cause an audible artifact, known as musical noise. The noise is perceived as tones with random frequencies that change from frame to frame.

Spectral floor (Berouti et al., 1978)

$$|\hat{X}_{tk}|^2 \approx \begin{cases} |Y_{tk}|^2 - \alpha\widehat{\sigma_k^2} & \text{if } |Y_{tk}|^2 > (\alpha+\beta)\widehat{\sigma_k^2} \\ \beta\widehat{\sigma_k^2} & \text{otherwise} \end{cases} .$$

- $\alpha > 1$ - over-subtraction factor, reducing wideband residual noise.
- $0 < \beta \ll 1$ - spectral floor parameter, masking narrowband residual noise (musical noise).

Wiener Filtering

- An alternative to spectral subtraction is to find a linear filter h(n) such that the sequence

$$\hat{x}(n) = y(n) * h(n)$$

minimizes the mean-squared error (MMSE)

$$\min_h E\left\{(\hat{x}(n) - x(n))^2\right\} = \min_h E\left\{(h(n) * [x(n] + d(n)] - x(n))^2\right\}$$

- In the STFT domain we have

$$\min_{H_{tk}} E\left|(H_{tk} - 1)X_{tk} + H_{tk}D_{tk}\right|^2 = \min_{H_{tk}} \left\{(H_{tk} - 1)^2 \Phi_{tk}^{xx} + H_{tk}^2 \Phi_{tk}^{dd}\right\}$$

where $\Phi_{tk}^{xx} = E\{|X_{tk}|^2\}$ and $\Phi_{tk}^{dd} = E\{|D_{tk}|^2\}$ are the time-varying power spectra of the desired signal and the background noise, respectively.

Wiener Filtering (cont.)

- Therefore, the time-varying Wiener filter is given by

$$H_{tk} = \frac{\Phi_{tk}^{xx}}{\Phi_{tk}^{xx} + \Phi_{tk}^{dd}}$$

- Define the *a priori* SNR $\xi_{tk} = \frac{\Phi_{tk}^{xx}}{\Phi_{tk}^{dd}} \Rightarrow H_{tk} = \frac{1}{1+\frac{1}{\xi_{tk}}}$.
- Direct estimation of the a priori SNR will be addressed later.
- The Wiener filter does not invoke an absolute thresholding as spectral subtraction.

Wiener Filtering (cont.)

Application of the Wiener filter

- Estimate $\Phi_{tk}^{dd} \approx \widehat{\sigma_k^2}$ from noise-only segments, assuming stationarity.
- Estimate Φ_{tk}^{xx} by using the already enhanced segment:
  1. $$\hat{X}_{tk} = H_{t-1,k} Y_{tk}$$
  2. $$\hat{\Phi}_{tk}^{xx} = \eta \hat{\Phi}_{t-1,k}^{xx} + (1-\eta)|\hat{X}_{tk}|^2$$

  where η $(0 < \eta < 1)$ is a smoothing constant.

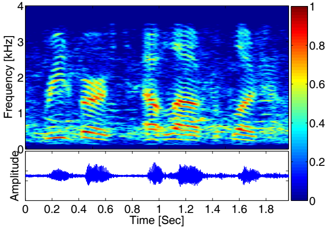
Wiener Filtering (cont.)

- The smoothing constant controls how fast we adapt to a nonstationary object spectrum.
- A fast adaptation, with a small smoothing constant, implies improved time resolution, but more noise in the spectral estimate, and thus more musicality in the synthesis.
- A large smoothing constant improves the spectral estimate in regions of stationarity, but it smears onsets and other rapid events.

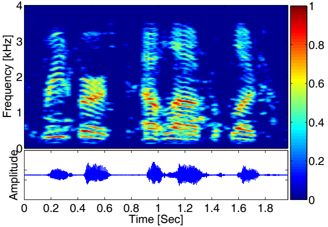
Experimental Results

Babble Noise, SNR=10dB, NOIZEUS database, Speech Enhancement, P. Loizou, 2007

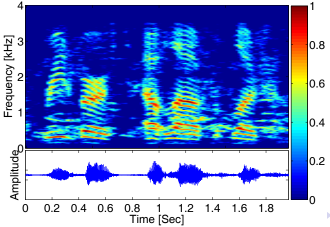
Noisy signal,



SSUB

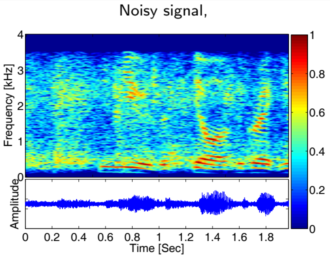


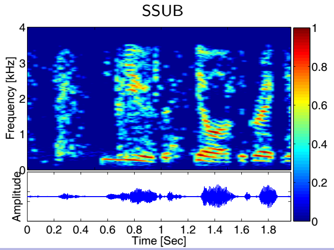
Wiener

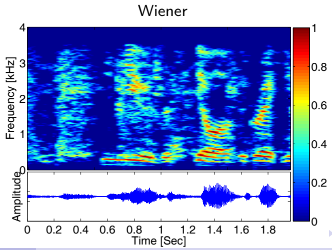


Experimental Results (cont.)

Train Noise, SNR=5dB, NOIZEUS database, Speech Enhancement, P. Loizou, 2007



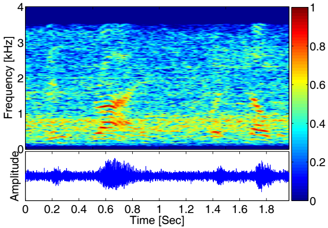




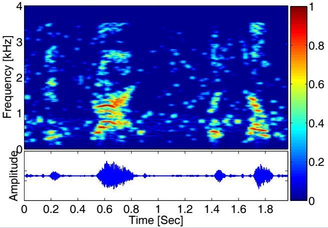
Experimental Results (cont.)

Car Noise, SNR=0dB, NOIZEUS database, Speech Enhancement, P. Loizou, 2007

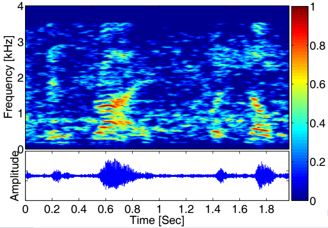
Noisy signal,



SSUB



Wiener



General Problem Formulation

Let $\{Y_{tk}\}$ denote a noisy speech signal in the STFT domain:

$$\begin{aligned} H_1^{tk}\text{ (speech present)}: \quad Y_{tk} &= X_{tk} + D_{tk} \\ H_0^{tk}\text{ (speech absent)}: \quad Y_{tk} &= D_{tk}\,. \end{aligned}$$

The spectral enhancement problem can be formulated as

$$\min_{\hat{X}_{tk}} E\left\{ d\left(X_{tk}, \hat{X}_{tk}\right) \,\middle|\, \hat{p}_{tk}\,, \hat{\lambda}_{tk}\,, \widehat{\sigma^2_{tk}}\,, \; Y_{tk} \right\}$$

- $d\left(X_{tk}, \hat{X}_{tk}\right)$ - distortion measure between X_{tk} and $\hat{X}_{tk}$
- $\hat{p}_{tk} = P\left(H_1^{tk} \mid \psi_t\right)$ - speech presence probability estimate
- $\hat{\lambda}_{tk} = E\left\{|X_{tk}|^2 \mid H_1^{tk}, \, \psi_t\right\}$ - speech spectral variance estimate
- $\widehat{\sigma^2_{tk}} = E\left\{|Y_{tk}|^2 \mid H_0^{tk}, \, \psi_t\right\}$ - noise spectral variance estimate
- ψ_t - information employed for estimation at frame t (*e.g.*, noisy data observed through time t)

Squared Error Distortion Measure

In particular, assuming a squared error distortion measure of the form

$$d\left(X_{tk}, \hat{X}_{tk}\right) = \left|g(\hat{X}_{tk}) - \tilde{g}(X_{tk})\right|^2$$

where $g(X)$ and $\tilde{g}(X)$ are specific functions of X (*e.g.*, X, $|X|$, $\log|X|$, $e^{j\angle X}$)

the estimator $\hat{X}_{tk}$ is calculated from

$$\begin{aligned} g(\hat{X}_{tk}) &= E\left\{\tilde{g}(X_{tk}) \,\middle|\, \hat{p}_{tk}\,, \hat{\lambda}_{tk}\,, \widehat{\sigma^2_{tk}}\,, Y_{tk}\right\} \\ &= \hat{p}_{tk}\, E\left\{\tilde{g}(X_{tk}) \,\middle|\, H_1^{tk}\,, \hat{\lambda}_{tk}\,, \widehat{\sigma^2_{tk}}\,, Y_{tk}\right\} \\ &\quad +(1-\hat{p}_{tk})\, E\left\{\tilde{g}(X_{tk}) \,\middle|\, H_0^{tk}\,, Y_{tk}\right\} . \end{aligned}$$

Estimator Specifications

The design of a particular estimator for X_{tk} requires the following specifications:

- Functions $g(X)$ and $\tilde{g}(X)$, which determine the fidelity criterion of the estimator.
- A conditional probability density function (pdf) $p\left(X_{tk} \mid \lambda_{tk}\,,\, H_1^{tk}\right)$ for X_{tk} under H_1^{tk} given its variance λ_{tk}, which determines the statistical model.
- An estimator $\hat{\lambda}_{tk}$ for the speech spectral variance.
- An estimator $\widehat{\sigma^2_{tk}}$ for the noise spectral variance.
- An estimator $\hat{p}_{tk|t} = P\left(H_1^{tk} \mid \psi_t\right)$ for the *a posteriori* speech presence probability, where ψ_t represents the information set known including the measurement Y_{tk}.

Fidelity Criteria

- Fidelity criteria that are of particular interest for speech enhancement applications are MMSE, MMSE of the spectral amplitude (MMSE-SA), and MMSE of the log-spectral amplitude (MMSE-LSA).
- The MMSE estimator is derived by using the functions

$$\begin{aligned} g(\hat{X}_{tk}) &= \hat{X}_{tk} \\ \tilde{g}(X_{tk}) &= \begin{cases} X_{tk}\,, & \text{under } H_1^{tk} \\ G_{\min}\, Y_{tk}\,, & \text{under } H_0^{tk} \end{cases} \end{aligned} \tag{1}$$

where $G_{\min} \ll 1$ represents a constant attenuation factor, which retains the noise naturalness during speech absence.

Fidelity Criteria (cont.)

- The MMSE-SA estimator is obtained by using the functions

$$\begin{aligned} g(\hat{X}_{tk}) &= |\hat{X}_{tk}| \\ \tilde{g}(X_{tk}) &= \begin{cases} |X_{tk}|\,, & \text{under } H_1^{tk} \\ G_{\min}|Y_{tk}|\,, & \text{under } H_0^{tk}\,. \end{cases} \end{aligned} \tag{2}$$

- The MMSE-LSA estimator is obtained by using the functions

$$\begin{aligned} g(\hat{X}_{tk}) &= \log|\hat{X}_{tk}| \\ \tilde{g}(X_{tk}) &= \begin{cases} \log|X_{tk}|\,, & \text{under } H_1^{tk} \\ \log\left(G_{\min}|Y_{tk}|\right)\,, & \text{under } H_0^{tk}\,. \end{cases} \end{aligned} \tag{3}$$

Gaussian Model

The Gaussian statistical model in the STFT domain relies on the following set of assumptions:

1. The noise spectral coefficients $\{D_{tk}\}$ are zero-mean statistically independent Gaussian random variables. The real and imaginary parts of D_{tk} are iid random variables $\sim \mathcal{N}\left(0, \frac{\sigma_{tk}^2}{2}\right)$.
2. Given $\{\lambda_{tk}\}$, the speech spectral coefficients $\{X_{tk}\}$ are zero-mean statistically independent Gaussian random variables. The real and imaginary parts of X_{tk} are iid random variables $\sim \mathcal{N}\left(0, \frac{\lambda_{tk}}{2}\right)$.

Signal Estimation

MMSE Spectral Estimation

Let

$$\xi_{tk} \triangleq \frac{\lambda_{tk}}{\sigma_{tk}^2}\,, \quad \gamma_{tk} \triangleq \frac{|Y_{tk}|^2}{\sigma_{tk}^2}\,,$$

represent the *a priori* and *a posteriori* SNRs, respectively, and let $G_{\mathrm{MSE}}(\xi,\gamma)$ denote a gain function that satisfies

$$E\left\{X_{tk} \,\middle|\, H_1^{tk}, \lambda_{tk}, \sigma_{tk}^2, Y_{tk}\right\} = G_{\mathrm{MSE}}(\xi_{tk}, \gamma_{tk})\, Y_{tk}\,.$$

Then,

$$\hat{X}_{tk} = \left[\hat{p}_{tk}\, G_{\mathrm{MSE}}\left(\hat{\xi}_{tk}, \hat{\gamma}_{tk}\right) + (1-\hat{p}_{tk})\, G_{\min}\right] Y_{tk}\,.$$

Signal Estimation (cont.)

Under a Gaussian model, the gain function is independent of the *a posteriori* SNR $\Rightarrow$ Wiener filter.

$$G_{\mathrm{MSE}}\left(\xi_{tk}\right) = \frac{\xi_{tk}}{1+\xi_{tk}} \, .$$

OM-LSA Estimation

In speech enhancement applications, estimators which minimize the MSE of the LSA have been found advantageous to MMSE spectral estimators.

let $G_{\mathrm{LSA}}\left(\xi,\,\gamma\right)$ denote a gain function that satisfies

$$\exp\left(E\left\{\log\left|X_{tk}\right| \;\middle|\; H_1^{tk}\,,\lambda_{tk}\,,\sigma_{tk}^2\,,Y_{tk}\right\}\right) = G_{\mathrm{LSA}}\left(\xi_{tk},\gamma_{tk}\right)\left|Y_{tk}\right| .$$

Signal Estimation (cont.)

Then,

$$\hat{X}_{tk} = \left[G_{\mathrm{LSA}}(\hat{\xi}_{tk}, \hat{\gamma}_{tk}) \right]^{\hat{p}_{tk}} G_{\min}^{1-\hat{p}_{tk}} Y_{tk}$$

where

$$G_{\mathrm{LSA}}(\xi, \gamma) \triangleq \frac{\xi}{1+\xi} \exp\left(\frac{1}{2} \int_{\vartheta}^{\infty} \frac{e^{-x}}{x} dx \right)$$

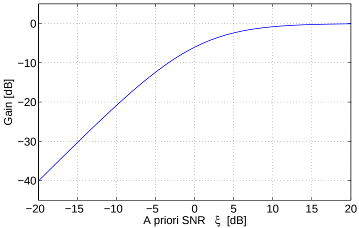
an ϑ is defined by $\vartheta \triangleq \xi \gamma / (1+\xi)$.

Similar to the MMSE spectral estimator, the OM-LSA estimator reduces to a constant attenuation of Y_{tk} when the signal is surely absent (*i.e.*, $\hat{p}_{tk} = 0$ implies $\hat{X}_{tk} = G_{\min} Y_{tk}$).

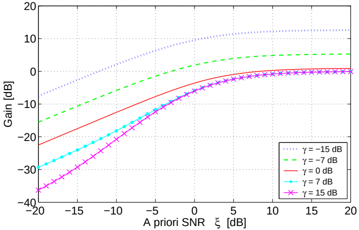
However, the characteristics of these estimators when the signal is present are readily distinctive.

Gain Function Comparison

MMSE gain function



LSA gain function



- For a fixed value of the *a posteriori* SNR γ, the LSA gain is a monotonically increasing function of ξ.
- However, for a fixed value of ξ, the LSA gain is a monotonically *decreasing* function of γ.

Gain Function Trends

- For $\gamma \gg 1$ $G_{\mathrm{LSA}}(\xi, \gamma) \to G_{\mathrm{MSE}}(\xi) = \frac{\xi}{1+\xi}$.
- For $\xi \gg 1$ and $\gamma > 0$, G_{LSA} exhibits low sensitivity to the value of γ.
- For low values of the *a priori* SNR ξ G_{LSA} is monotonically decreasing (!) as a function of the *a posteriori* SNR γ.
- For low and fixed values of ξ:
  - An instantaneous SNR (γ) increase can be attributed to noise components. The resulting lower G_{LSA} can have a positive effect on musical noise suppression.
  - Higher G_{LSA} compensates for the decrease in the instantaneous SNR γ.

Distortion measures

- Segmental SNR (SegSNR)

$$\mathrm{SegSNR} = \frac{1}{T} \sum_{t=0}^{T-1} \mathcal{C}\left(\mathrm{SNR}_t\right)$$

where

$$\mathrm{SNR}_t = 10 \log_{10} \frac{\sum_{n=tM}^{tM+N-1} x^2(n)}{\sum_{n=tM}^{tM+N-1} \left[x(n) - \hat{x}(n)\right]^2}$$

represents the SNR in the t-th frame.
The operator $\mathcal{C}$ confines the SNR at each frame to perceptually meaningful range between 35 dB and -10 dB ($\mathcal{C}x \triangleq \min[\max(x, -10), 35]$).

Distortion measures (cont.)

- Log-spectral distortion (LSD)

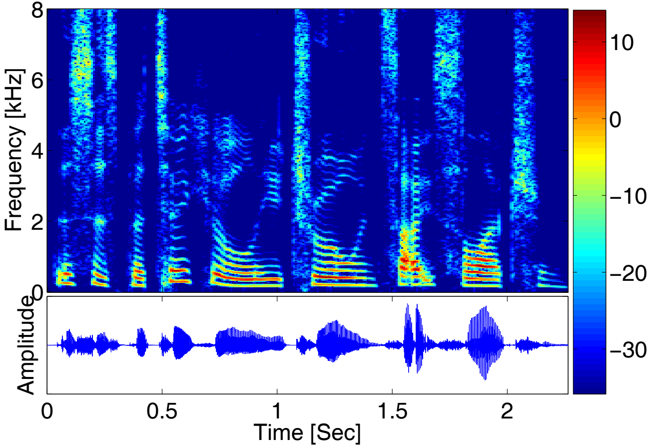
$$\mathrm{LSD} = \frac{1}{T}\sum_{t=0}^{T-1}\left[\frac{2}{N}\sum_{k=1}^{N/2}\left(\mathcal{L}X_{tk} - \mathcal{L}\hat{X}_{tk}\right)^2\right]^{\frac{1}{2}}$$

where $\mathcal{L}X_{tk} \triangleq \max\left\{20\log_{10}|X_{tk}|, \delta\right\}$ is the log spectrum confined to about 50 dB dynamic range (that is, $\delta = \max_{tk}\left\{20\log_{10}|X_{tk}|\right\} - 50$).

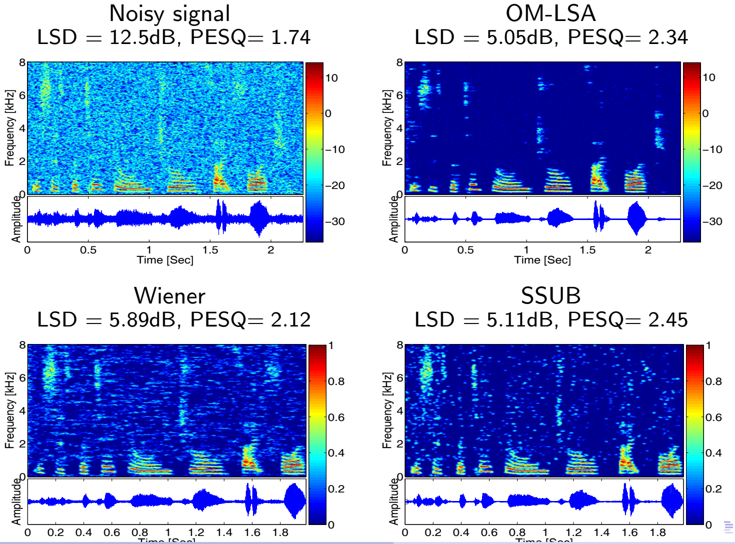
- Perceptual evaluation of speech quality (PESQ) score (ITU-T P.862).

Experimental Results - Clean Signal

"This is particularly true in site selection"

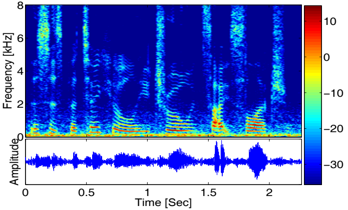


Experimental Results - White Gaussian Noise

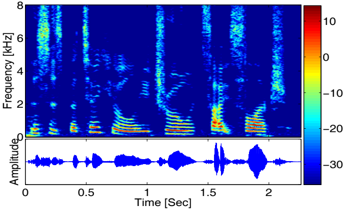


Experimental Results - Car Interior Noise

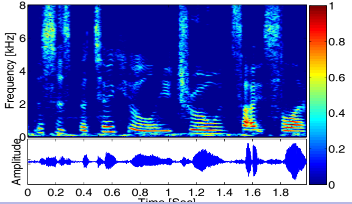
Noisy signal
LSD = 3.17dB, PESQ= 2.47



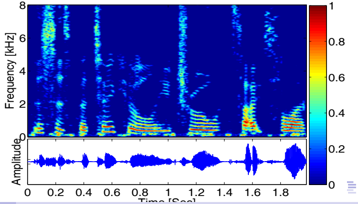
OM-LSA
LSD = 2.67dB, PESQ= 3.00



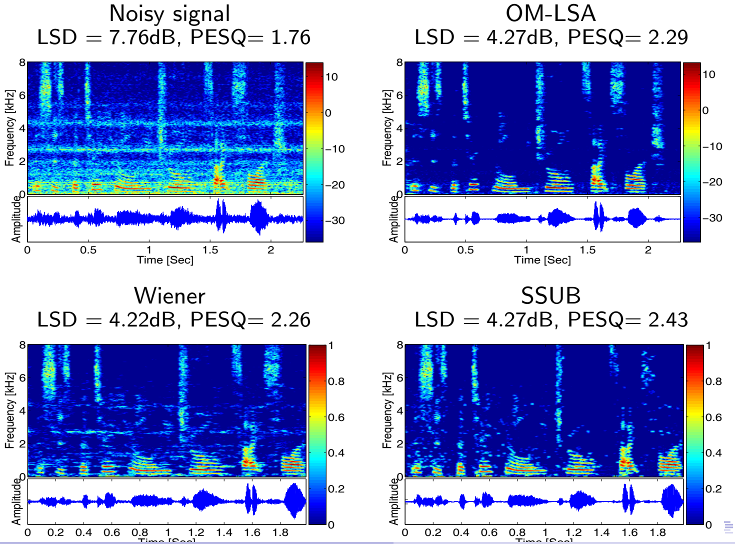
Wiener
LSD = 2.60dB, PESQ= 2.86



SSUB
LSD = 3.21dB, PESQ= 2.76

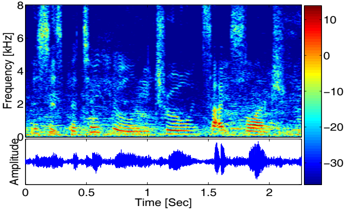


Experimental Results - F16 Cockpit Noise

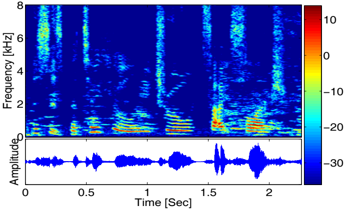


Experimental Results - Babble Noise

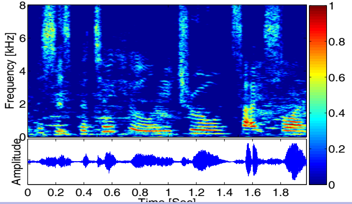
Noisy signal
LSD = 5.64dB, PESQ= 1.87



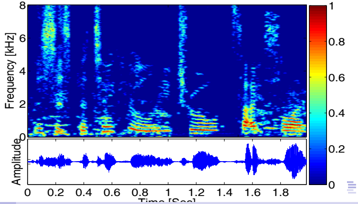
OM-LSA
LSD = 4.20dB, PESQ= 2.13



Wiener
LSD = 4.10dB, PESQ= 2.08



SSUB
LSD = 4.32dB, PESQ= 2.06



Conclusions

- The OM-LSA gain function is obtained by modifying the gain function of the conventional LSA estimator.
- The modification includes:
  - A lower bound for the gain (determined by a subjective criteria for the noise naturalness)
  - Exponential weights (conditional speech presence probability)
  - Improved a priori SNR estimate (under speech presence uncertainty)
- The OM-LSA demonstrates improved noise suppression, while retaining weak speech components and avoiding the musical residual noise phenomena.
- A free MATLAB code is available on: http://www.ee.technion.ac.il/people/IsraelCohen/

Alternative Approaches

- Model based:
  - Speech modeled as an Autoregressive (AR) process:
    - Iterative procedure (EM procedure).
    - Frequency-domain using Wiener filter (Lim, Oppenheim, 1978).
    - Time-domain using Kalman filter (Gannot, Burshtein, Weinstein, 1998).
  - GARCH model (Cohen, 2004).
- Subspace methods (Ephraim, Van Trees, 1995; Hu, Loizou, 2003):
  - Clean speech is confined to a subspace of the noisy Euclidean space.
  - Use methods from Linear Algebra (EVD, SVD or Karhunen-Loève transform) to project the noisy signal onto the "clean" subspace.
- Codebook based (Burshtein, Gannot, 2001):
  - Use training data for clean speech signals.
  - Use GMM to model log-spectrum of clean speech.
  - Approximate addition in linear domain by maximization in log-spectrum domain.