Part I: Background
Traditional Speech Enhancement

Israel Cohen[1], Sharon Gannot[2] and Ronen Talmon[3]

[1]Elect. Eng. Dept., Technion - Israel Inst. of Tech., Israel
[2]Faculty of Engineering, Bar-Ilan University, Israel
[3]Mathematics Department, Yale University, CT

ICASSP 2012

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.

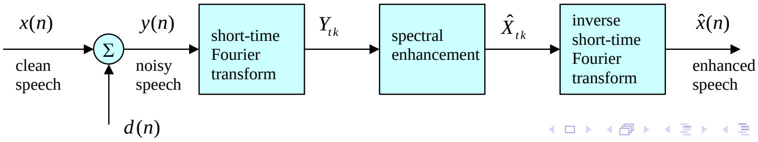
Spectral Enhancement

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).

General Problem Formulation

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

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 $\widehat{\sigma_{tk}^2}$ 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

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

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,

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

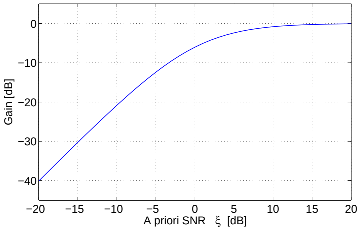
where

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

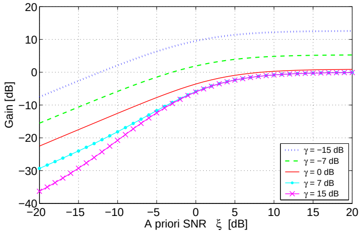
an ϑ is defined by $\vartheta \triangleq \xi \, \gamma / \left(1+\xi\right)$.
Similar to the MMSE spectral estimator, the OM-LSA estimator reduces to a constant attenuation of Y_{tk} when the signal is surely absent (*i.e.*, $\hat{p}_{tk} = 0$ implies $\hat{X}_{tk} = G_{\min} \, Y_{tk}$).
However, the 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 $\gamma \gg 1$ $G_{\mathrm{LSA}}(\xi, \gamma) \rightarrow G_{\mathrm{MSE}}(\xi) = \frac{\xi}{1+\xi}$.
- 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.

Noise Spectrum Estimation

Minima Controlled Recursive Averaging (MCRA)

- A common noise estimation technique is to recursively average past spectral power values of the noisy measurement during periods of speech absence:

$$H_0^{tk} : \; \bar{\sigma}^2_{t+1,k} = \alpha_d \, \bar{\sigma}^2_{tk} + (1-\alpha_d) \, |Y_{tk}|^2$$
$$H_1^{tk} : \; \bar{\sigma}^2_{t+1,k} = \bar{\sigma}^2_{tk}$$

  where α_d $(0 < \alpha_d < 1)$ denotes a smoothing parameter.
- Under speech presence uncertainty

$$\bar{\sigma}^2_{t+1,k} = \tilde{p}_{tk} \, \bar{\sigma}^2_{tk} + (1-\tilde{p}_{tk}) \left[\alpha_d \, \bar{\sigma}^2_{tk} + (1-\alpha_d) \, |Y_{tk}|^2 \right]$$

  where $\tilde{p}_{tk}$ is an estimator for the conditional speech presence probability $p_{tk} = P\left(H_1^{tk} \mid Y_{tk}\right)$.

- Equivalently

$$\bar{\sigma}^2_{t+1,k} = \tilde{\alpha}_{tk}\,\bar{\sigma}^2_{tk} + (1-\tilde{\alpha}_{tk})\,|Y_{tk}|^2$$

  where

$$\tilde{\alpha}_{tk} \stackrel{\triangle}{=} \alpha_d + (1-\alpha_d)\,\tilde{p}_{tk}$$

  is a time-varying frequency-dependent smoothing parameter, adjusted by the speech presence probability.

- Deciding speech is absent (H_0) when speech is present (H_1) is more destructive when estimating the speech than when estimating the noise.
- Hence, we make a distinction between the estimator $\hat{p}_{tk}$ used for estimating the clean speech, and the estimator $\tilde{p}_{tk}$, which controls the adaptation of the noise spectrum. Generally $\hat{p}_{tk} \geq \tilde{p}_{tk}$.

- The estimator $\tilde{p}_{tk}$ is biased toward higher values, since deciding speech is absent when speech is present results ultimately in the attenuation of speech components.
- Accordingly, we include a bias compensation factor in the noise estimator

$$\hat{\sigma}^2_{t+1,k} = \beta \cdot \bar{\sigma}^2_{t+1,k}$$

  such that the factor β ($\beta \geq 1$) compensates the bias when speech is absent:

$$\beta \triangleq \left. \frac{\sigma^2_{tk}}{E\left\{\bar{\sigma}^2_{tk}\right\}} \right|_{H_0} .$$

- The value of β is completely determined by the particular estimator for the *a priori* speech absence probability.

Minimum Statistics

- Let α_s $(0 < \alpha_s < 1)$ be a smoothing parameter, and let b denote a normalized window function of length $2w+1$, *i.e.*, $\sum_{i=-w}^{w} b_i = 1$.
- The frequency smoothing of the noisy power spectrum in each frame is defined by

$$S_{tk}^{f} = \sum_{i=-w}^{w} b_i \left| Y_{t,k-i} \right|^2 .$$

- Subsequently, smoothing in time is performed by a first order recursive averaging:

$$S_{tk} = \alpha_s S_{t-1,k} + (1-\alpha_s) S_{tk}^{f} .$$

- The minima values of S_{tk} are picked within a finite window of length D, for each frequency bin:

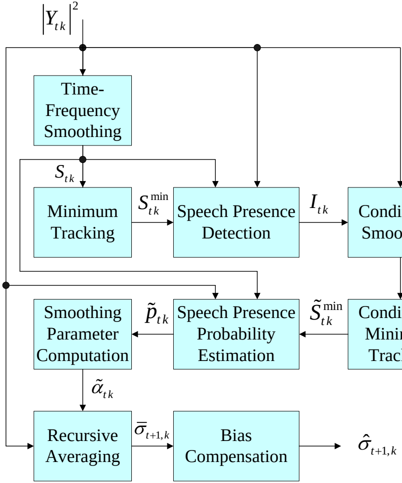
$$S_{tk}^{\min} \triangleq \min \left\{ S_{t',k} \mid t - D + 1 \leq t' \leq t \right\} .$$

- It follows that there exists a constant factor $B_{\min}$, independent of the noise power spectrum, such that

$$E\left\{ S_{tk}^{\min} \mid H_0 \right\} = B_{\min}^{-1} \cdot \sigma_{tk}^2 .$$

- The factor $B_{\min}$ represents the bias of a minimum noise estimate, and generally depends on the values of D, α_s, b and the spectral analysis parameters (type, length and overlap of the analysis windows)
- The value of $B_{\min}$ can be estimated by generating a white Gaussian noise, and computing the inverse of the mean of $S_{tk}^{\min}$.

Block diagram of the IMCRA noise estimator



Implementation

A free MATLAB code is available on:
http://www.ee.technion.ac.il/people/IsraelCohen/

Initialization at the first frame for all frequency-bins $k = 1, \ldots, N/2$:

$$\hat{\sigma}^2_{0k} = |Y_{0k}|^2; \quad \bar{\sigma}^2_{0k} = |Y_{0k}|^2; \quad S_{0k} = S^f_{0k}; \quad S^{\min}_{0k} = S^f_{0k};$$

For all short-time frames $t = 0, 1, \ldots$

For all frequency-bins $k = 1, \ldots, N/2$

1) Compute the *a posteriori* SNR $\hat{\gamma}_{tk}$ and the *a priori* SNR $\hat{\xi}_{tk}$ with the initial condition $\hat{\xi}_{0k} = \alpha + (1 - \alpha) \max\{\hat{\gamma}_{0k} - 1, 0\}$.

2) Compute the conditional spectral estimate under the hypothesis of speech presence $\hat{X}_{tk|H_1} = G_{\mathrm{LSA}}(\hat{\xi}_{tk}, \hat{\gamma}_{tk})\, Y_{tk}$.

3) Compute the smoothed power spectrum S_{tk} and update its running minimum: $S_{tk}^{\min} = \min\left\{S_{t-1,k}^{\min},\, S_{tk}\right\}$.

4) Compute the speech presence probability $\tilde{p}_{tk}$, and the smoothing parameter $\tilde{\alpha}_{tk}$.

5) Update the noise spectrum estimate $\hat{\sigma}_{t+1,k}^2$.

6) Compute the speech presence probability $\hat{p}_{tk}$.

7) Compute the speech spectral estimate $\hat{X}_{tk}$.

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)

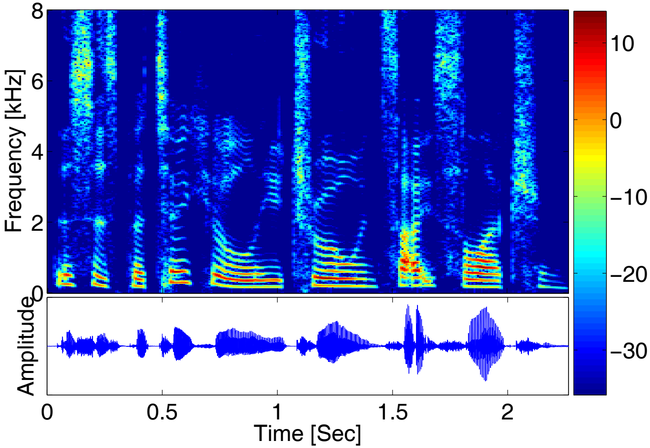
$$\mathrm{LSD} = \frac{1}{T}\sum_{t=0}^{T-1}\left[\frac{2}{N}\sum_{k=1}^{N/2}\left(\mathcal{L}X_{tk} - \mathcal{L}\hat{X}_{tk}\right)^2\right]^{\frac{1}{2}}$$

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

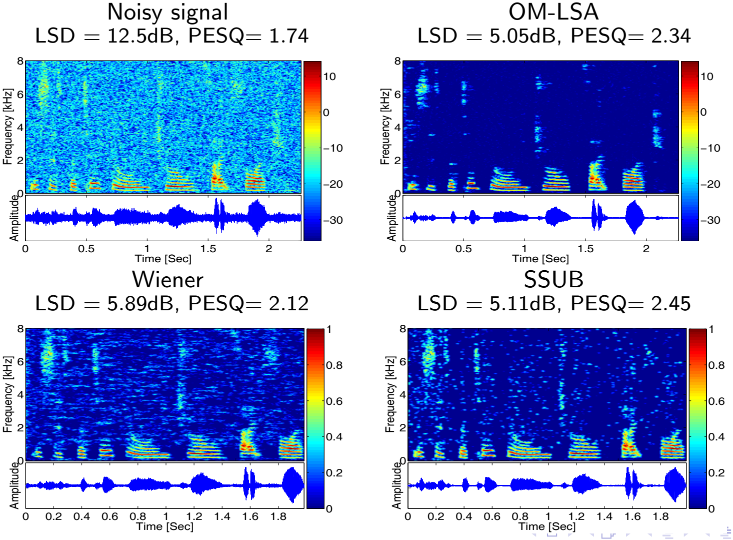
- Perceptual evaluation of speech quality (PESQ) score (ITU-T P.862).

Experimental Results - Clean Signal

"This is particularly true in site selection"

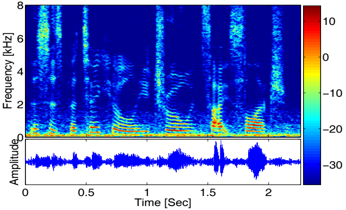


Experimental Results - White Gaussian Noise

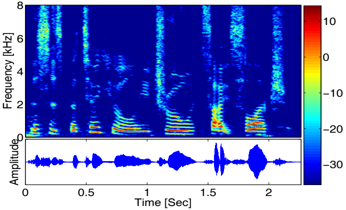


Experimental Results - Car Interior Noise

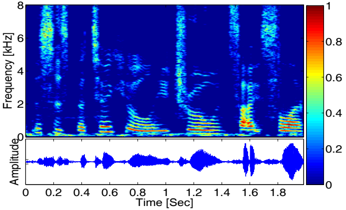
Noisy signal
LSD = 3.17dB, PESQ= 2.47



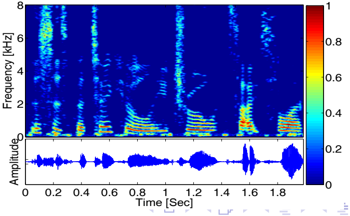
OM-LSA
LSD = 2.67dB, PESQ= 3.00



Wiener
LSD = 2.60dB, PESQ= 2.86

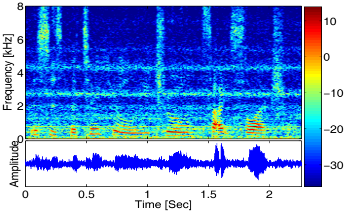


SSUB
LSD = 3.21dB, PESQ= 2.76

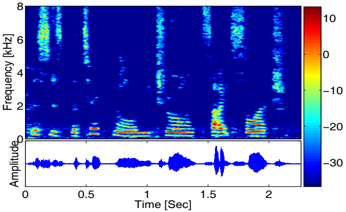


Experimental Results - F16 Cockpit Noise

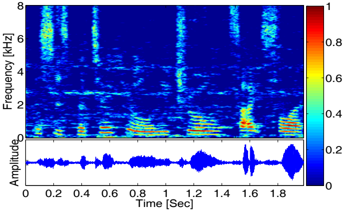
Noisy signal
LSD = 7.76dB, PESQ= 1.76



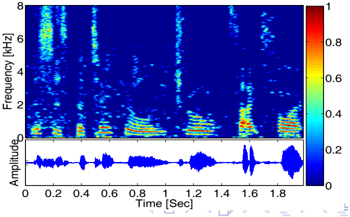
OM-LSA
LSD = 4.27dB, PESQ= 2.29



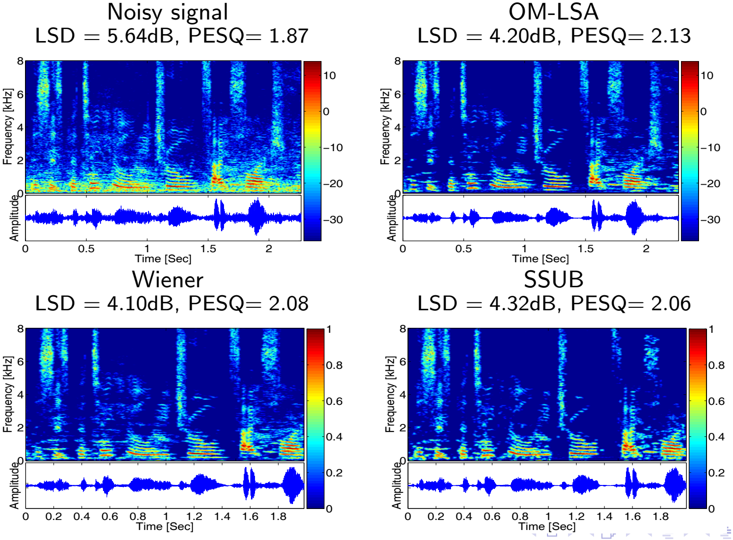
Wiener
LSD = 4.22dB, PESQ= 2.26



SSUB
LSD = 4.27dB, PESQ= 2.43



Experimental Results - Babble Noise



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.