PROSE: Perceptual Risk Optimization for Speech Enhancement

[1]Jishnu Sadasivan and [2]Chandra Sekhar Seelamantula
[1]Department of Electrical Communication Engineering, [2]Department of Electrical Engineering
Indian Institute of Science, Bangalore 560012, India
Emails : jishnus@ece.iisc.ernet.in, chandra.sekhar@ee.iisc.ernet.in

1. Overview

- We address the problem of suppressing noise from noisy speech within a risk minimization framework.
- The clean signal is estimated by minimizing an unbiased estimate of the risk function.
- We develop unbiased estimates of perceptual distortion functions.
- Minimize risk estimates to obtain the optimal denoising functions.
- For input SNR greater than 5 dB, the proposed algorithms outperform three benchmarking algorithms in terms of PESQ and SSNR scores.

2. Risk estimation principle

- **Observation model**:

$$x_n = s_n + w_n \quad n = 1, 2, \cdots, N.$$

- **Parameter estimation**: Obtain an estimate $\hat{s}$, of the (non-random) parameter that minimizes the risk:

$$\mathcal{R} = \mathcal{E}\left\{d\left(s, \hat{s}\right)\right\},$$

d measures the closeness between s and $\hat{s}$.

- Risk estimation approach: Since $\mathcal{R}$ depends on s, we estimate $\mathcal{R}$ and minimize it.
- **SURE**: An unbiased estimate of the MSE under i.i.d. gaussian assumption [1].
- **Our contribution**: Under the assumption a priori SNR is high and additive noise is a truncated gaussian, we develop perceptual risk estimates.
- Perceptual risk estimate is minimized to obtain the optimum shrinkage estimator.

3. Perceptual risk estimation

- Itakura-Saito distortion:

$$\mathcal{R}_{\text{IS}} := \mathcal{E}\left\{d_{\text{IS}}(s_k, \hat{s}_k) \Big| |w_k| < |x_k|\right\} \quad \text{where}$$

$$\begin{aligned} d_{\text{IS}}(s_k, \hat{s}_k) &= \frac{\hat{s}_k}{s_k} - \log\left(\frac{\hat{s}_k}{s_k}\right) - 1 \\ &= \frac{\hat{s}_k}{x_k}\left(1 - \frac{w_k}{x_k}\right)^{-1} - log\left(\hat{s}_k\right) + log\left(s_k\right) - 1 \\ &= \frac{\hat{s}_k}{x_k}\sum_{n=0}^{\infty}\left(\frac{w_k}{x_k}\right)^n - log\left(\hat{s}_k\right) + log\left(s_k\right) - 1. \end{aligned}$$

- Shrinkage estimator: $\hat{s}_k = a_k x_k$
- Truncating the series beyond $n = 4$ yields

$$\mathcal{R}_{\text{IS}} \approx \sum_{n=0}^{4} \mathcal{E}\left(a_k \frac{w_k^n}{x_k^n}\right) - \mathcal{E}\left\{log\left(a_k x_k\right)\right\} + log\left(s_k\right) - 1.$$

- Generalized Stein's Lemma: Let W be a real random variable with p.d.f

$$p\left(w; c_1, c_2, \sigma\right) = \frac{1}{\sqrt{2\pi}\sigma K}\exp\left(-\frac{w^2}{2\sigma^2}\right)\mathbb{1}_{\{-c_1\sigma < w < c_2\sigma\}}$$

where $K = \frac{1}{\sqrt{2\pi}\sigma}\int_{-c_1\sigma}^{c_2\sigma}\exp\left(-\frac{u^2}{2\sigma^2}\right)\mathrm{d}u$ and let $f: \mathbb{R} \rightarrow \mathbb{R}$ be an n-fold indefinite integral of the Lebesgue measurable function $f^{(n)}$, which is the n^{th} derivative of f. Suppose also that $\mathcal{E}\left\{|W^{(n-k)}f^{(k)}(W)|\right\} < \infty$, $c_1\sigma, c_2\sigma >> \sigma$, and $f^{(k)}\left(W\right)$ belongs to a class of functions such that $-\sigma^2 f^{(k)}(w)p\left(w; c_1, c_2, \sigma\right)\Big|_{-c_1\sigma}^{c_2\sigma} \approx 0, k = 1, 2, \cdots, n$. Then,

$$\mathcal{E}\{W^n f\left(W\right)\} \approx \sigma^2\mathcal{E}\{f'(W)W^{n-1}\} + \sigma^2(n-1)f\left(W\right)W^{n-2}\}.$$

- Using Lemma, the $\mathcal{R}_{\text{IS}}$ is

$$\mathcal{R}_{\text{IS}} = \mathcal{E}\left\{a_k\left(1 + 60\frac{\sigma^6}{x_k^6} + 840\frac{\sigma^8}{x_k^8}\right) - log(a_k x_k)\right\} - log(s_k) - 1.$$

- The unbiased estimate of $\mathcal{R}_{\text{IS}}$ is

$$\hat{\mathcal{R}}_{\text{IS}} = a_k\left(1 + 60\frac{\sigma^6}{x_k^6} + 840\frac{\sigma^8}{x_k^8}\right) - log(a_k x_k) - log(s_k) - 1.$$

- Differentiating $\mathcal{R}_{\text{IS}}$ with respect to a_k and equating to zero, we get that

$$a_{k,\text{opt}} = \left[1 + \frac{60}{\xi_k^3} + \frac{840}{\xi_k^4}\right]^{-1} \quad \text{where} \quad \xi_k = \frac{x_k^2}{\sigma^2}.$$

Table 1: Optimal shrinkage parameters corresponds to different perceptual risk estimate, where $[x]_+ = \max(0, x)$.

Risk	$d(s_k, \hat{s}_k)$	$a_{k,\text{opt}}$
MSE	$(\hat{s}_k - s_k)^2$	$\left[1 - \frac{1}{\xi_k}\right]_+$
log MSE	$\left(\log\frac{\hat{s}_k}{s_k}\right)^2$	$\exp\left(\frac{0.5}{\xi_k} - \frac{0.75}{\xi_k^2} - \frac{10}{\xi_k^3} - \frac{210}{\xi_k^4}\right)$
WE	$\frac{(\hat{s}_k - s_k)^2}{s_k}$	$\left[1 + \frac{1}{\xi_k} - \frac{1}{\xi_k^2} + \frac{48}{\xi_k^3} + \frac{360}{\xi_k^4}\right]_+^{-1}$
IS-II	$\frac{\hat{s}_k^2}{s_k^2} - \log\frac{\hat{s}_k^2}{s_k^2} - 1$	$\left[1 - \frac{1}{\xi_k} + \frac{24}{\xi_k^2} + \frac{360}{\xi_k^3} + \frac{4200}{\xi_k^4}\right]_+^{-\frac{1}{2}}$
COSH	$\frac{1}{2}\left[\frac{s_k}{\hat{s}_k} + \frac{\hat{s}_k}{s_k}\right] - 1$	$\sqrt{1 + \frac{1}{\xi_k}} \Big/ \sqrt{1 + 60\frac{1}{\xi_k^3} + 840\frac{1}{\xi_k^4}}$
WCOSH	$\left[\frac{s_k}{\hat{s}_k} + \frac{\hat{s}_k}{s_k} - 1\right]\frac{1}{s_k^p}$	$\left[1 - \frac{1}{\xi_k} + \frac{3}{\xi_k^2} + \frac{420}{\xi_k^3} + \frac{9450}{\xi_k^4}\right]_+^{-\frac{1}{2}}$

Implementation details:

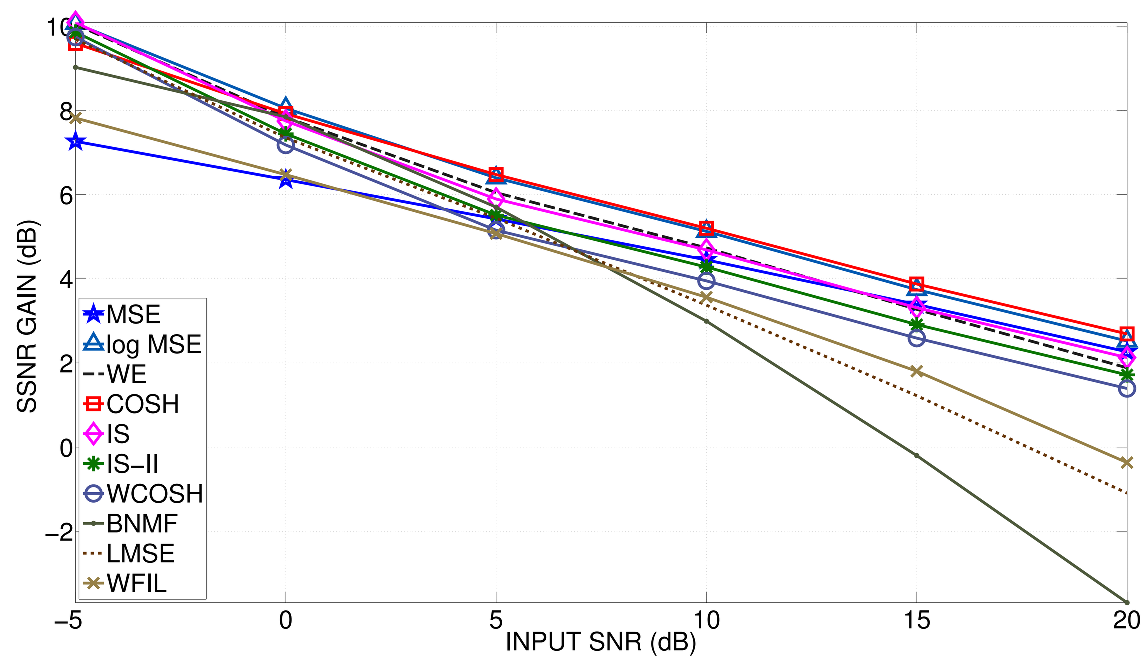
- We apply shrinkage estimator in DCT domain.
- Framewise processing: Frame length = 40 ms, 75% Overlap, Fs=8 kHz.

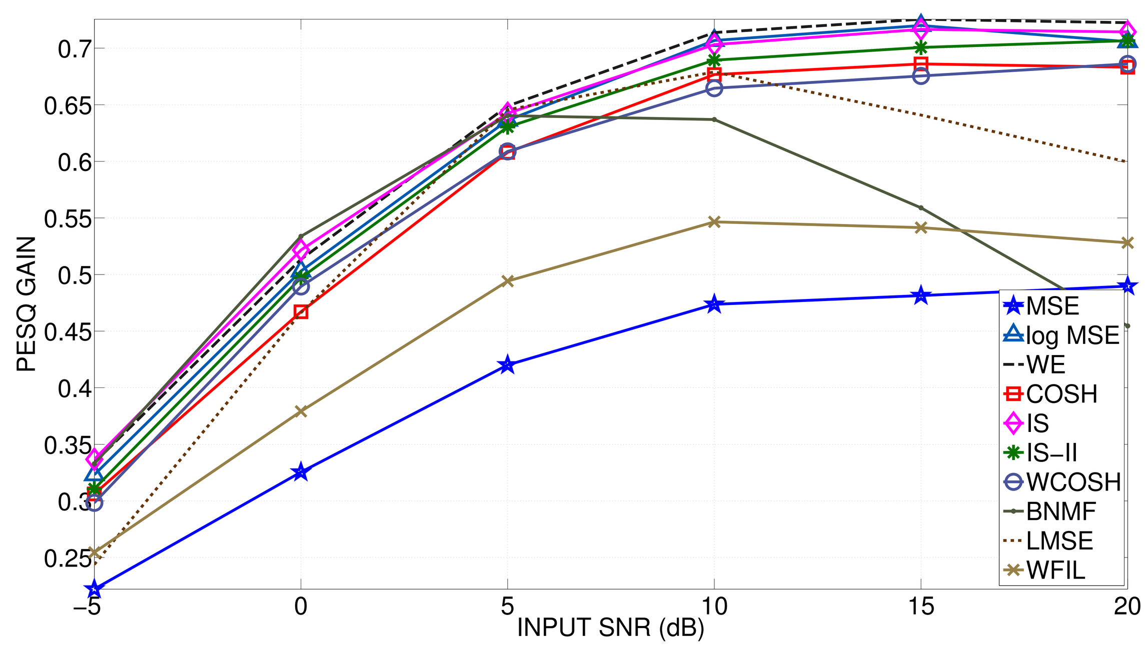
Benchmarking denoising algorithms: WFIL [3], LMSE [4], and BNMF [5].

4. Performance Comparison

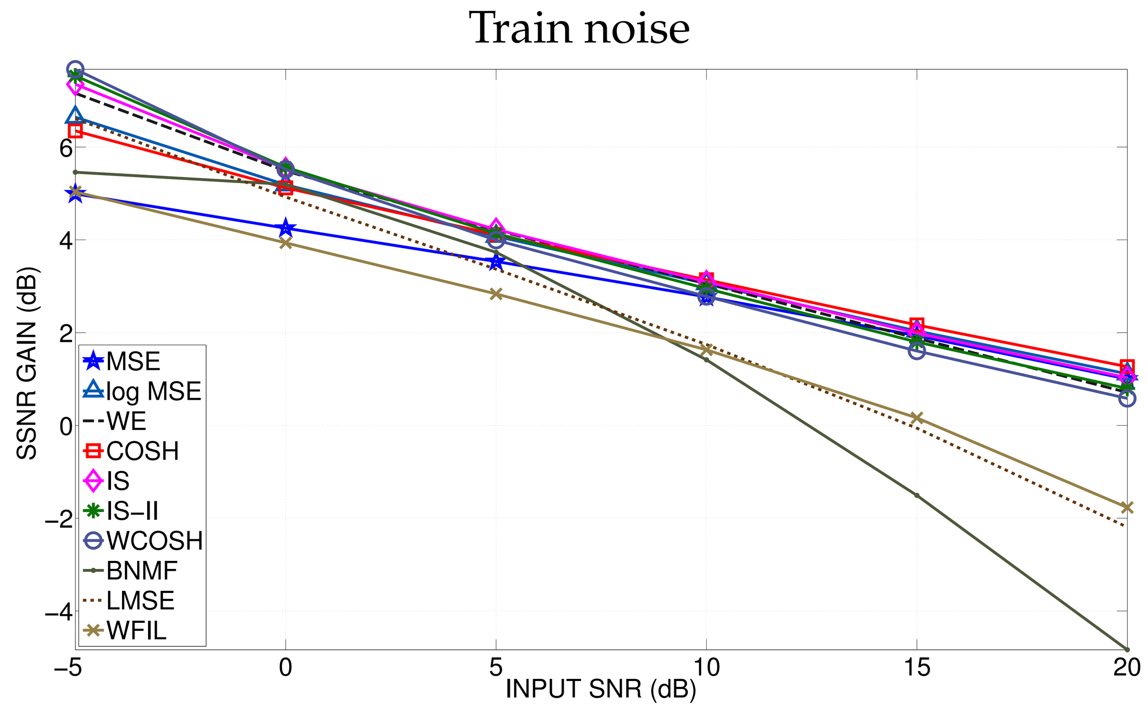
Results averaged over 10 different speech files and 50 different noise realizations (NOIZEUS database)

White noise





Train noise



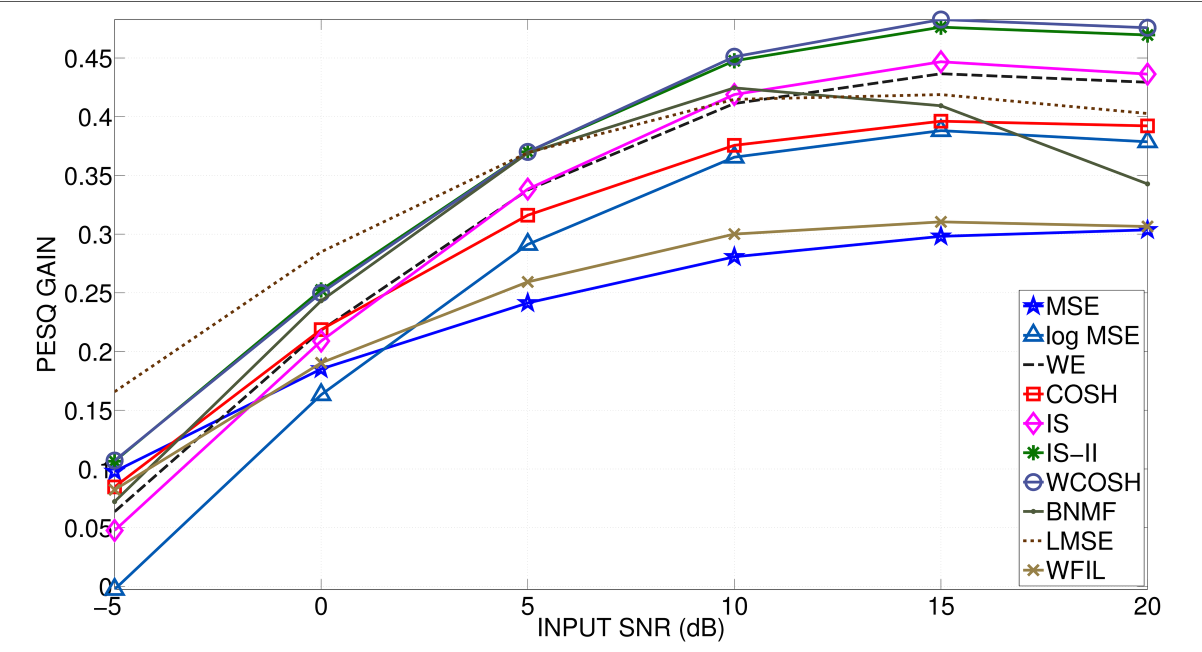


Figure 1: Performance comparison of the denoising algorithms.

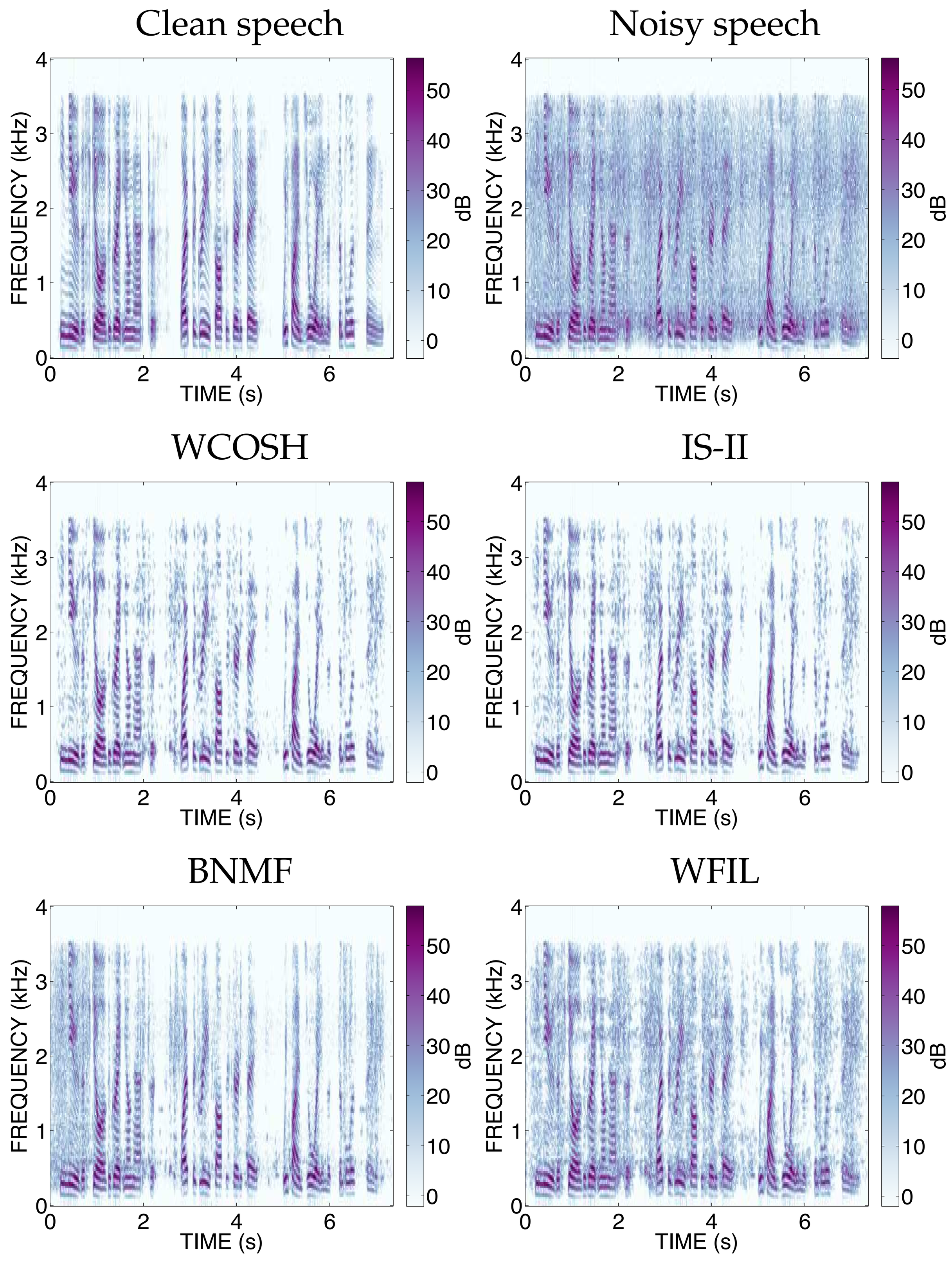


Figure 2: Spectrograms of denoised speech signals where noise corrupted is train noise with 10 dB input SNR .

Demo available online at http://spectrumee.wix.com/prose

5. Conclusion

- Introduced the notion of risk estimation for single-channel speech enhancement.
- Proposed unbiased estimates for perceptual distortion functions.
- Minimize risk estimates to obtain the optimum denoising functions.
- For SNR greater than 5 dB, the proposed approach resulted in better denoising performance than the benchmarking techniques.

6. References

[1] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *Ann. Stat.*, vol. 9, no. 6, pp. 1135–1151, Nov. 1981.

[2] R. M. Gray, A. Buzo, A. H. Gray, Jr., and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust. Speech Sig. Proc.*, vol. ASSP-28, pp. 367–376, Aug. 1980.

[3] P. Scalart, and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* , vol. 2, pp. 629–632, May. 1996.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-squared error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[5] N. Mohammadiha, P. Smaragdis, and A. Leijon," Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no.10, pp. 2140–2151, Oct. 2013.

[6] ITU-T Rec. P.862, "Perceptual Evaluation Of Speech Quality (PESQ), An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," International Telecommunication Union, Feb. 2001.

PROSE: Perceptual Risk Optimization for Speech Enhancement

[1] Jishnu S.
Supervisor : [2]Dr. Chandra Sekhar Seelamantula

[1]Department of Electrical Communication Engineering
[2]Department of Electrical Engineering
Indian Institute of Science
Bangalore – 560012, India

jishnus@ece.iisc.ernet.in

April 7, 2017

Outline

- Problem statement
- SURE
- Perceptual risk estimation
- Perceptual risk optimization for speech enhncement
- Conclusions

Problem statement

- Consider samples of a signal s_n, distorted by additive random noise w_n. The observation model is given by:

$$x_n = s_n + w_n. \qquad n = 1, 2 \cdots$$

- Goal: To estimate s_n from x_n, by minimizing a suitable distortion metric.

Risk estimation

- **Conventional method :** Obtain an estimate of s by minimizing the distortion function (risk) between estimate $\hat{s} = h(x)$ and s,

$$\hat{s} = \arg\min_{h(x)} \underbrace{\mathcal{E}\left\{d\left(h\left(x\right), s\right)\right\}}_{\mathcal{R}},$$

  where d measure the closeness between $h(x)$ and s.
- Direct minimization of cost requires the knowledge of underlying clean signal.
- **Risk Estimation :** Minimize an unbiased estimate of $\mathcal{R}$ to obtain $\hat{s}$.

Basic SURE formulation

- Consider MSE

$$\mathcal{R} = \mathcal{E}\left\{d\left(h(x), s\right)\right\} = \mathcal{E}\left\{\left(h\left(x\right) - s\right)^2\right\}$$
$$= \mathcal{E}\left\{s^2\right\} - 2\mathcal{E}\left\{h\left(x\right)s\right\} + \mathcal{E}\left\{h\left(x\right)^2\right\}.$$

where $x \sim \mathcal{N}(s, \sigma^2)$.

- SURE is an unbiased estimate of MSE obtained using Stein's lemma.
  (Stein, 1981)
  Let Y be a real random variable $\mathcal{N}(0, \sigma^2)$ and let $h : \mathbb{R} \to \mathbb{R}$ be an indefinite integral of the Lebesgue measurable function h', essentially the derivative of h. Suppose also that $\mathcal{E}_Y\left\{|h'(Y)|\right\} < \infty$. Then

$$\mathcal{E}_Y\left\{Yh(Y)\right\} = \sigma^2\mathcal{E}_Y\left\{h'(Y)\right\}$$

SURE

- Using Stein's lemma: $\mathcal{E}\left\{h\left(x\right)s\right\} = \mathcal{E}\left\{h\left(x\right)x\right\} - \sigma^2\mathcal{E}\left\{h'\left(x\right)\right\}$.
- Unbiased estimate of $\mathcal{R}$ becomes

$$\hat{\mathcal{R}} = s^2 - 2h\left(x\right)x + 2\sigma^2 h'\left(x\right) + h\left(x\right)^2$$

i.e. $\mathcal{R} = \mathcal{E}[\hat{\mathcal{R}}]$. Minimize $\hat{\mathcal{R}}$ to obtain $h\left(x\right)$.

- Clean speech DCT coefficient estimate, $h\left(x_k\right) = a_k x_k$, where $a_k \in [0, 1]$ and x_k is noisy DCT coefficient.
- Optimum pointwise shrinkage parameter $a_{k,opt} = \arg\min_{a_k} \hat{\mathcal{R}}$

$$a_{k,opt} = \left[1 - \frac{\sigma^2}{x_k^2}\right]_+ \text{ where } \quad [x]_+ = \max(0, x).$$

Perceptual risk estimation

- Perceptual distortion functions: Itakura-Saito distortion, hyperbolic-cosine (cosh) distortion, weighted cosh distortion, etc. [2].
- Practical noise types are bounded, hence one can model the noise using a truncated Gaussian distribution.
- Assuming observation distribution is truncated gaussian and SNR is high, we propose risk estimate for perceptual distortion functions.
- Minimize perceptual risk estimates to obtain optimum shrinkage estimators.

Itakura Saito(IS) Distortion

- $\mathcal{R}_{\mathcal{IS}} := \mathcal{E}\left\{ d_{IS}(s_k, \hat{s}_k) \middle| |w_k| < |x_k| \right\}$ where

$$
\begin{aligned}
d_{IS}(s_k, \hat{s}_k) &= \frac{\hat{s}_k}{s_k} - \log\left(\frac{\hat{s}_k}{s_k}\right) - 1 \\
&= \frac{\hat{s}_k}{x_k}\left(1 - \frac{w_k}{x_k}\right)^{-1} - log\left(\hat{s}_k\right) + log\left(s_k\right) - 1 \\
&= \frac{\hat{s}_k}{x_k}\sum_{n=0}^{\infty}\left(\frac{w_k}{x_k}\right)^{n} - log\left(\hat{s}_k\right) + log\left(s_k\right) - 1.
\end{aligned}
$$

- Truncating the series beyond n=4 using $\hat{s}_k = a_k x_k$ yields

$$
\mathcal{R}_{\mathcal{IS}} \approx \sum_{n=0}^{4} \mathcal{E}\left(a_k \frac{w_k^n}{x_k^n}\right) - \mathcal{E}\left\{log\left(a_k x_k\right)\right\} + log\left(s_k\right) - 1.
$$

Lemma 1

Let W be a real random variable with p.d.f

$$p\left(w; c_1, c_2, \sigma\right) = \frac{1}{\sqrt{2\pi}\sigma K} \exp\left(-\frac{w^2}{2\sigma^2}\right) \mathbb{1}_{\{-c_1\sigma < w < c_2\sigma\}}$$

where $K = \frac{1}{\sqrt{2\pi}\sigma} \int_{-c_1\sigma}^{c_2\sigma} \exp\left(-\frac{u^2}{2\sigma^2}\right) \mathrm{d}u$ and let $f : \mathbb{R} \to \mathbb{R}$ be an n-fold indefinite integral of the Lebesgue measurable function $f^{(n)}$, which is the n^{th} derivative of f. Suppose also that $\mathcal{E}\left\{|W^{(n-k)} f^{(k)}(W)|\right\} < \infty$, $c_1\sigma, c_2\sigma >> \sigma$, and $f^{(k)}\left(W\right)$ belongs to a class of functions such that $-\sigma^2 f^{(k)}(w) p\left(w; c_1, c_2, \sigma\right)\Big|_{-c_1\sigma}^{c_2\sigma} \approx 0, k = 1, 2, \cdots, n$. Then

$$\mathcal{E}\{W^n f\left(W\right)\} \approx \sigma^2 \mathcal{E}\{f'(W) W^{n-1}\} + \sigma^2 (n-1) \mathcal{E}\{f\left(W\right) W^{n-2}\}.$$

- Using Lemma 1, the risk $\mathcal{R}_{IS}$ is

$$\mathcal{R}_{\mathcal{IS}} = \mathcal{E}\left\{a_k\left(1 + 60\frac{\sigma^6}{x_k^6} + 840\frac{\sigma^8}{x_k^8}\right) - log(a_k x_k)\right\} - log(s_k) - 1.$$

- The unbiased estimate of $\mathcal{R}_{\mathcal{IS}}$ is

$$\hat{\mathcal{R}}_{\mathcal{IS}} = a_k\left(1 + 60\frac{\sigma^6}{x_k^6} + 840\frac{\sigma^8}{x_k^8}\right) - log(a_k x_k) - log(s_k) - 1.$$

- Differentiating $\mathcal{R}_{\mathcal{IS}}$ with respect to a_k and equating to zero, we get that

$$a_{k,opt} = \left[1 + \frac{60}{\xi_k^3} + \frac{840}{\xi_k^4}\right]^{-1}$$

where $\xi_k = \dfrac{x_k^2}{\sigma^2}$.

Table: Optimal shrinkage parameters for different perceptual risk estimates.

risk	$d(s_k, \hat{s}_k)$	a_{opt}
log MSE	$\left(\log \frac{\hat{s}_k}{s_k}\right)^2$	$\exp\left(\frac{0.5}{\xi_k} - \frac{0.75}{\xi_k^2} - \frac{10}{\xi_k^3} - \frac{210}{\xi_k^4}\right)$
WE	$\frac{(\hat{s}_k - s_k)^2}{s_k}$	$\left[1 + \frac{1}{\xi_k} - \frac{1}{\xi_k^2} + \frac{48}{\xi_k^3} + \frac{360}{\xi_k^4}\right]_+^{-1}$
IS-II	$\frac{\hat{s}_k^2}{s_k^2} - \log \frac{\hat{s}_k^2}{s_k^2} - 1$	$\left[1 - \frac{1}{\xi_k} + \frac{24}{\xi_k^2} + \frac{360}{\xi_k^3} + \frac{4200}{\xi_k^4}\right]_+^{-\frac{1}{2}}$
COSH	$\frac{1}{2}\left[\frac{s_k}{\hat{s}_k} + \frac{\hat{s}_k}{s_k}\right] - 1$	$\sqrt{1 + \frac{1}{\xi_k}} \Big/ \sqrt{1 + 60\frac{1}{\xi_k^3} + 840\frac{1}{\xi_k^4}}$
WCOSH	$\left[\frac{s_k}{\hat{s}_k} + \frac{\hat{s}_k}{s_k} - 1\right]\frac{1}{s_k^p}$	$\left[1 - \frac{1}{\xi_k} + \frac{3}{\xi_k^2} + \frac{420}{\xi_k^3} + \frac{9450}{\xi_k^4}\right]_+^{-\frac{1}{2}}$

where $\xi_k = \frac{x_k^2}{\sigma^2}$.

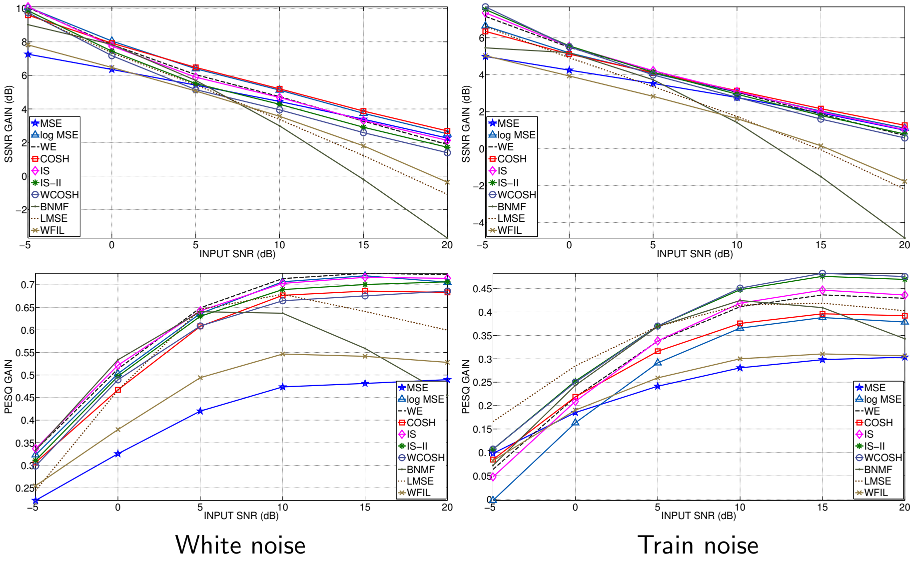


Figure: Performance comparison of different denoising algorithms.

Conclusion

- Introduced the notion of risk estimation for single-channel speech enhancement.
- We proposed risk estimates for perceptual distortion metrics and minimize to obtain the optimum denoising function.
- For SNR greater than 5 dB, the proposed approach resulted in better denoising performance than the benchmarking techniques.

References

[1] C.M Stein, "Estimation of the mean of a multivariate normal distribution," *Ann. Stat.*, vol. 9, no. 6, pp. 1135-1151, Nov. 1981.

[2] R. M. Gray, A. Buzo, A. H. Gray, Jr., and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust. Speech Sig. Proc.*, vol. ASSP-28, pp. 367–376, Aug. 1980.

[3] P. Scalart, and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* , vol. 2, pp. 629–632, May. 1996.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-squared error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[5] N. Mohammadiha, P. Smaragdis, and A. Leijon," Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no.10, pp. 2140–2151, Oct. 2013.

THANK YOU