Technion-Israel Institute of Technology
Department of Electrical Engineering

Signal and Image Processing Lab

Wavelet-Based Denoising of Speech

Arkady Bron

supervised by

Prof. Shalom Raz and Prof. David Malah

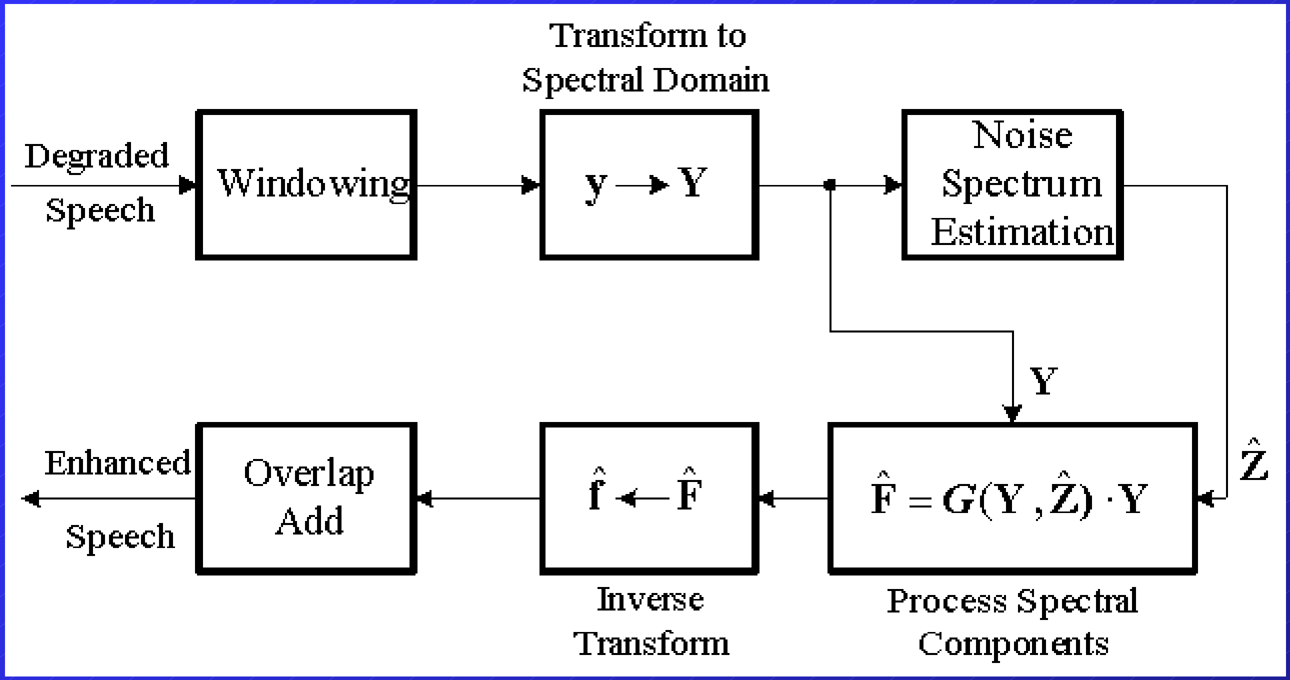
Outline

- Why do we need to enhance speech?
- State of the art of speech denoising algorithms
- Joint time-frequency representations
- Wavelet-based denoising techniques
- The proposed speech denoising algorithms
- A comparative performance analysis
- Summary and conclusions

Why do we need to enhance speech?

- Improvement in the quality and comprehension of speech.
- Preprocessing stage in coding and recognition techniques.

State of the Art of Speech Denoising



$$G(Y_k, \hat{Z}_k) = \begin{cases} \left(1 - \alpha\left[\dfrac{|\hat{Z}_k|}{|Y_k|}\right]^{\gamma_1}\right)^{\gamma_2}, & \left[\dfrac{|\hat{Z}_k|}{|Y_k|}\right]^{\gamma_1} < \dfrac{1}{\alpha+\beta}, \\ \left(\beta\left[\dfrac{|\hat{Z}_k|}{|Y_k|}\right]^{\gamma_1}\right)^{\gamma_2}, & \text{otherwise.} \end{cases}$$

α – *oversubtraction factor*

β – *spectral flooring factor*

γ_1, γ_2 – *exponent parameters*

- Choosing $\alpha = 1$, $\beta = 0$, $\gamma_1 = 2$, $\gamma_2 = 1$ we obtain the so-called non-causal Wiener filter

Ephraim-Malah (E-M) Speech Denoising Algorithm (1984/5)

- 1984 – Spectral Amplitude Estimator
- 1985 – Log-Spectral Amplitude Estimator

$$E\left\{\left(\log A_k - \log \hat{A}_k\right)^2\right\} \rightarrow \min$$

$$\xi_k = \frac{E\{|F_k|^2\}}{E\{|Z_k|^2\}} \; (a\ priori\ SNR) \qquad \gamma_k = \frac{|Y_k|^2}{E\{|Z_k|^2\}} \; (a\ posteriori\ SNR)$$

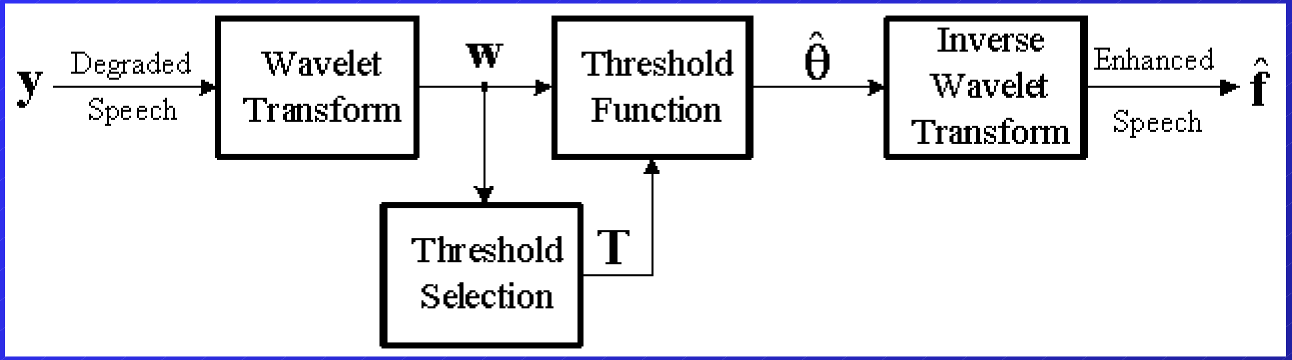
"Decision Directed" a priori SNR Estimation

$$\hat{\xi}_k(n) = \alpha \frac{\left|\hat{F}_k(n-1)\right|^2}{E\{|Z_k(n-1)|^2\}} + (1-\alpha)\eta_s\left(\gamma_k(n),1\right)$$

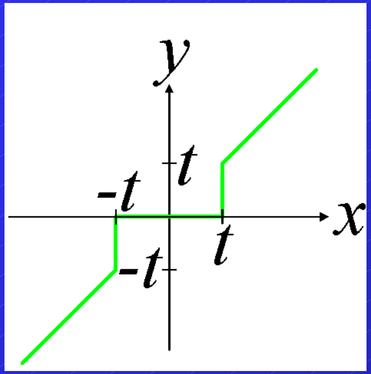
$$\eta_s\left(\gamma_k(n),1\right) = \begin{cases} \gamma_k(n)-1, & \gamma_k(n) \geq 1 \\ 0, & \gamma_k(n) < 1 \end{cases} \qquad \hat{\xi}_k(0) = \alpha + (1-\alpha)\eta_s\left(\gamma_k(0),1\right)$$

Wavelet-Based Denoising Techniques

The Donoho-Johnstone Algorithm (1994/5)



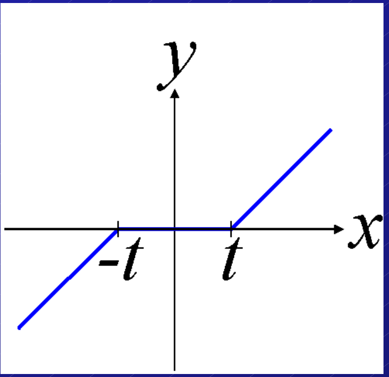
$\mathbf{w} = \{w_{\ell,n,k}\}$ $\mathbf{T} = \{t_{\ell,n,k}\}$ $\hat{\boldsymbol{\theta}} = \{\hat{\theta}_{\ell,n,k}\}$



$y = \eta_h(x,t)$

$$\eta_h(x,t) = \begin{cases} x, & |x| > t \\ 0, & |x| \le t \end{cases}$$

$$\eta_s(x,t) = \operatorname{sign}(x) \cdot \begin{cases} |x| - t, & |x| > t \\ 0, & |x| \le t \end{cases}$$



$y = \eta_s(x,t)$

$$\hat{\theta}_{\ell,n,k} = \eta_s\left(w_{\ell,n,k}, t_{\ell,n,k}\right)$$

Implementation and Quality Measures

All examinations were done for 3 following sentences, each pronounced by a male and a female:

- A lathe is a big tool
- An icy wind raked the beach
- Joe brought a young girl

Each sentence was sampled at 8 KHz sampling frequency and has 16384 samples (J=14).

$$SNR = 10\log_{10}\left(\frac{\|\mathbf{f}\|_2^2}{\|\mathbf{f}-\hat{\mathbf{f}}\|_2^2}\right)\ [dB]$$

$$SEGSNR = \frac{1}{M}\sum_{i=1}^{M} SNR_i,\quad SNR_i = 10\log_{10}\left(\frac{\|\mathbf{f}_i\|_2^2}{\|\mathbf{f}_i-\hat{\mathbf{f}}_i\|_2^2}+1\right)\ [dB]$$

$$LSD = \frac{1}{M}\sum_{i=1}^{M} D_i,\quad D_i = \left[\frac{1}{N}\sum_{k=1}^{N}\left(10\log_{10}|F_i(k)| - 10\log_{10}|\hat{F}_i(k)|\right)^2\right]^{\frac{1}{2}}\ [dB]$$

$$F_i(k) = DFT\{\mathbf{f}_i\}(k),\quad \hat{F}_i(k) = DFT\{\hat{\mathbf{f}}_i\}(k)$$

WPD-Based Denoising of Speech (1)

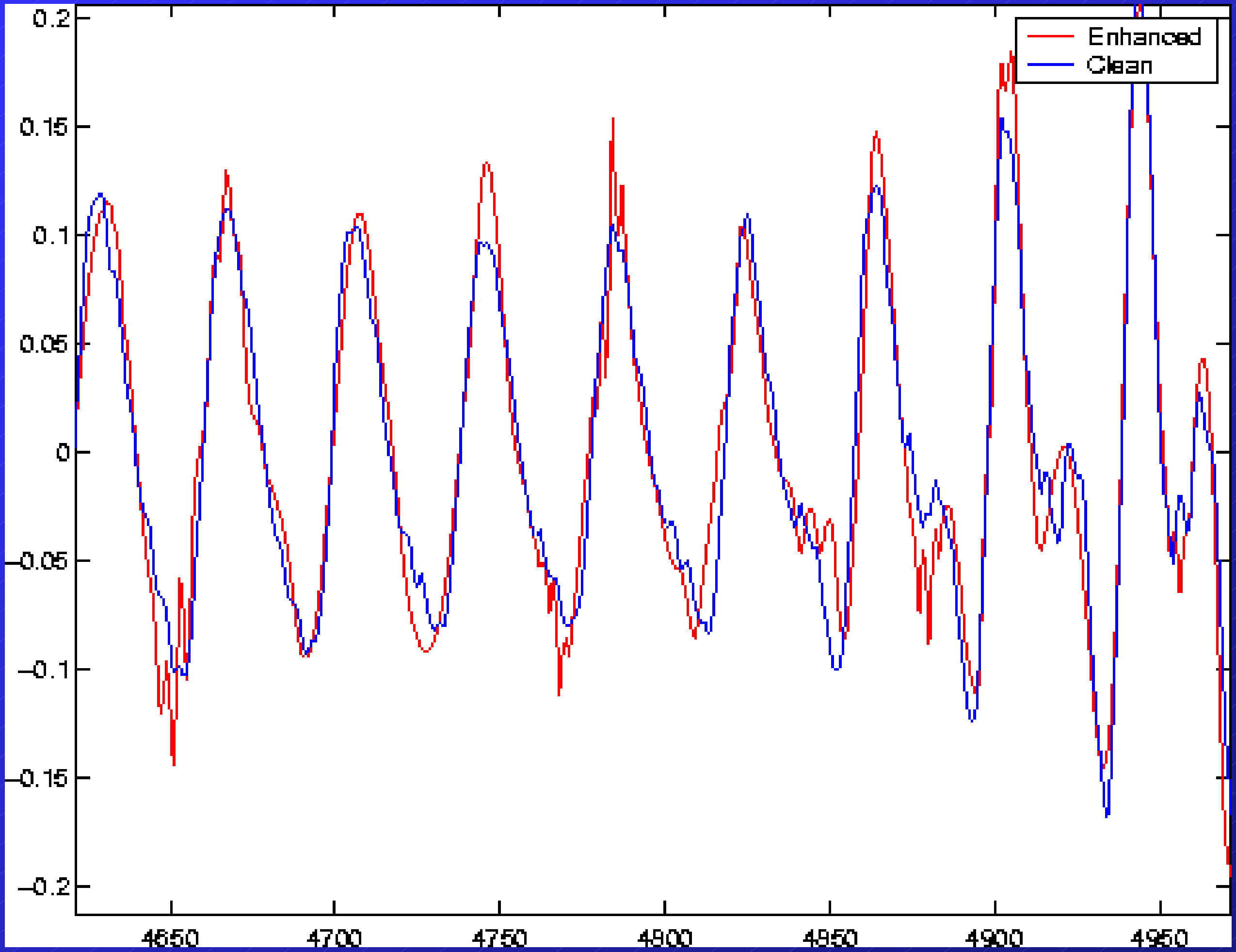
- Daubechies nearly symmetric mother wavelet of 8'th order (DNS(8))
- Entropy-based best-basis selection (*L*=6)
- Soft-thresholding

Test sentence #2, pronounced by a female • Clean Speech • Noisy Speech

Estimator type	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
VisuShrink	10	6.68	9.47	10.08	6.76	6.88
RiskShrink	10	6.68	9.47	12.69	8.13	6.47
SureShrink	10	6.68	9.47	14.73	9.28	6.35
Wiener	10	6.68	9.47	13.35	8.63	7.34

- Thresholding-based algorithms – oversmoothing and artifacts
- Use of Shift-Invariant WPD (Cohen, Raz and Malah, 1997) didn't improve denoising performance

Oversmoothing and Artifacts in Thresholding-Based Denoising



Oversmoothing in speech, enhanced by SureShrink

Artifacts in speech, enhanced by RiskShrink

Suppression of Artifacts

Increasing temporal support of basis functions:

- Choosing appropriate cost function
- Increasing temporal support of mother wavelet

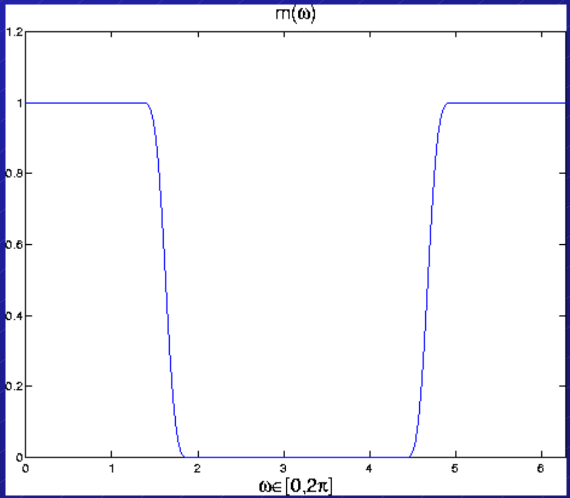
Influence of Cost Function

- There is no significant difference in the quality of enhanced speech
- Full subband WPD-based denoising attains the highest SNR

Increasing Temporal Support of Mother Wavelet

- Generalized Meyer mother wavelet

$$m(\omega)=\begin{cases}1, & |\omega|\le\frac{\pi}{2}(1-r)\\ \cos\left[\frac{\pi}{2}\nu\left(\dfrac{|\omega|-\frac{\pi}{2}(1-r)}{\pi r}\right)\right], & \frac{\pi}{2}(1-r)\le|\omega|\le\frac{\pi}{2}(1+r)\\ 0, & |\omega|\ge\frac{\pi}{2}(1+r)\end{cases}$$



$r-roll-off$ $\qquad m_0(\omega)=m(\omega)|_{r=\frac{1}{3}}$ $\qquad r=\frac{1}{5}$

Temporal Support and Frequency Localization

- Entropy-based best-basis selection algorithm (L=6)
- SureShrink and Wiener estimators

DNS(8) mother wavelet

Estimator type	Input SNR	Output SNR
SureShrink	10	14.73
Wiener	10	13.35

Generalized Meyer mother wavelet

Estimator type	N, r	Output SNR
SureShrink	32, 1/3	14.81
SureShrink	64, 1/5	14. 92
Wiener	32, 1/3	13.53
Wiener	64, 1/5	13. 63

- Increasing temporal support suppress the artifacts
- Improving frequency localization improves the resulting SNR

Framing

- Denoising without framing: output SNR=13.63[dB]
- Framing (Hanning window, 50% overlapping, 256 samples per frame): output SNR=15.69[dB]
- Framing improves resulting SNR
- Smoothing of gains fluctuations is needed

Utilization of the "Decision Directed" A Priori SNR Estimation

- Tracking a priori SNR for decomposition tree terminal nodes: the full subband decomposition is the optimal choice (*L*=*J*)

$$G(\mathbf{w}_{\ell,n}(j),\hat{\mathbf{z}}_{\ell,n}(j)) = \frac{\hat{\xi}_{\ell,n}(j)}{\hat{\xi}_{\ell,n}(j)+1} \qquad \xi_{\ell,n}(j) = \frac{\|\boldsymbol{\theta}_{\ell,n}(j)\|_2^2}{\|\hat{\mathbf{z}}_{\ell,n}(j)\|_2^2} \ \ (a\ priori\ SNR)$$

$$\hat{\xi}_{\ell,n}(j) = \alpha\frac{\|\hat{\boldsymbol{\theta}}_{\ell,n}(j-1)\|_2^2}{\|\hat{\mathbf{z}}_{\ell,n}(j-1)\|_2^2} + (1-\alpha)\eta_s(\gamma_{\ell,n}(j),1), \ \ j=2,3,\ldots M$$

$$\gamma_{\ell,n}(j) = \frac{\|\mathbf{w}_{\ell,n}(j)\|_2^2}{\|\hat{\mathbf{z}}_{\ell,n}(j)\|_2^2} \ \ (a\ posteriori\ SNR) \qquad \hat{\xi}_{\ell,n}(1) = \alpha + (1-\alpha)\eta_s(\gamma_{\ell,n}(1),1)$$

Proposed WPD-Based Speech Denoising Algorithm

- Wiener estimator, combined with the "decision directed" a priori SNR estimation (α=0.9, Hanning window, 50% overlapping, 256 samples per frame)

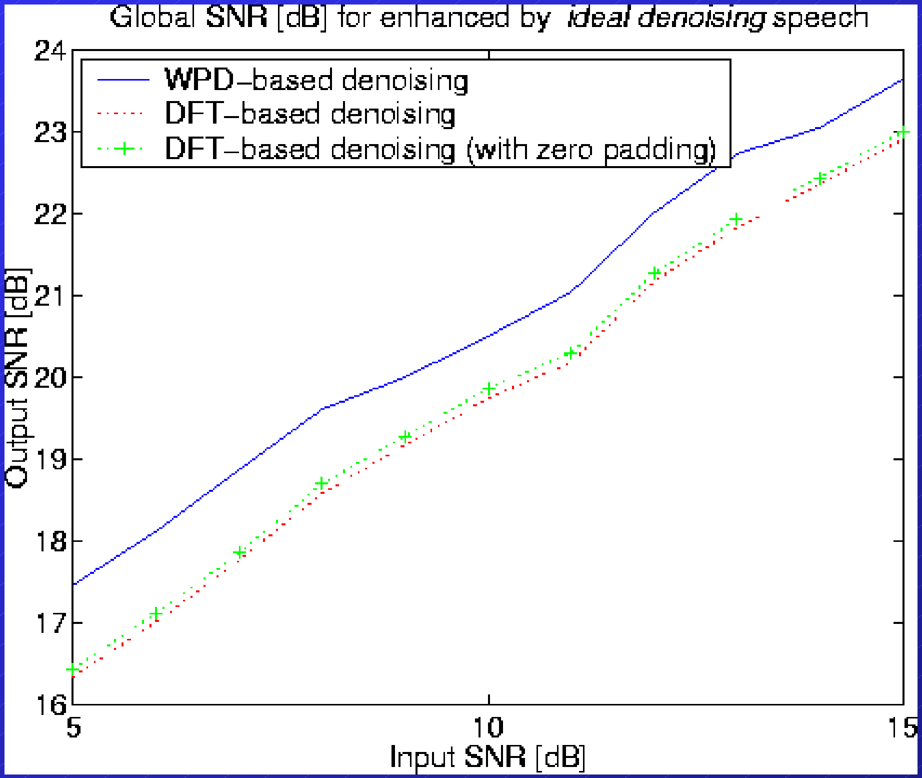
- Full Subband decomposition (L=J=8)

- Generalized Meyer mother wavelet (N = 64, r = 0.1)

#	Speaker	Decomposition type	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
1	Female	*WPD*	10	6.06	11.51	17.37	9.58	8.52
1	Male	*WPD*	10	5.96	11.53	15.95	8.48	9.62
2	Female	*WPD*	10	6.68	9.47	16.01	9.58	6.62
2	Male	*WPD*	10	6.73	9	15.05	9.54	6.31
3	Female	*WPD*	10	6.17	11.11	16.56	9.01	9.12
3	Male	*WPD*	10	5.92	11.51	15.7	7.94	9.96

"Ideal" Denoising

- "Ideal" denoising –assuming prior knowledge of noise squared-spectral amplitude exact value
- Results of "ideal" denoising:



- Better frequency resolution when compared to DFT-based denoising (zero padding for DFT-based denoising improves resulting SNR)
- Exact phase reconstruction

A Comparative Performance Analysis (1)

- Results of practical denoising:

Estimator type, decomposition type	**Input SNR**	**Output SNR**
Wiener, WPD	10	17.37
Wiener, CPD	10	16.69
Wiener, WPD(DCT)	10	16.49

Estimator type, decomposition type	**Input SNR**	**Output SNR**
Wiener, DFT	10	17.83
E-M, DFT	10	17.22

- DFT-based Wiener estimator attains the highest SNR and is characterized by the lowest level of the residual background noise
- E-M algorithm is characterized by approximately white residual background noise and by the best quality of enhanced speech
- WPD-based denoising algorithm attains SNRs, close to resulting by E-M algorithm SNR
- Denoising algorithms, based on LTD and WPD applied to DCT coefficients, attain the lowest SNRs, comparing to other transforms; speech quality is comparable to other algorithms

DFT-Based Denoising vs. Real-Valued Transforms-Based Denoising

- Given only noisy observations and estimated noise squared-spectral components, the phase of clean speech can not be any more exactly reconstructed using real-valued transform

- The variance of noise squared-spectral components, obtained by real-valued transform, is twice the variance of noise squared-spectral components, obtained by DFT (except the DC coefficient)

Summary

- Thresholding-based denoising techniques using WPD (or LTD) have low performance when applied to speech (hoarseness and artifacts)

- We have proposed speech denoising algorithms, that are based on WPD and LTD

- Enhanced speech quality is good, and resulting quantitave measures are close to benchmark DFT-based speech denoising algorithms

- Proposed WPD-based speech denoising algorithm is recommended for using with WPD-based speech coding techniques

- Proposed LTD-based speech denoising algorithm is characterized by lower complexity than WPD-based while obtaining good quality of enhanced speech and is recommended for combined speech denoising and segmentation

- We have presented results of theoretical investigations