



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.

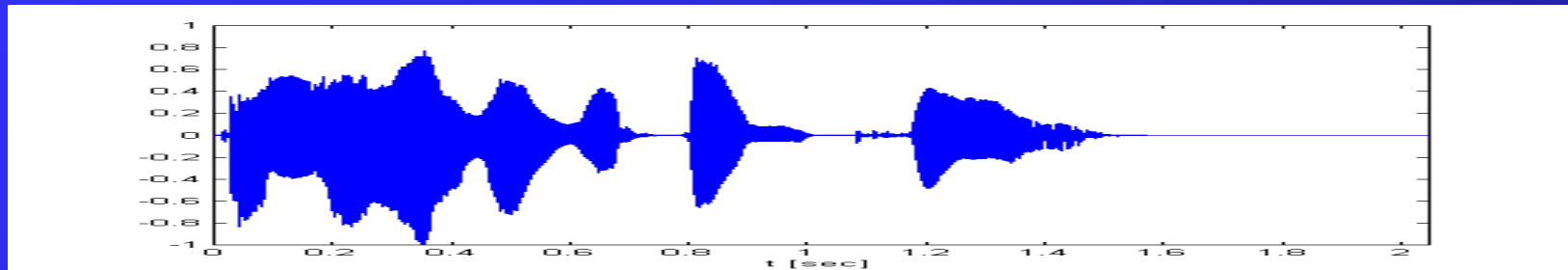


Speech Examples

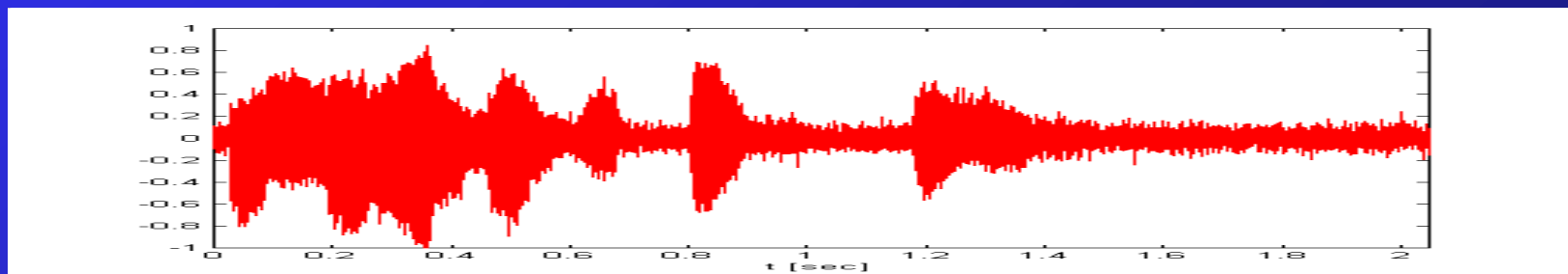
“An icy wind raked the beach” pronounced by a female



Clean speech $\mathbf{f} = \{f_i\}_{i=0}^{N-1}$

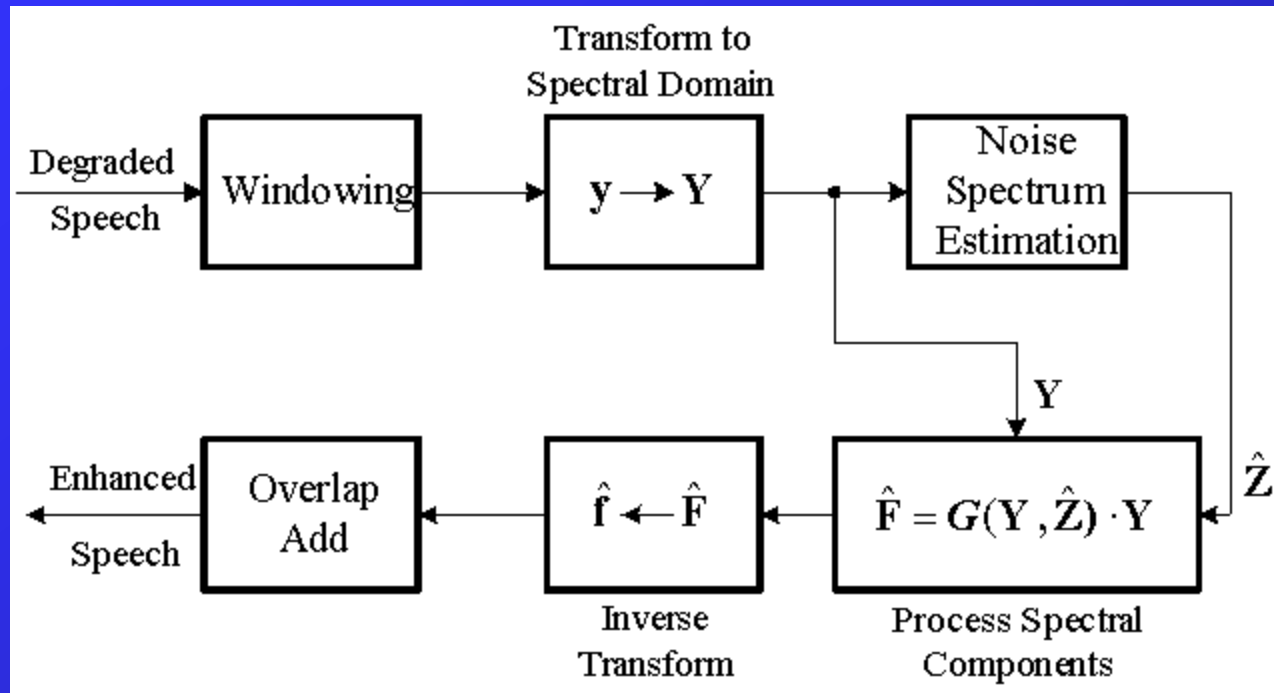


Noisy speech



$$\mathbf{y} = \{y_i\}_{i=0}^{N-1} = \{f_i + e_i\}_{i=0}^{N-1} \quad e_i \sim N(0, \sigma^2) \quad SNR = 10dB$$

State of the Art of Speech Denoising (1)



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

α – oversubtraction factor
 β – spectral flooring factor
 γ_1, γ_2 – exponent parameters

Non-casual Wiener filter

$$G_w(Y_k, \hat{Z}_k) = \begin{cases} \left(1 - \left[\frac{|\hat{Z}_k|}{|Y_k|}\right]^2\right), & \left[\frac{|\hat{Z}_k|}{|Y_k|}\right]^2 < 1, \\ 0, & \text{otherwise.} \end{cases} \quad \alpha = 1, \beta = 0, \gamma_1 = 2, \gamma_2 = 1$$

Amplitude Spectral Subtraction

$$G_A(Y_k, \hat{Z}_k) = \begin{cases} \left(1 - \frac{|\hat{Z}_k|}{|Y_k|}\right), & \frac{|\hat{Z}_k|}{|Y_k|} < 1, \\ 0, & \text{otherwise.} \end{cases}$$

$$\alpha = 1, \beta = 0, \gamma_1 = \gamma_2 = 1$$

Power Spectral Subtraction

$$G_P(Y_k, \hat{Z}_k) = \begin{cases} \left(1 - \left[\frac{|\hat{Z}_k|}{|Y_k|}\right]^2\right)^{\frac{1}{2}}, & \left[\frac{|\hat{Z}_k|}{|Y_k|}\right]^2 < 1, \\ 0, & \text{otherwise.} \end{cases}$$

$$\alpha = 1, \beta = 0, \gamma_1 = 2, \gamma_2 = 1/2$$

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

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

$$E\left\{\left(\log A_k - \log \hat{A}_k\right)^2\right\} \rightarrow \min \quad G(\xi_k, \gamma_k) = \frac{\xi_k}{\xi_k + 1} \exp\left\{\frac{1}{2} \int_{\nu_k}^{\infty} \frac{e^{-t}}{t} dt\right\}$$

$$\nu_k = \frac{\xi_k}{\xi_k + 1} \gamma_k \quad \xi_k = \frac{E\{|F_k|^2\}}{E\{|Z_k|^2\}} \text{ (a priori SNR)} \quad \gamma_k = \frac{|Y_k|^2}{E\{|Z_k|^2\}} \text{ (a posteriori SNR)}$$

“Decision Directed” a priori SNR Estimation

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

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

Joint Time-Frequency Representations (1)

Wavelet Packet Decomposition (WPD)

$$\mathcal{B} = \left\{ \psi_{\ell,n,k}(t) = 2^{\ell/2} \psi_n(2^\ell t - k) : \ell \in \mathbb{Z}_-, n \in \mathbb{Z}_+, k \in \mathbb{Z} \right\} \quad (\text{library of wavelet packets})$$

ℓ – scaling parameter (resolution level) n – oscillation parameter

k – time – domain position index

$$\psi_{2n}(t) \equiv \sqrt{2} \sum_k h_k \psi_n(2t - k) \equiv H \psi_n(t) \quad (\text{low-pass filtering, followed by decimation (2:1)})$$

$$\psi_{2n+1}(t) \equiv \sqrt{2} \sum_k g_k \psi_n(2t - k) \equiv G \psi_n(t) \quad (\text{high-pass filtering, followed by decimation (2:1)})$$

$$\sum_l h_{l-2k}^* g_{l-2n} = 0 \quad \sum_k h_k = \sqrt{2} \quad \sum_k g_k = 0 \quad (\text{orthogonality, perfect reconstruction and "admissibility" conditions})$$

$$\sum_l [h_{k-2l} h_{m-2l}^* + g_{k-2l} h_{m-2l}^*] = \delta_{m,k}$$

$$g_k = (-1)^k h_{1-k}$$

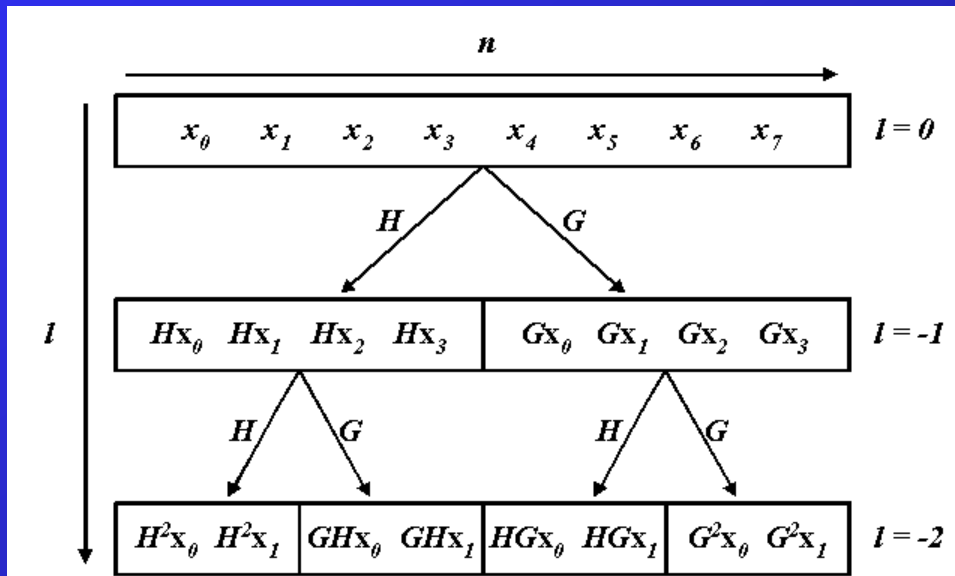
Wavelet Packet Decomposition (cont'd)

$\psi_0(t) \equiv \varphi(t)$ (characteristic (scaling) function)

$\psi(t) \equiv \sqrt{2} \sum_k g_k \varphi(2t - k) = \psi_1(t)$ (mother wavelet)

$\{I_{\ell,n}\} = \{[2^\ell n, 2^\ell(n+1)) : (\ell,n) \in E\}$ has to be a disjoint cover of $[0,1)$

$B = \{\psi_{\ell,n,k}(t) = 2^{\ell/2} \psi_n(2^\ell t - k) : (\ell,n) \in E, k \in \mathbb{Z}\}$



$E \subset \{(-1,0), (-1,1), (-2,0),$
 $(-2,1), (-2,2), (-2,3)\}$

$O(rN \log_2 N)$

Local Trigonometric Decomposition (LTD)

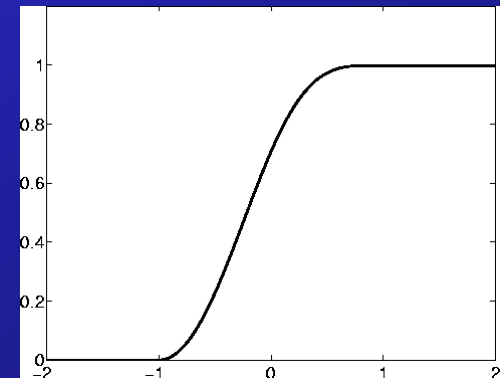
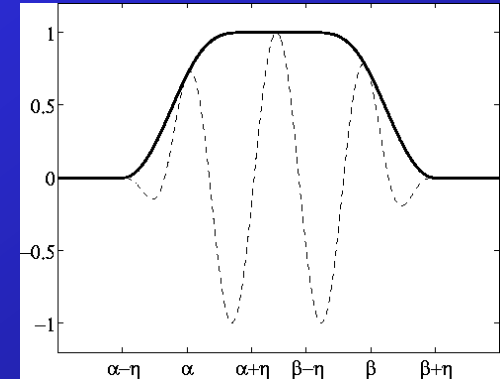
$$\Psi_j^k(t) \equiv b_j(t) F_j^k(t) \quad (\text{local trigonometric basis function})$$

$$b_j(t) \equiv r\left(\frac{t-a_j}{\eta}\right) r\left(\frac{a_{j+1}-t}{\eta}\right) \quad (\text{window function})$$

$$I_j \equiv [a_j, a_{j+1}) \quad R = \bigcup_{j \in \mathbb{Z}} I_j \quad 0 < \eta \leq \frac{|I_j|}{2}, \quad \forall j$$

$$r(t) = \begin{cases} 0, & \text{if } t \leq -1 \\ 1, & \text{if } t > 1 \end{cases} \quad (\text{"right cut-off function"})$$

$$F_j^k(t) \equiv \frac{\sqrt{2}}{\sqrt{|I_j|}} \sin\left(\pi\left(k + \frac{1}{2}\right) \frac{t-a_j}{|I_j|}\right) \quad (\text{DST-IV trigonometric basis function})$$

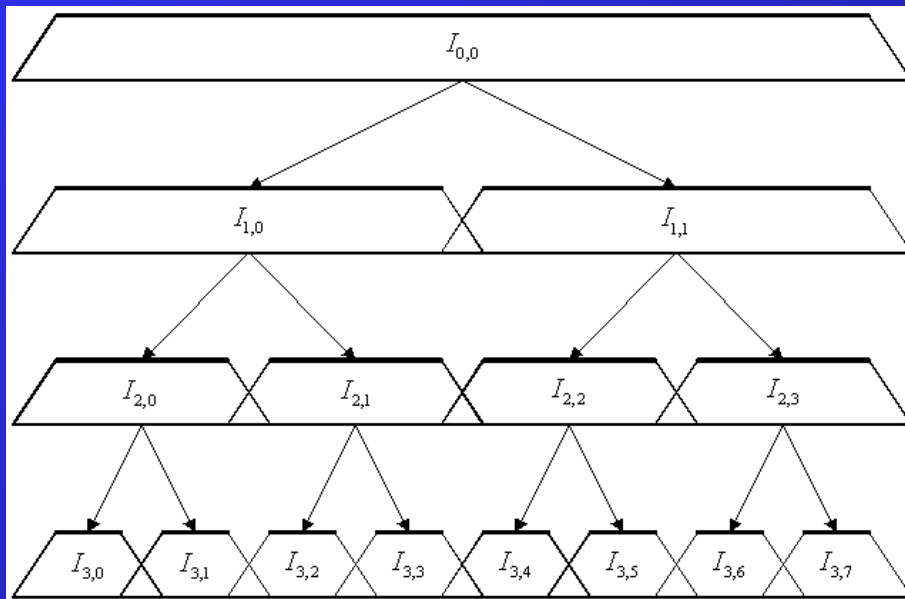


Local Trigonometric Decomposition (cont'd)

$$f(t) = \sum_j P_{I_j} f(t) = \sum_j c_j^k \Psi_j^k(t)$$

$$c_j^k \equiv \langle f(t), \Psi_j^k(t) \rangle = \langle \mathbf{T}f(t), \mathbf{1}_{I_j} \Psi_j^k(t) \rangle \quad (\text{expansion coefficients})$$

$$\mathbf{1}_{I_j} \equiv \begin{cases} 1, & t \in I_j \\ 0, & t \notin I_j \end{cases} \quad (\text{indicator function})$$

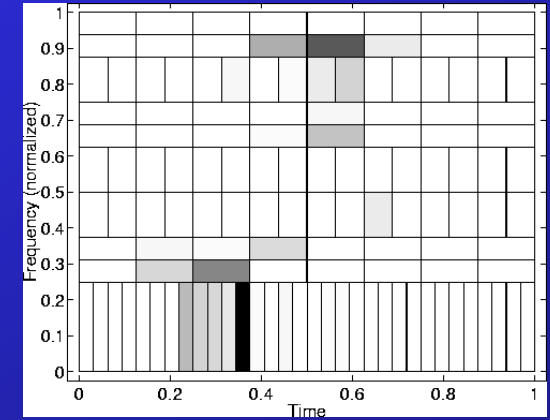
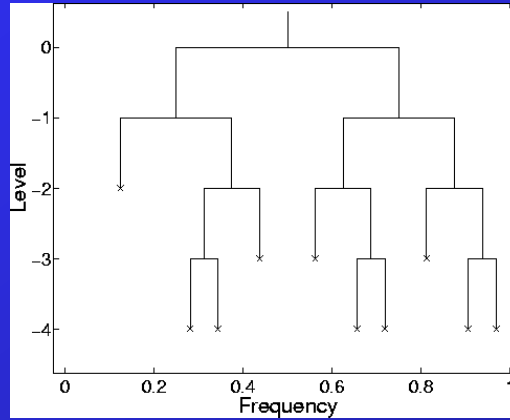
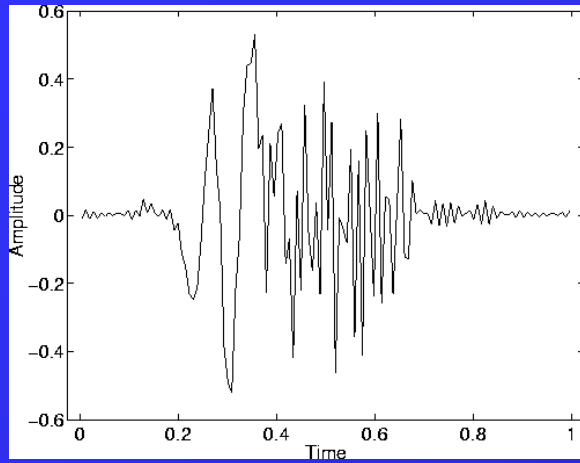


$$O(LN \log_2 N)$$

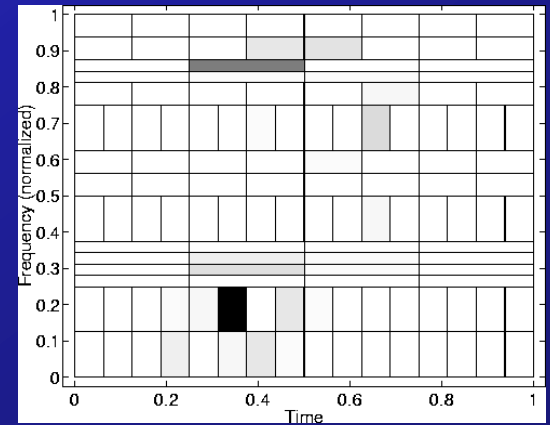
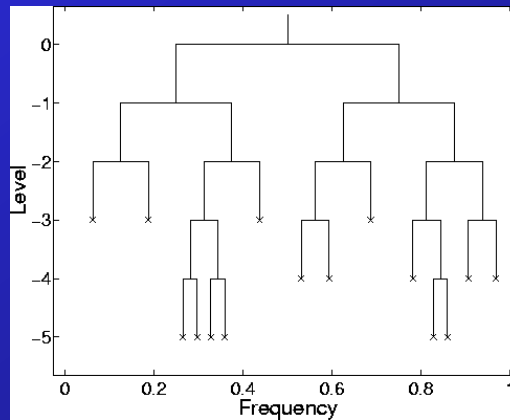
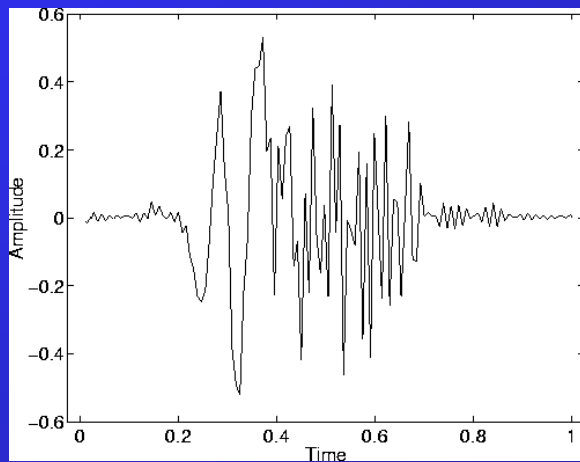
Joint Time-Frequency Representations (5)

Shift-Invariance

- Wavelet Packet Decomposition (WPD) is translation-variant



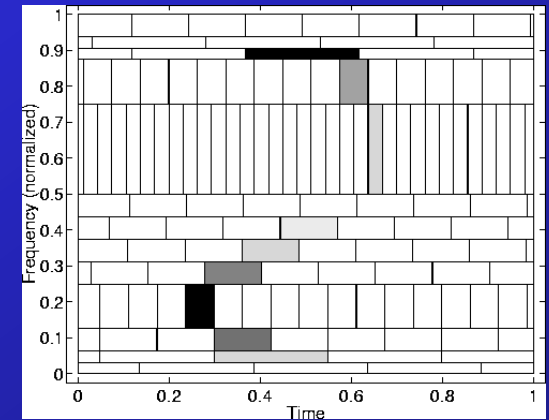
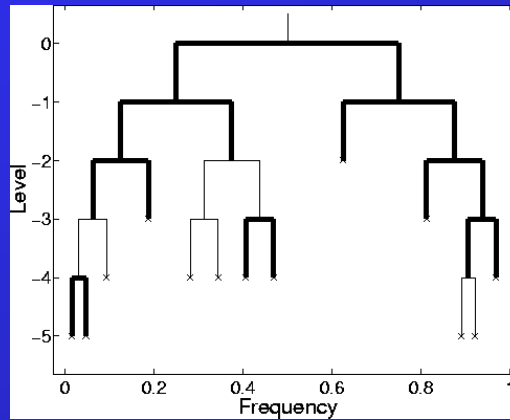
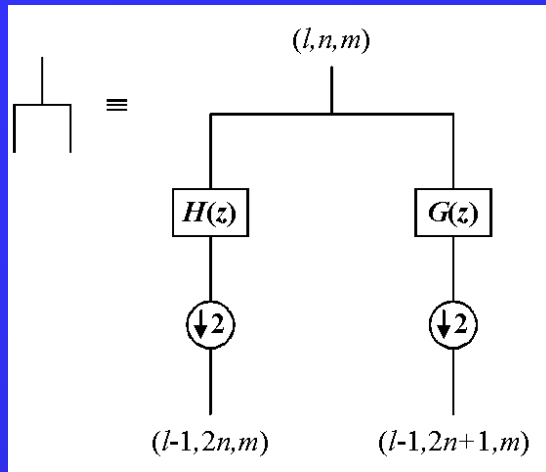
Entropy = 2.84



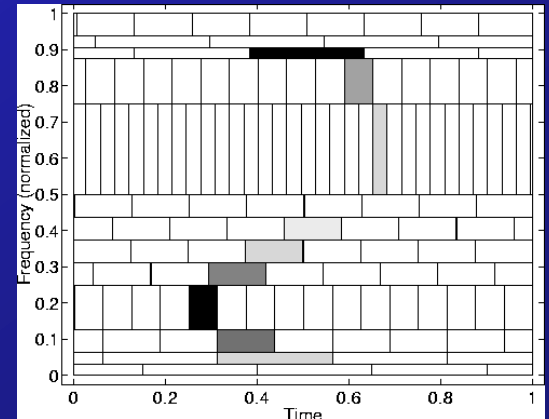
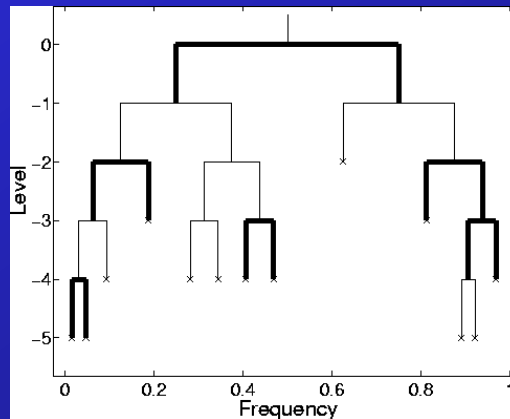
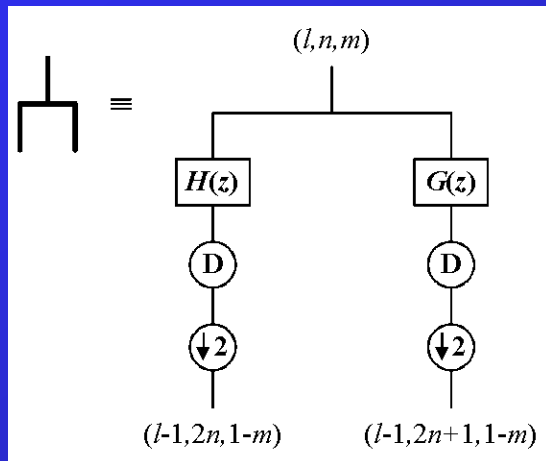
Entropy = 2.59

Shift-Invariance (cont'd)

- Shift-Invariant Wavelet Packet Decomposition (SIWPD) possesses shift-invariance property and characterizes by lower information cost (Cohen *et. al.*)



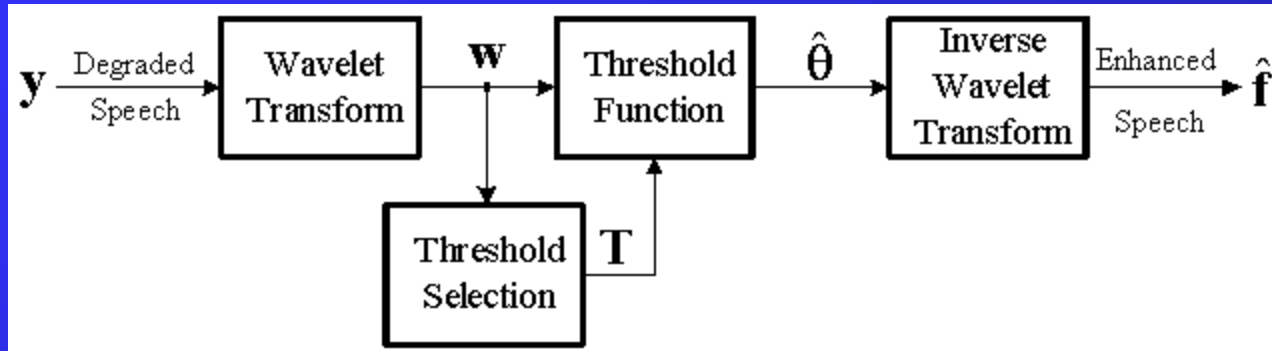
Entropy = 1.92



Entropy = 1.92

Wavelet-Based Denoising Techniques (1)

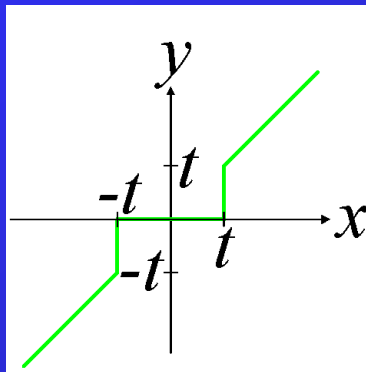
The Donoho-Johnstone Algorithm (1994/5)



$$\mathbf{w} = \{w_{\ell,n,k}\}$$

$$\mathbf{T} = \{t_{\ell,n,k}\}$$

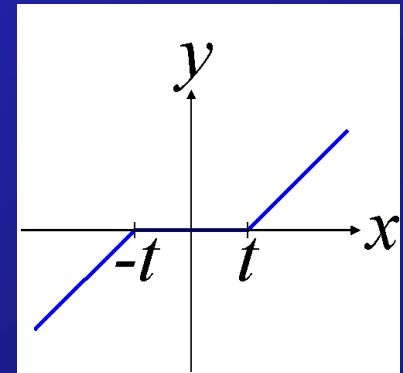
$$\hat{\boldsymbol{\theta}} = \{\hat{\theta}_{\ell,n,k}\}$$



$$y = \eta_h(x, t)$$

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

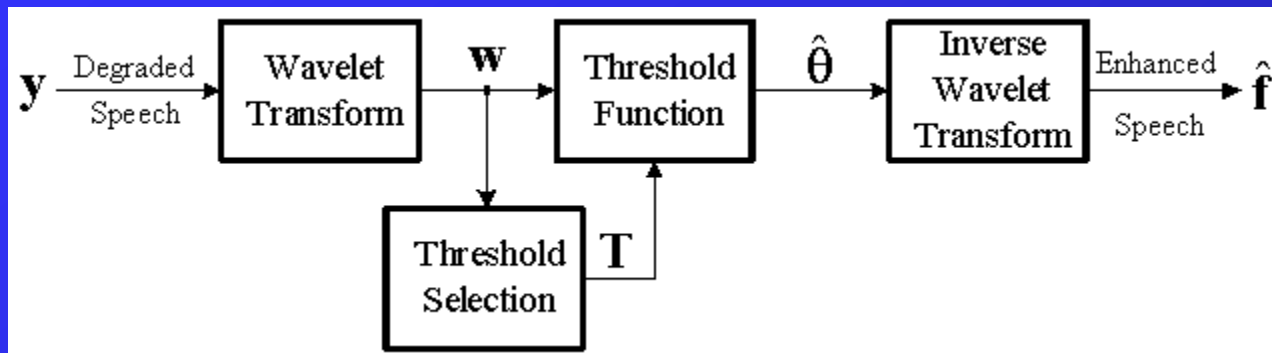
$$\eta_s(x, t) = \text{sign}(x) \cdot \begin{cases} |x| - t, & |x| > t \\ 0, & |x| \leq t \end{cases}$$



$$y = \eta_s(x, t)$$

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

The Donoho-Johnstone Algorithm (cont'd.)



$$\mathbf{W} = \{w_{\ell,n,k}\}$$

$$\mathbf{T} = \{t_{\ell,n,k}\}$$

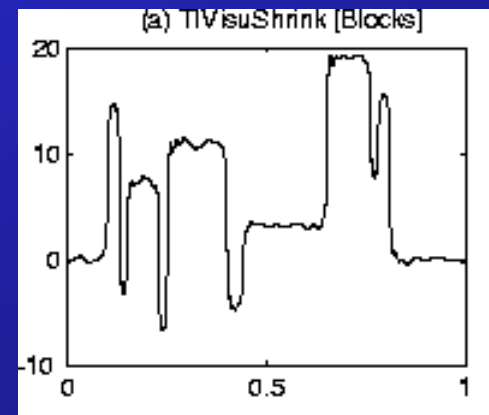
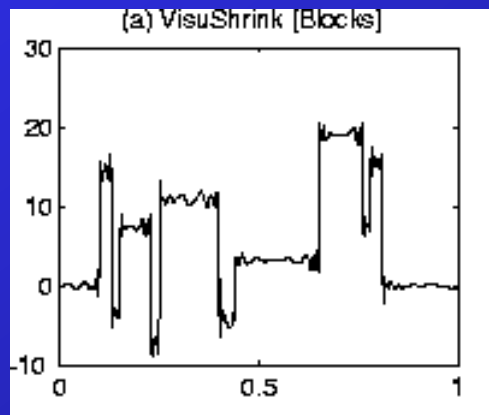
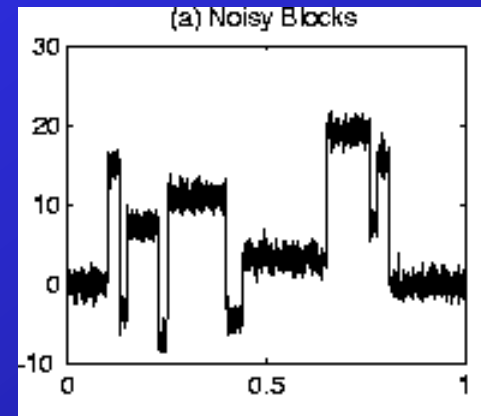
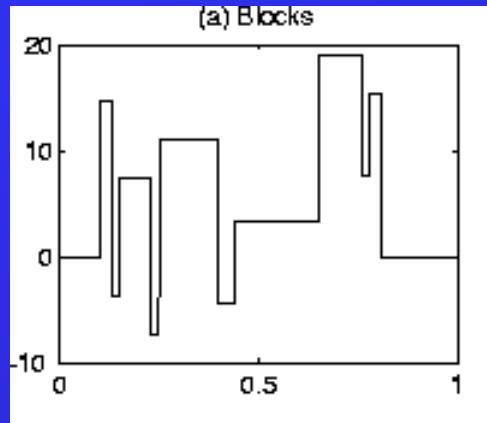
$$\hat{\boldsymbol{\theta}} = \{\hat{\theta}_{\ell,n,k}\}$$

$$\text{RiskShrink} : t_{\ell,n,k} = \lambda(\ell)\sigma$$

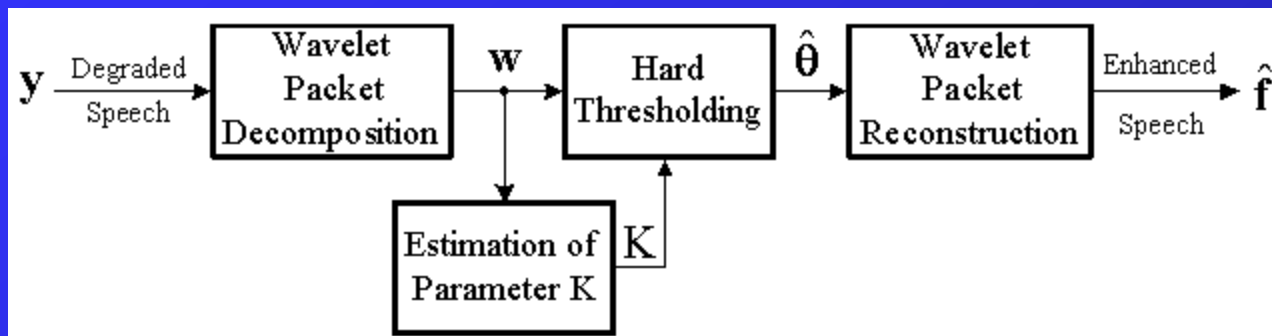
$$\text{VisuShrink} : t_{\ell,n,k} = \lambda_d(\ell)\sigma = \sigma\sqrt{2(\ell+J)\ln 2}$$

SureShrink : *adaptive threshold selection based on Stein unbiased estimate of risk*

Coifman-Donoho Translation-Invariant Denoising (1994/5)



Saito Adaptive Estimator (1994)



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

$$\mathcal{L}(\mathbf{f}, \Gamma_p, p) = \mathcal{L}(p) + \mathcal{L}(\Gamma_p | p) + \mathcal{L}(\mathbf{f} | \Gamma_p, p)$$

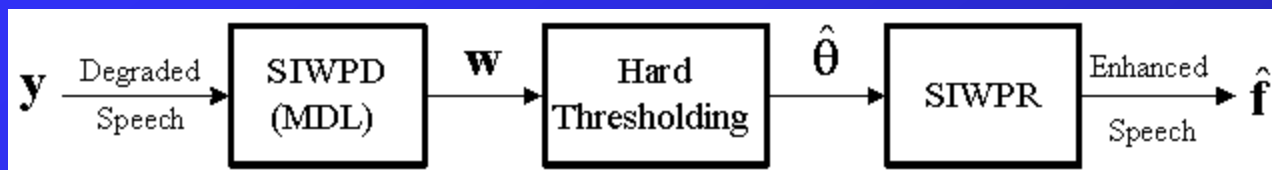
$$K = \arg \min_{0 \leq k \leq N-1} \{MDL(\mathbf{w}, k)\}$$

$$MDL(\mathbf{w}, k) = \frac{3}{2} K \log_2 N + \frac{N}{2} \log_2 \|\mathbf{w} - \eta^{(k)}(\mathbf{w})\|_2^2$$

$\eta^{(k)}$ – hard thresholding operation, which keeps the k largest (in absolute value) elements intact and sets all other elements to zero

$$\hat{\boldsymbol{\theta}} = \eta^{(K)}(\mathbf{w})$$

Cohen-Raz-Malah Adaptive Estimator (1998)



$$\mathbf{w} = \{w_{\ell,n,k}^{(m)}\} \quad \mathbf{w}_{\ell,n}^{(m)} = \{w_{\ell,n,k}^{(m)}\}_{k=0}^{2^{(\ell+J)}-1} \quad \hat{\boldsymbol{\theta}} = \{\hat{\theta}_{\ell,n,k}^{(m)}\}$$

$$K_{opt} = \#\{(w_{\ell,n,k}^{(m)})^2 > 3\sigma^2 \ln N\}$$

$$MDL(\mathbf{w}_{\ell,n}^{(m)}) = 3 + \frac{1}{2\sigma^2 \ln 2} \sum_{k=0}^{2^{(\ell+J)}-1} \min\{(w_{\ell,n,k}^{(m)})^2, 3\sigma^2 \ln N\}$$

$$\hat{\boldsymbol{\theta}} = \eta_h(\mathbf{w}, 3\sigma^2 \ln N)$$

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)

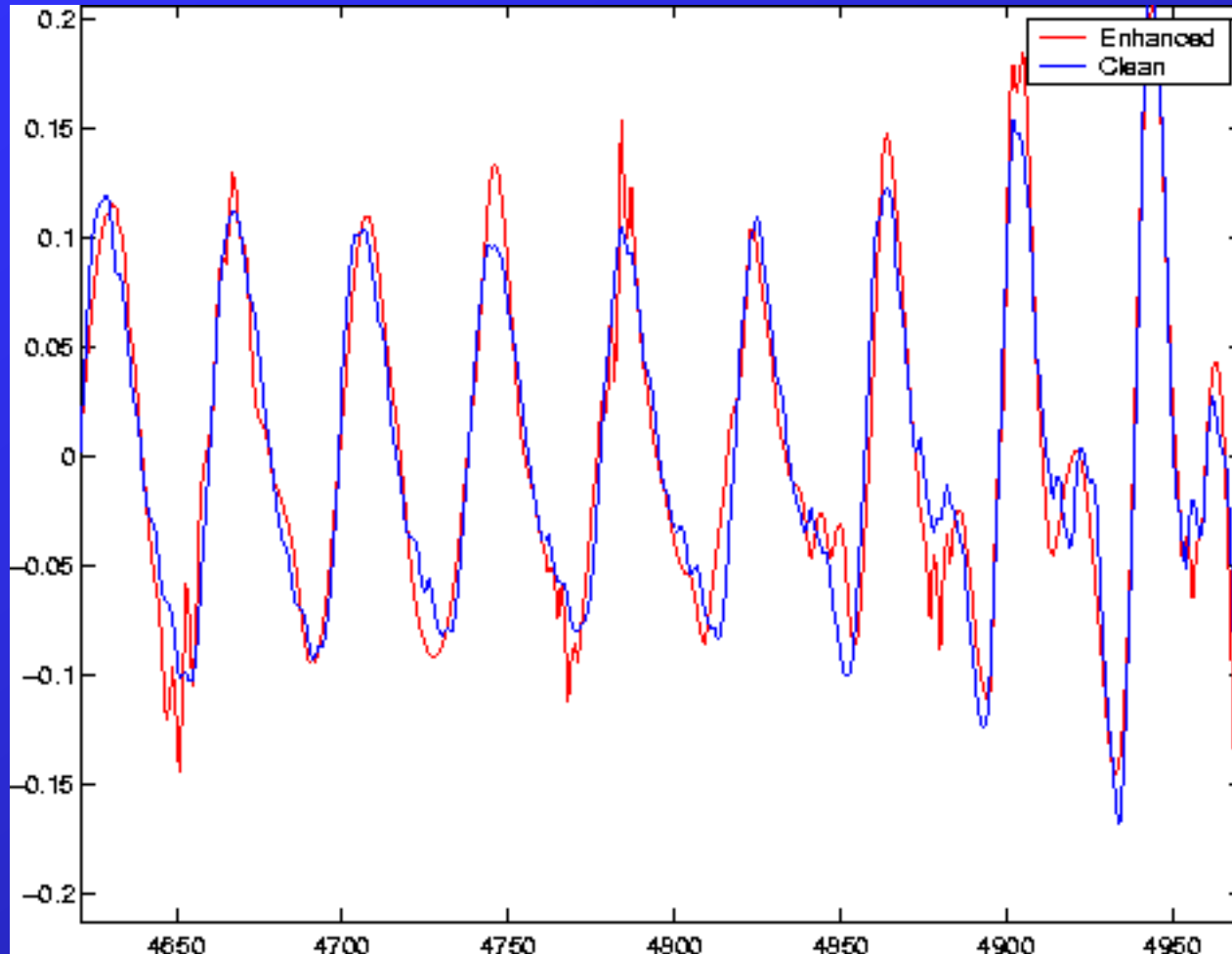
- Daubechies nearly symmetric mother wavelet of 8'th order (DNS(8))
- Entropy-based best-basis selection ($L=6$)
- Soft-thresholding

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
<i>VisuShrink</i>	10	6.68	9.47 	10.08	6.76	6.88
<i>RiskShrink</i>	10	6.68	9.47 	12.69	8.13	6.47
<i>SureShrink</i>	10	6.68	9.47 	14.73	9.28	6.35
<i>Saito</i>	10	6.68	9.47 	9.55	6.52	7.44
<i>Cohen</i>	10	6.68	9.47 	11.28	7.62	6.85
<i>Wiener</i>	10	6.68	9.47 	13.35	8.63	7.34

- Thresholding-based algorithms – oversmoothing and artifacts

Oversmoothing and Artifacts in Thresholding-Based Denoising



Over-smoothing speech enhanced by SRSK sink

Suppression of Artifacts

Increasing temporal support of basis functions:

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

Influence of Cost Function

- DNS(8)
- Best-basis selection algorithm ($L=6$)
- Wiener estimator

Cost function	Input SNR	Output SNR
<i>Entropy</i>	10	13.35
<i>Log-Energy</i>	10	13.53
l^1	10	13.49

- There is no significant difference in the quality of enhanced speech

<i>Full Subband</i>	10	13.55
---------------------	----	-------

- Full subband WPD-based denoising attains the highest SNR

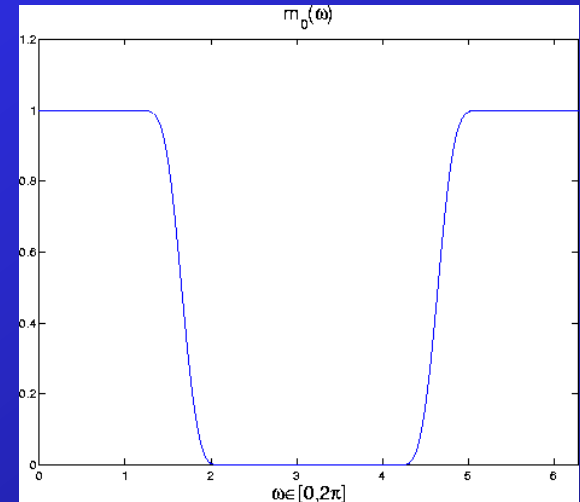
Increasing Temporal Support of Mother Wavelet

- Meyer mother wavelet

$$m_0(\omega) = \begin{cases} 1, & |\omega| \leq \frac{\pi}{3} \\ \cos\left[\frac{\pi}{2} v\left(\frac{3}{\pi} |\omega| - 1\right)\right], & \frac{\pi}{3} \leq |\omega| \leq \frac{2\pi}{3} \\ 0, & |\omega| \geq \frac{2\pi}{3} \end{cases}$$

$v(x)$ – auxiliary function, $x \in [0,1]$

$$v(x) = 35x^4 - 84x^5 + 70x^6 - 20x^7$$

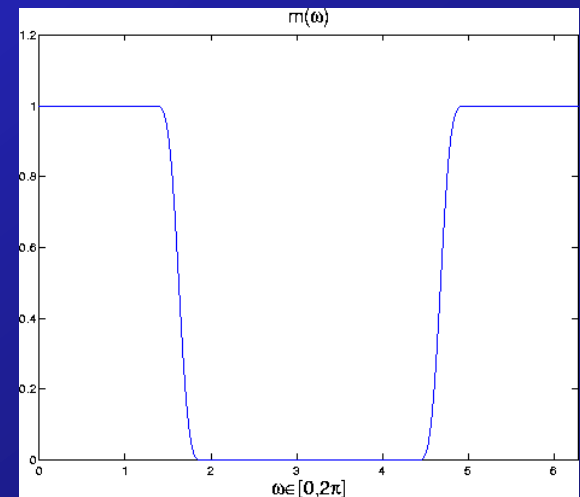


- Generalized Meyer mother wavelet

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

r – roll-off

$$m_0(\omega) = m(\omega) \Big|_{r=\frac{1}{3}}$$



$$r = \frac{1}{5}$$

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
----------------	-----------	------------

<i>SureShrink</i>	10	14.73
-------------------	----	-------

<i>Wiener</i>	10	13.35
---------------	----	-------

Generalized Meyer mother wavelet

Estimator type	N, r	Output SNR
----------------	--------	------------

<i>SureShrink</i>	32, 1/3	14.81
-------------------	---------	-------

<i>SureShrink</i>	64, 1/5	14.92
-------------------	---------	-------

<i>Wiener</i>	32, 1/3	13.53
---------------	---------	-------

<i>Wiener</i>	64, 1/5	13.63
---------------	---------	-------

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

SIWPD-Based Denoising

- Generalized Meyer mother wavelet ($N = 64, r = 0.2$)
- Full subband decomposition ($L=6$)
- Wiener estimator

Speaker	Decomposition type	Input SNR	Output SNR
Female	WPD	10	13.63
Female	SIWPD	10	13.77

Speaker	Decomposition type	Input SNR	Output SNR
Male	WPD	10	12.79
Male	SIWPD	10	12.78

Test signals of Donoho:

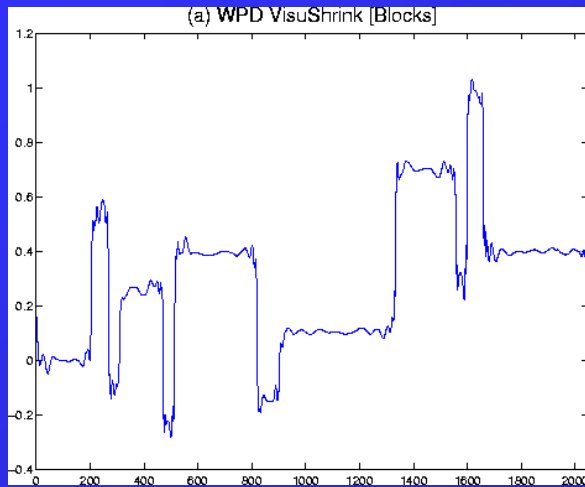
- DNS(8) mother wavelet
- Entropy-based best-basis selection ($L=6$)
- SureShrink estimator

Test Signal	Decomposition type	Input SNR	Output SNR
Blocks	WPD	17	18.23
Blocks	SIWPD	17	18.24

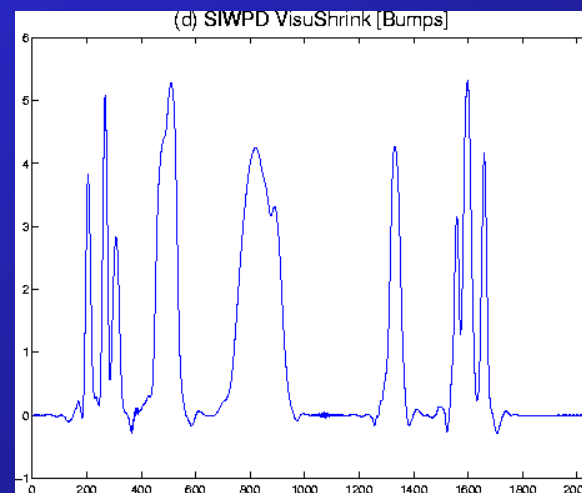
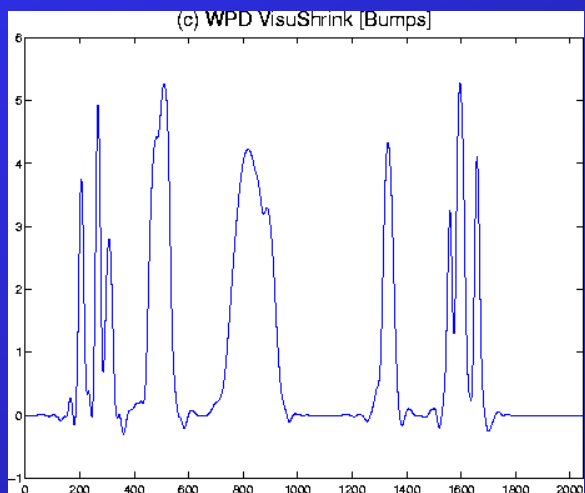
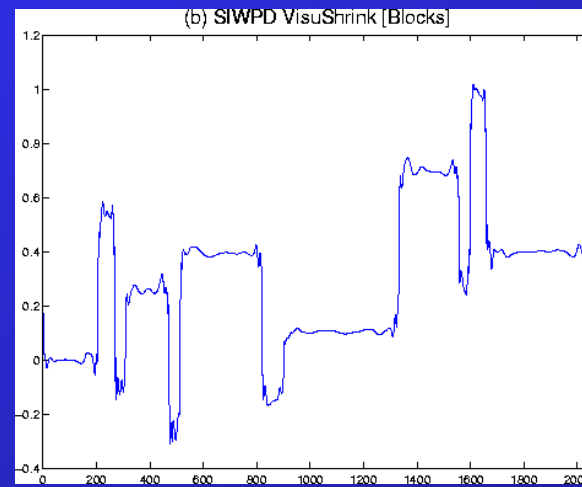
Test Signal	Decomposition type	Input SNR	Output SNR
Bumps	WPD	17	21.34
Bumps	SIWPD	17	21.39

SIWPD-Based Denoising (cont'd)

WPD-based VisuShrink



SIWPD-based VisuShrink




- Property of shift-invariance does not improve denoising performance


Framing

- Denoising without framing
- Full subband decomposition ($L=6$)

Speaker	Input SNR	Output SNR
Female	10	13.63




Speaker	Input SNR	Output SNR
Male	10	12.79




- Framing (Hanning window, 50% overlapping, 256 samples per frame)
- Full subband decomposition ($L=5$)

Speaker	Input SNR	Output SNR
Female	10	15.69



Speaker	Input SNR	Output SNR
Male	10	14.95



- Framing improves resulting SNR
- Smoothing of gains fluctuations is needed

Utilization of the “Decision Directed” A Priori SNR Estimation

- Framing: optimal frame length – 256 samples
- Tracking a priori SNR for decomposition tree terminal nodes: the full subband decomposition is the optimal choice ($L=J$)

$$G(\mathbf{w}_{\ell,n}(j), \hat{\mathbf{z}}_{\ell,n}(j)) = \frac{\hat{\xi}_{\ell,n}(j)}{\hat{\xi}_{\ell,n}(j) + 1}$$

$$\xi_{\ell,n}(j) = \frac{\|\boldsymbol{\theta}_{\ell,n}(j)\|_2^2}{\|\hat{\mathbf{z}}_{\ell,n}(j)\|_2^2} \quad (\text{a priori SNR})$$

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

$$\gamma_{\ell,n}(j) = \frac{\|\mathbf{w}_{\ell,n}(j)\|_2^2}{\|\hat{\mathbf{z}}_{\ell,n}(j)\|_2^2} \quad (\text{a posteriori SNR})$$

$$\hat{\xi}_{\ell,n}(1) = \alpha + (1-\alpha)\eta_s(\gamma_{\ell,n}(1), 1)$$

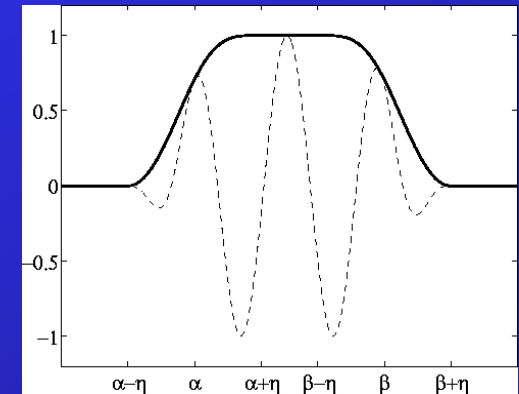
Proposed WPD-Based Speech Denoising Algorithm

- Wiener estimator, combined with the “decision directed” a priori SNR estimation ($\alpha=0.9$, 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

LTD-Based Denoising of Speech (1)

- Cosine Packet Decomposition (CPD) (DCT-IV) with Wickerhauser symmetric bell ($\eta = 6$)
- 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	WPD-based denoising
<i>VisuShrink</i>	10	6.68	9.47 	10.34	6.91	6.81	 10.08
<i>SureShrink</i>	10	6.68	9.47 	14.48	9.06	5.9	 14.73
<i>Wiener</i>	10	6.68	9.47 	12.92	8.05	6.93	 13.35

- Thresholding-based algorithms – oversmoothing and artifacts
- Artifacts, that characterize LTD-based denoising, are less annoying

Influence of Cost Function

- There is no significant difference in the quality of enhanced speech
- Full “subsegment” CPD-based denoising attains the highest SNR

Temporal Support and Frequency Localization

$$a_{j+1} - a_j \geq 2\eta > 0$$

$$\eta_{\max} = 2^{J-L} / 2$$

$$\eta_{\max} |_{J=14, L=6} = 128$$

- CPD with Wickerhauser symmetric bell ($\eta = \eta_{\max}$)
- Full “subsegment” decomposition ($L=6$)
- Wiener estimator

Speaker	η	Input SNR	Output SNR
Female	6	10	12.99
Female	128	10	13.24

Speaker	η	Input SNR	Output SNR
Male	6	10	12.73
Male	128	10	12.89

Utilization of the “Decision Directed” A Priori SNR Estimation

- Tracking a priori SNR for decomposition tree nodes on the same decomposition level
- The full “subsegment” decomposition is the optimal choice ($L=J$)

$$G(w_{\ell,n,k}, \hat{z}_{\ell,n,k}) = \frac{\hat{\xi}_{\ell,n,k}}{\hat{\xi}_{\ell,n,k} + 1}$$

$$\xi_{\ell,n,k} = \frac{|\theta_{\ell,n,k}|^2}{|\hat{z}_{\ell,n,k}|^2} \quad (\text{a priori SNR})$$













$$\hat{\xi}_{\ell,n,k} = \alpha \frac{|\hat{\theta}_{\ell,n-1,k}|^2}{|\hat{z}_{\ell,n-1,k}|^2} + (1-\alpha)\eta_s(\gamma_{\ell,n,k}, 1)$$

$$\gamma_{\ell,n,k} = \frac{|w_{\ell,n,k}|^2}{|\hat{z}_{\ell,n,k}|^2} \quad (\text{a posteriori SNR})$$

$$\hat{\xi}_{\ell,0,k} = \alpha + (1-\alpha)\eta_s(\gamma_{\ell,0,k}, 1)$$

Proposed LTD-Based Speech Denoising Algorithm













- Wiener estimator, combined with the “decision directed” a priori SNR estimation ($\alpha=0.9$)
- Full “subsegment” decomposition ($L=6$)
- Improved frequency localization and increased time support ($\eta = \eta_{\max} = 2^{J-L-1}$)

#	Speaker	Decomposition type	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
1	Female	CPD	 10	6.06	11.51 	16.69	9.17	8.56
1	Male	CPD	 10	5.96	11.53 	15.44	8.1	9.69
2	Female	CPD	 10	6.68	9.47 	15.13	9.43	6.76
2	Male	CPD	 10	6.73	9 	14.37	9.06	6.39
3	Female	CPD	 10	6.17	11.11 	15.94	8.59	9.2
3	Male	CPD	 10	5.92	11.51 	15.24	7.68	9.99

WPD Applied to DCT-I Coefficients

Proposed Speech Denoising Algorithm

- Wiener estimator, combined with the “decision directed” a priori SNR estimation ($\alpha=0.9$)
- Full “subsegment” decomposition ($L=6$)
- DNS mother wavelet (4'th order)

#	Speaker	Decomposition type	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
1	Female	WPD(DCT) 	10	6.06	11.51 	16.49	9.04	8.81
1	Male	WPD(DCT) 	10	5.96	11.53 	15.31	8.07	9.73
2	Female	WPD(DCT) 	10	6.68	9.47 	15.02	9.35	6.95
2	Male	WPD(DCT) 	10	6.73	9 	14.22	9	6.49
3	Female	WPD(DCT) 	10	6.17	11.11 	15.83	8.42	9.35
3	Male	WPD(DCT) 	10	5.92	11.51 	15.16	7.53	10.09

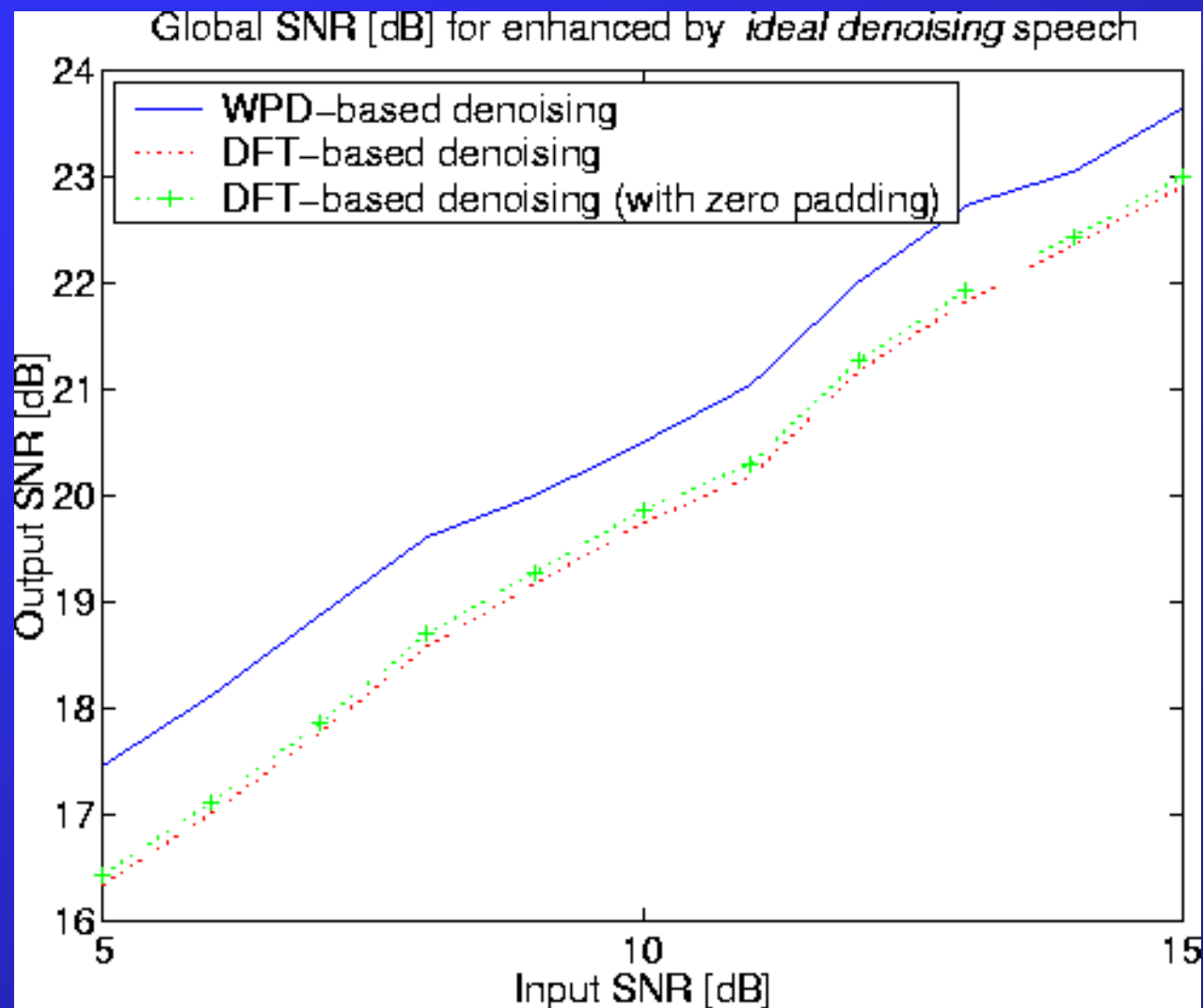
“Ideal” Denoising (1)

- Decision directed approach smoothes gains fluctuations from frame to frame, caused by fluctuations of noise squared spectral amplitude
- “Ideal” denoising –assuming prior knowledge of noise squared-spectral amplitude exact value
- Results of “ideal” denoising:
 - Denoising, based on WPD : the proposed algorithm with $\alpha=0$
 - Denoising, based on CPD (or WPD applied to DCT coefficients):
 - * Hanning window, 512 samples per frame with 25% overlapping
 - * Decomposition with $L=1$

Estimator type, decomposition type	Input SNR	Output SNR
<i>Wiener, WPD</i>	10	20.48
<i>Wiener, DCT</i>	10	20.45
<i>Wiener, CPD</i>	10	20.36

Estimator type, decomposition type	Input SNR	Output SNR
<i>Wiener, WPD(DCT)</i>	10	20.5
<i>Wiener, DFT_{2N}</i>	10	19.86
<i>E-M, DFT_{2N}</i>	10	19.51

DFT-Based “Ideal” Denoising vs. Real-Valued Transforms-Based “Ideal” Denoising



Advantages of Real-Valued Transform-Based “Ideal” Denoising

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

Temporal Support and Frequency Localization

- Increasing L and decreasing r improves performance of the WPD-based “ideal” denoising
- Increasing η improves performance of the LTD-based “ideal” denoising

A Comparative Performance Analysis (1)

- Results of practical denoising:

Estimator type, decomposition type	Input SNR	Output SNR
<i>Wiener, WPD</i>	10	17.37
<i>Wiener, CPD</i>	10	16.69
<i>Wiener, WPD(DCT)</i>	10	16.49

Estimator type, decomposition type	Input SNR	Output SNR
<i>Wiener, DFT</i>	10	17.83
<i>E-M, DFT</i>	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
- 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 quantitative measures are close to benchmark DFT-based speech denoising algorithms
- Proposed WPD-based speech denoising algorithm attains relatively high SNR and 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
- We have presented results of theoretical investigations

Future Work

- Techniques for better estimation of noise spectral components
- Combined LTD-based segmentation and denoising
- Applying SILTD for speaker identification/verification

The End