Source Separation Tutorial Mini-Series I

Speech enhancement algorithms

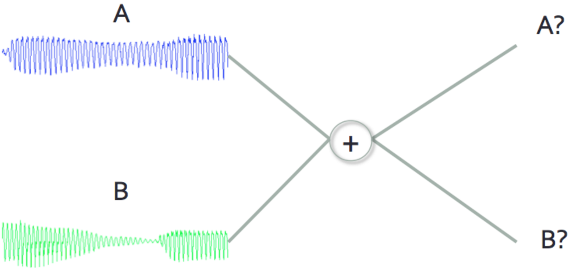
Eunjoon Cho

Stanford University, EE

April 2nd, 2013

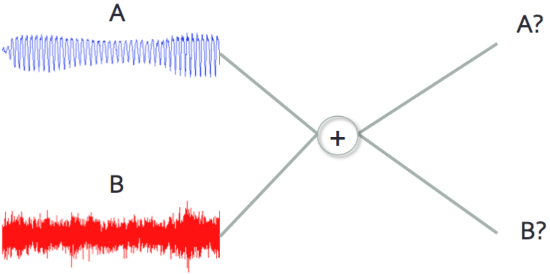
Speech enhancement: A source separation perspective

- Source separation: Decoupling of two or more sources with no, little or some prior information.



Speech enhancement: A source separation perspective

- Source separation: Decoupling of two or more sources with no, little or some prior information.
- Speech enhancement is a natural application for source separation.



Speech enhancement: The application perspective

- Goal: Increase the intelligibility/quality of noisy speech.

Speech enhancement: The application perspective

- Goal: Increase the intelligibility/quality of noisy speech.
- Practical constraints.

Speech enhancement: The application perspective

- Goal: Increase the intelligibility/quality of noisy speech.
- Practical constraints.
  - Computationally efficient: Real-time applications on mobile phones, teleconferences.

Speech enhancement: The application perspective

- Goal: Increase the intelligibility/quality of noisy speech.
- Practical constraints.
  - Computationally efficient: Real-time applications on mobile phones, teleconferences.
  - A solution independent of the noise environment.

Speech enhancement: The application perspective

- Goal: Increase the intelligibility/quality of noisy speech.
- Practical constraints.
  - Computationally efficient: Real-time applications on mobile phones, teleconferences.
  - A solution independent of the noise environment.
  - Stronger emphasis on reconstructing speech (different objective/subjective measures).

Objective

- Present some of the well-established methods in the speech enhancement literature and discuss the relationship between them.

Objective

- Present some of the well-established methods in the speech enhancement literature and discuss the relationship between them.
- Shed insight on how such methods differ in approach and assumptions with methods that rely on matrix factorization and/or prior training of sources.

Objective

- Present some of the well-established methods in the speech enhancement literature and discuss the relationship between them.
- Shed insight on how such methods differ in approach and assumptions with methods that rely on matrix factorization and/or prior training of sources.
- Working code that can act as baselines for any speech enhancement work.

Speech enhancement: Model sources

Under-determined problem: $Y(\omega) = X(\omega) + D(\omega)$

Speech enhancement: Model sources

Under-determined problem: $Y(\omega) = X(\omega) + D(\omega)$

- Train dictionaries (bases) of noise and/or speech to use as prior.
    - Model of the noise can be inaccurate.
    - Training online can be computationally expensive.

Speech enhancement: Model sources

Under-determined problem: $Y(\omega) = X(\omega) + D(\omega)$

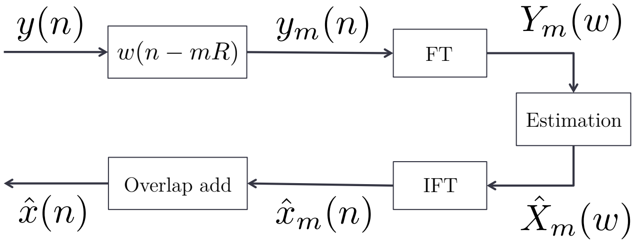
- Train dictionaries (bases) of noise and/or speech to use as prior.
  - Model of the noise can be inaccurate.
  - Training online can be computationally expensive.
- Assumption that noise varies more slowly compared to speech and that speech is temporally sparse.
  - Use voice activity detectors and estimate noise when there is no speech.
  - Keep track of the minimum level of spectrum at certain frequency.

Speech enhancement: Separate sources

- Based on an estimate of the noise, how can we estimate the speech.
  - Spectral subtraction [Bol79]
  - Wiener filtering: MMSE Estimator [LO79]
  - Spectral Amplitude MMSE Estimator [EM84]

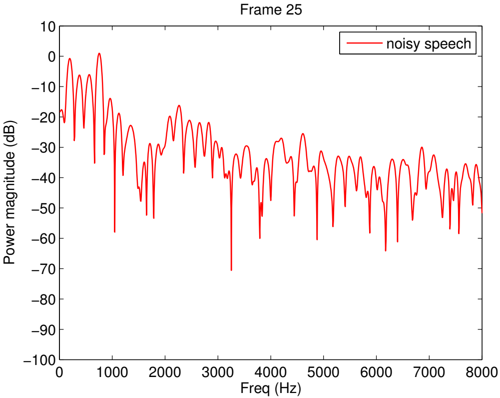
Frequency domain approaches

- Discuss approaches in the frequency domain: $Y(\omega) = X(\omega) + D(\omega)$
- Overall flow of STFT processing



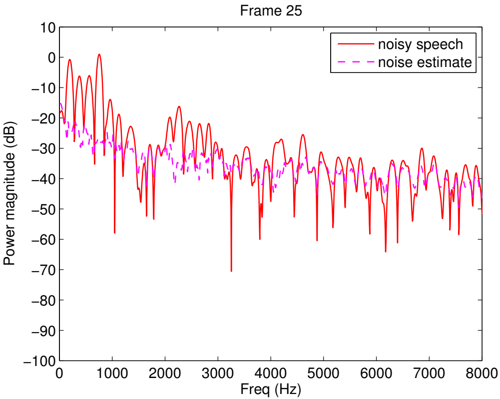
Spectral subtraction

$$|Y_m(\omega)|$$



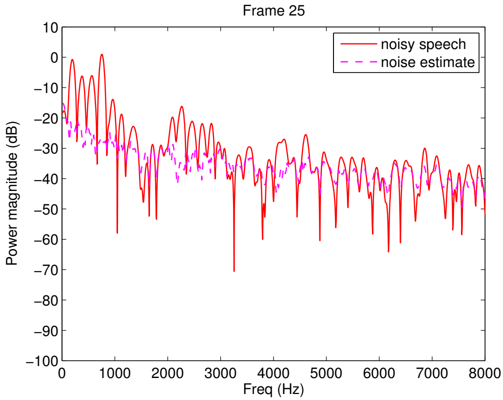
Spectral subtraction

$$|Y_m(\omega)|, |\hat{D}_m(\omega)|$$



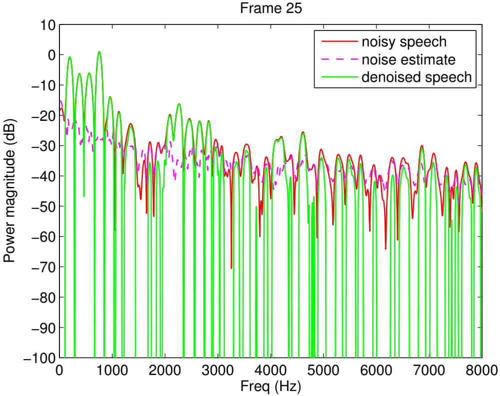
Spectral subtraction

$$|Y_m(\omega)|, E\left[|D_m(\omega)|\right]$$



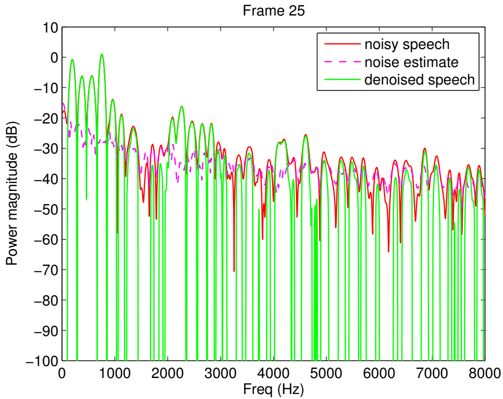
Spectral subtraction

$$|\hat{X}_m(\omega)| = |Y_m(\omega)| - E\left[|D_m(\omega)|\right]$$



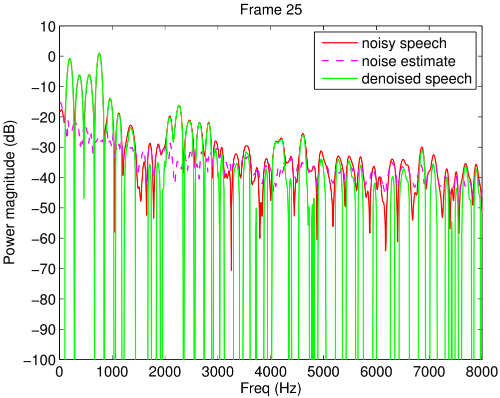
Spectral subtraction

$$|\hat{X}_m(\omega)| = \max\{|Y_m(\omega)| - E\left[|D_m(\omega)|\right], 0\}$$



Spectral subtraction

$$|\hat{X}_m(\omega)| = \max\{|Y_m(\omega)| - E\left[|D_m(\omega)|\right], 0\}$$
$$\angle\hat{X}_m(\omega) = \angle Y_m(\omega)$$



Power spectral subtraction

$$|\hat{X}_m(\omega)|^\alpha = \max\{|Y_m(\omega)|^\alpha - E\left[|D_m(\omega)|^\alpha\right], 0\}$$

Power spectral subtraction

$$|\hat{X}_m(\omega)|^\alpha = \max\{|Y_m(\omega)|^\alpha - E\left[|D_m(\omega)|^\alpha\right], 0\}$$

- The power spectral subtraction: $\alpha = 2$

$$|\hat{X}_m(\omega)|^2 = \max\{|Y_m(\omega)|^2 - E\left[|D_m(\omega)|^2\right], 0\}$$

Gain for power spectral subtraction

- From the power spectral subtraction,

$$|\hat{X}_m(\omega)|^2 = |Y_m(\omega)|^2 - E\left[|D_m(\omega)|^2\right]$$

the gain can be expressed as follows

$$H_m(\omega) = \frac{|\hat{X}_m(\omega)|}{|Y_m(\omega)|}$$

Gain for power spectral subtraction

- From the power spectral subtraction,

$$|\hat{X}_m(\omega)|^2 = |Y_m(\omega)|^2 - E\left[|D_m(\omega)|^2\right]$$

the gain can be expressed as follows

$$H_m(\omega) = \frac{|\hat{X}_m(\omega)|}{|Y_m(\omega)|} = \sqrt{\frac{|Y_m(\omega)|^2 - E\left[|D_m(\omega)|^2\right]}{|Y_m(\omega)|^2}} = \sqrt{\frac{\gamma(\omega) - 1}{\gamma(\omega)}}$$

Gain for power spectral subtraction

- From the power spectral subtraction,

$$|\hat{X}_m(\omega)|^2 = |Y_m(\omega)|^2 - E\left[|D_m(\omega)|^2\right]$$

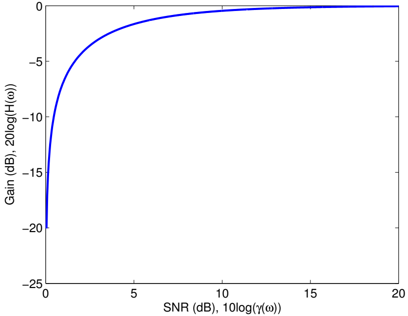
  the gain can be expressed as follows

$$H_m(\omega) = \frac{|\hat{X}_m(\omega)|}{|Y_m(\omega)|} = \sqrt{\frac{|Y_m(\omega)|^2 - E\left[|D_m(\omega)|^2\right]}{|Y_m(\omega)|^2}} = \sqrt{\frac{\gamma(\omega) - 1}{\gamma(\omega)}}$$

- $\gamma(\omega)$ is called the a-posteriori SNR. $\gamma(\omega) = \frac{|Y_m(\omega)|^2}{E[|D_m(\omega)|^2]}$

Gain for power spectral subtraction

$$H_m(\omega) = \sqrt{\frac{\gamma(\omega) - 1}{\gamma(\omega)}}$$

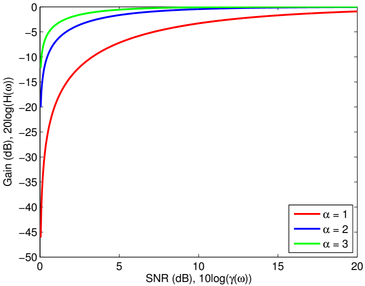


Gain for general spectral subtraction

$$|\hat{X}_m(\omega)|^\alpha = |Y_m(\omega)|^\alpha - E\left[|D_m(\omega)|^\alpha\right]$$

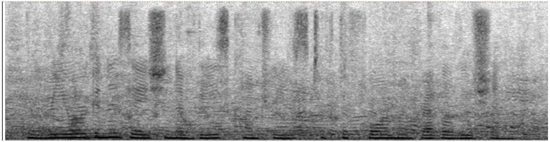
- Different gain functions for various α

$$H_m(\omega) = \frac{|\hat{X}_m(\omega)|}{|Y_m(\omega)|} = \left(\frac{\gamma(\omega)^{\alpha/2} - 1}{\gamma(\omega)^{\alpha/2}}\right)^{1/\alpha}$$

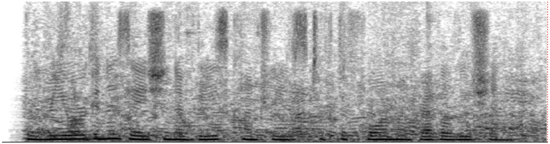


Example of spectral subtraction

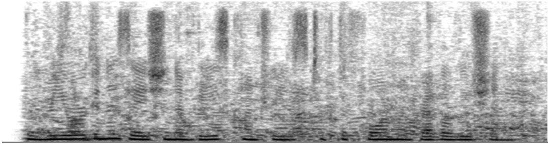
Noisy speech



Power spectral subtraction



Magnitude spectral subtraction



Musical noise

- Q. Why do we get musical noise?

Musical noise

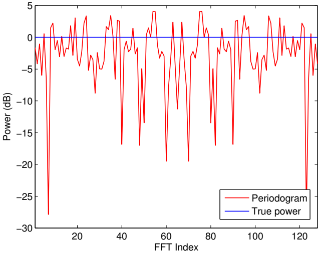
- Q. Why do we get musical noise?
- A1. Inaccurate estimate of unknown variables.
  - From $Y_m(\omega) = X_m(\omega) + D_m(\omega)$,

$$|X_m(\omega)|^2 = |Y_m(\omega)|^2 - |D_m(\omega)|^2 \\ - (X_m(\omega)D_m^*(\omega) + X_m^*(\omega)D_m(\omega))$$

  - $|D_m(\omega)|^2 \approx E\left[|D_m(\omega)|^2\right]$
  - $2Re\{X_m(\omega)D_m^*(\omega)\} \approx E\left[2Re\{X_m(\omega)D_m^*(\omega)\}\right] = 0$

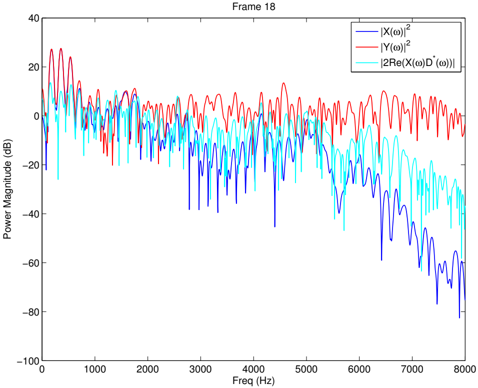
Noise estimation

- Simple method is to average power spectra when there is no speech activity $|\hat{D}_m(\omega)|^2 = E\left[|Y_m(\omega)|^2\right] = E\left[|D_m(\omega)|^2\right]$
  - Need an accurate voice activity detector (VAD)
  - Issues with non-stationary noise (babble noise) conditions
- Issues with $|\hat{D}_m(\omega)|^2 = E\left[|D_m(\omega)|^2\right]$
  - The noise power spectrum (periodogram) has high variance with respect to the underlying power spectral density



Cross term spectrum

- $2Re\{X_m(\omega)D_m^*(\omega)\}$ requires the phase information which is difficult to estimate

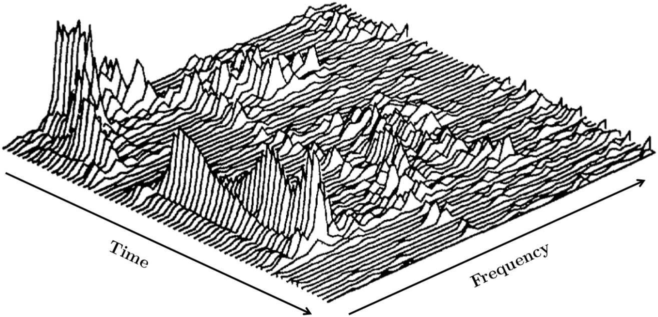


Musical noise

- Q. Why do we get musical noise?
- A1. In accurate estimate of unknown variables.
- A2. How we engineer situations when we have a bad estimate.
    - Half rectify negative values.

$$|\hat{X}_m(\omega)|^2 = \max\{|Y_m(\omega)|^2 - E\left[|D_m(\omega)|^2\right], 0\}$$

Artifacts from half-wave rectifying



1

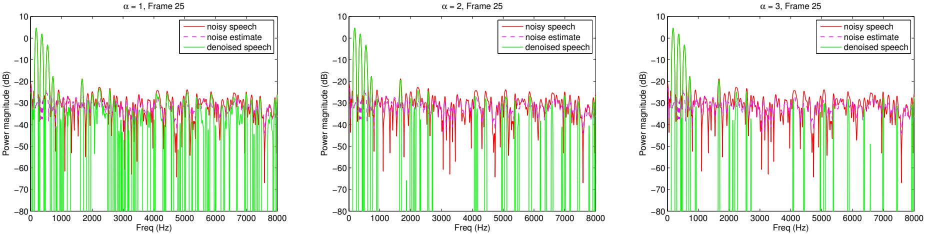
[1] Image from [Bol79]

Oversubtraction [BSM79]

- Over-subtract the noise estimate to reduce noise peaks

$$|\hat{X}_m(\omega)|^2 = |Y_m(\omega)|^2 - \alpha E\left[|D_m(\omega)|^2\right]$$

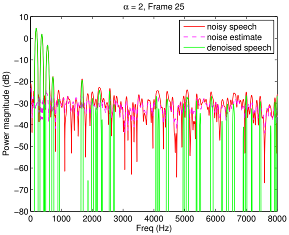
- Comes at the expense of attenuating the underlying signal

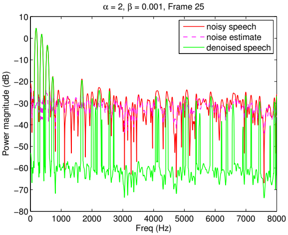


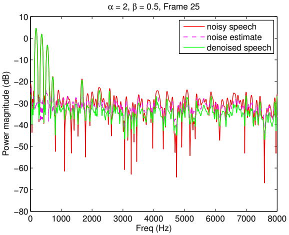
Oversubtraction [BSM79]

- Fill in valleys at frequencies to mask residue noise

$$|\hat{X}_m(\omega)|^2 = \max\{|Y_m(\omega)|^2 - \alpha E\left[|D_m(\omega)|^2\right], \\ \beta E\left[|D_m(\omega)|^2\right]\}$$

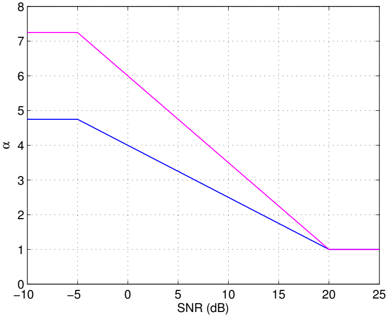






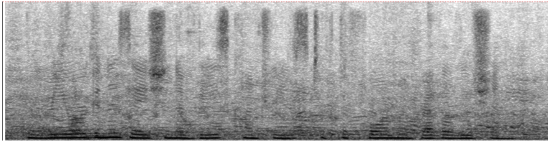
Oversubtraction [BSM79]

- α should be dependent on the frame segmental SNR (γ)
- Less attenuation (small α) for high SNR, and more attenuation (large α) for low SNR

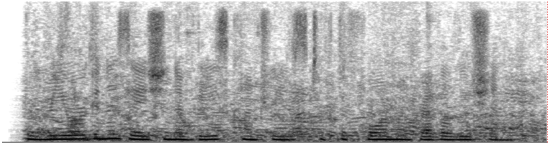


Example of over spectral subtraction

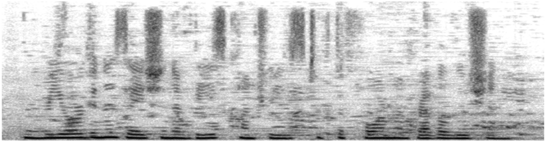
Noisy speech



Power spectral subtraction



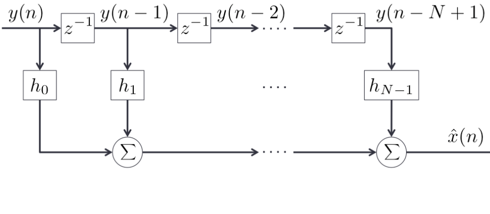
Spectral over subtraction: $\alpha = 15, \beta = 0.01$



Wiener filter

- Find optimal linear filter that outputs the desired signal (clean speech)

$$e(n) = x(n) - \hat{x}(n) = x(n) - \sum_{k=0}^{M-1} h_k y(n-k)$$



- Find $\boldsymbol{h}^*$ that minimizes $E\left[e^2(n)\right]$ by solving $\frac{\partial E\left[e^2(n)\right]}{\partial \boldsymbol{h}} = 0$
- $\boldsymbol{h}^* = \boldsymbol{R}_{\boldsymbol{yy}}^{-1}\boldsymbol{r}_{\boldsymbol{yx}} = (\boldsymbol{R}_{\boldsymbol{xx}} + \boldsymbol{R}_{\boldsymbol{dd}})^{-1}\boldsymbol{r}_{\boldsymbol{xx}}$

Wiener filter in frequency domain

- If we assume a non-causal IIR filter, using the convolution theorem, i.e., $x(n) * h(n) \leftrightarrow X(w)H(w)$

$$E(\omega) = X(w) - H(w)Y(w)$$

- If we minimize $E\left[|E(\omega)|^2\right]$ with respect to $H(\omega)$, we have

$$\begin{aligned} H(\omega) &= \frac{E\left[X(\omega)Y^*(\omega)\right]}{E\left[|Y(w)|^2\right]} = \frac{E\left[X(\omega)(X(\omega)^* + D(\omega)^*)\right]}{E\left[|Y(w)|^2\right]} \\ &= \frac{E\left[|X(\omega)|^2\right]}{E\left[|Y(\omega)|^2\right]} = \frac{E\left[|X(\omega)|^2\right]}{E\left[|X(\omega)|^2\right] + E\left[|D(\omega)|^2\right]} \end{aligned}$$

Parametric wiener filters

- Wiener filter

$$H(\omega) = \frac{E\left[|X(\omega)|^2\right]}{E\left[|Y(\omega)|^2\right]} = \frac{E\left[|X(\omega)|^2\right]}{E\left[|X(\omega)|^2\right] + E\left[|D(\omega)|^2\right]}$$

- More generally,

$$H(\omega) = \left(\frac{E\left[|X(\omega)|^2\right]}{E\left[|X(\omega)|^2\right] + \alpha E\left[|D(\omega)|^2\right]}\right)^{\beta}$$

Connection with spectral subtraction

$$H(\omega) = \left(\frac{E\left[|X(\omega)|^2\right]}{E\left[|X(\omega)|^2\right] + \alpha E\left[|D(\omega)|^2\right]} \right)^{\beta}$$

- $E\left[|X(\omega)|^2\right]$ is unknown

Connection with spectral subtraction

$$H(\omega) = \left(\frac{E\left[|X(\omega)|^2\right]}{E\left[|X(\omega)|^2\right] + \alpha E\left[|D(\omega)|^2\right]} \right)^\beta$$

- $E\left[|X(\omega)|^2\right]$ is unknown
- If $\alpha = 1$, $\beta = 1/2$, and $E\left[|X(\omega)|^2\right] = |\hat{X}(\omega)|^2$ then...

Connection with spectral subtraction

$$H(\omega) = \left(\frac{E\left[|X(\omega)|^2\right]}{E\left[|X(\omega)|^2\right] + \alpha E\left[|D(\omega)|^2\right]} \right)^{\beta}$$

- $E\left[|X(\omega)|^2\right]$ is unknown
- If $\alpha = 1$, $\beta = 1/2$, and $E\left[|X(\omega)|^2\right] = |\hat{X}(\omega)|^2$ then...

$$|\hat{X}(\omega)| = H(\omega)|Y(\omega)| = \sqrt{\frac{|\hat{X}(\omega)|^2}{|\hat{X}(\omega)|^2 + E\left[|D(\omega)|^2\right]}}|Y(\omega)|$$

Connection with spectral subtraction

$$H(\omega) = \left(\frac{E\left[|X(\omega)|^2\right]}{E\left[|X(\omega)|^2\right] + \alpha E\left[|D(\omega)|^2\right]} \right)^{\beta}$$

- $E\left[|X(\omega)|^2\right]$ is unknown
- If $\alpha = 1$, $\beta = 1/2$, and $E\left[|X(\omega)|^2\right] = |\hat{X}(\omega)|^2$ then...

$$|\hat{X}(\omega)| = H(\omega)|Y(\omega)| = \sqrt{\frac{|\hat{X}(\omega)|^2}{|\hat{X}(\omega)|^2 + E\left[|D(\omega)|^2\right]}}|Y(\omega)|$$

$$|\hat{X}(\omega)|^2(|\hat{X}(\omega)|^2 + E\left[|D(\omega)|^2\right]) = |\hat{X}(\omega)|^2|Y(\omega)|^2$$

Connection with spectral subtraction

$$H(\omega) = \left(\frac{E\left[|X(\omega)|^2\right]}{E\left[|X(\omega)|^2\right] + \alpha E\left[|D(\omega)|^2\right]} \right)^{\beta}$$

- $E\left[|X(\omega)|^2\right]$ is unknown
- If $\alpha = 1$, $\beta = 1/2$, and $E\left[|X(\omega)|^2\right] = |\hat{X}(\omega)|^2$ then...

$$|\hat{X}(\omega)| = H(\omega)|Y(\omega)| = \sqrt{\frac{|\hat{X}(\omega)|^2}{|\hat{X}(\omega)|^2 + E\left[|D(\omega)|^2\right]}}|Y(\omega)|$$

$$|\hat{X}(\omega)|^2(|\hat{X}(\omega)|^2 + E\left[|D(\omega)|^2\right]) = |\hat{X}(\omega)|^2|Y(\omega)|^2$$

- gives two solutions $|\hat{X}(\omega)|^2 = |\hat{Y}(\omega)|^2 - E\left[|D(\omega)|^2\right]$ or $|\hat{X}(\omega)|^2 = 0$.
- which is essentially the power spectral subtraction algorithm

Wiener filter gain

- If we replace $E[|X(\omega)|^2] = |\hat{Y}(\omega)|^2 - E\left[|D(\omega)|^2\right]$
- Wiener filter

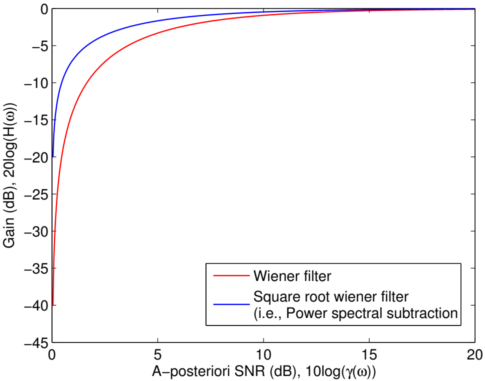
$$H(\omega) = \frac{E[|X(\omega)|^2]}{E[|X(\omega)|^2] + E\left[|D(\omega)|^2\right]} = \frac{\gamma(\omega) - 1}{\gamma(\omega)}$$

, where $\gamma(\omega) = \frac{|\hat{Y}(\omega)|^2}{E[|D(\omega)|^2]}$.

- The square root wiener filter = power spectral subtraction

$$H(\omega)_{\frac{1}{2}} = \sqrt{\frac{\gamma(\omega) - 1}{\gamma(\omega)}}$$

Wiener filter gain



Connection with spectral subtraction

- If $\alpha \neq 1$, using the same method

$$|\hat{X}(\omega)|^2 = |\hat{Y}(\omega)|^2 - \alpha E\left[|D(\omega)|^2\right]$$

  which is the spectral over subtraction method
- Note the Wiener filter is zero-phase, and thus $\angle\hat{X}(\omega) = \angle Y(\omega)$, just like the spectral subtraction method

MMSE-STSA Estimator

- Suggested by Eprahim and Malah [EM84]
- Estimator that minimizes the mean square error of the spectral magnitude
- Given $X(\omega_k) = X_k e^{j\angle X(\omega_k)}$,

$$\min E\left[(X_k - \hat{X}_k)^2\right]$$

Comparison of MMSE-STSA Estimator with Wiener filter

1. MMSE in the complex spectrum vs. magnitude spectrum
   - Wiener: $\min E\left[(X(\omega_k) - \hat{X}(\omega_k))^2)\right]$
   - MMSE-STSA: $\min E\left[(X_k - \hat{X}_k)^2\right]$
2. Linear assumption vs. assumption on distribution of X_k
   - Wiener: $\min E\left[(X(\omega_k) - H(\omega_k)Y(\omega_k))^2\right]$
   - MMSE-STSA: $\min E\left[(X_k - \hat{X}_k)^2\right]$, where expectation is taken over $p(Y(\omega_k), X_k)$

MMSE-STSA Estimator

$$\min E\left[(X_k - \hat{X}_k)^2\right]$$

- From Bayesian statistics the optimal MMSE estimator is,

$$\begin{aligned}
\hat{X}_k &= E\left[X_k | Y(\omega_k)\right] \\
&= \int_0^\infty x_k p(x_k | Y(\omega_k)) dx_k \\
&= \frac{\int_0^\infty x_k p(Y(\omega_k) | x_k) p(x_k) dx_k}{p(Y(\omega_k))}
\end{aligned}$$

- We need knowledge on the distribution of $X(\omega_k)$ and $Y(\omega_k)$

Distribution assumption for $X(w_k)$, $D(w_k)$, and $Y(w_k)$

- Fourier transform coefficients (of both speech and noise) are Gaussian distributed.
  - From central limit theorem: $Y(\omega_k) = \sum_{n=0}^{N-1} y(n)e^{-j\omega_k n}$
  - CLT holds for weakly dependent signals too
  - The variance of the distribution $E|Y(\omega_k)|^2$ is time varying
- $X(w_k) \sim \mathcal{N}(0, E\left[|X(w_k)|^2\right])$
- $D(w_k) \sim \mathcal{N}(0, E\left[|D(w_k)|^2\right])$
- $Y(w_k) \sim \mathcal{N}(0, E\left[|X(w_k)|^2\right] + E\left[|D(w_k)|^2\right])$
- $X_k \sim \text{Rayleigh}(\sigma)$, with $\sigma = \sqrt{E\left[|X(w_k)|^2\right]/2}$

Spectral gain of MMSE-STSA estimator

- The spectral gain can be represented with two variables.
  - The a-priori SNR: $\xi_k = \frac{E\left[|X(\omega_k)|^2\right]}{E[|D(\omega_k)|^2]}$
  - The a-posteriori SNR: $\gamma_k = \frac{|Y(\omega_k)|^2}{E[|D(\omega_k)|^2]}$
- Using a temporary variable, $\nu_k = \frac{\xi_k}{1+\xi_k}\gamma_k$,

$$\hat{X}_k = \frac{\sqrt{\pi}}{2}\frac{\sqrt{\nu_k}}{\gamma_k}\exp\left(-\frac{\nu_k}{2}\right)\left[(1+\nu_k)I_0\left(\frac{\nu_k}{2}\right)+\nu_k I_1\left(\frac{\nu_k}{2}\right)\right]Y_k$$
$$= G(\xi_k, \gamma_k)Y_k$$

Gain as a function of a-priori SNR

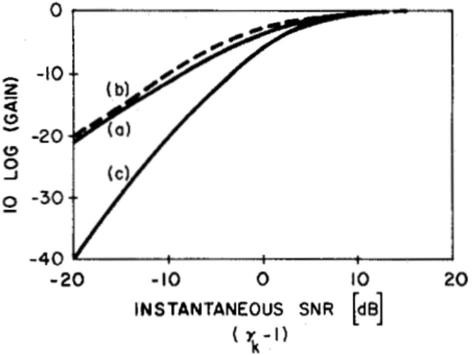


Fig. 6. Gain curves describing (a) MMSE gain function $G_{MMSE}(\xi_k, \gamma_k)$ defined by (7) and (14), with $\xi_k = \gamma_k - 1$, (b) "spectral subtraction" gain function (46) with $\beta = 1$, and (c) Wiener gain function $G_w(\xi_k, \gamma_k)$ (15) with $\xi_k = \gamma_k - 1$.

Estimating the a-priori SNR

$$\xi_k = \frac{E\left[|X(\omega_k)|^2\right]}{E\left[|D(\omega_k)|^2\right]}$$

- Instantaneous SNR: $\hat{\xi}_k = \frac{|Y(\omega_k)|^2 - E\left[|D(\omega_k)|^2\right]}{E\left[|D(\omega_k)|^2\right]} = \frac{|Y(\omega_k)|^2}{E\left[|D(\omega_k)|^2\right]} - 1$
- Decision directed approach

$$\hat{\xi}_k(m) = a\frac{\hat{X}_k^2(m-1)}{E\left[|D(\omega_k, m-1)|^2\right]} + (1-a)\left(\frac{|Y(\omega_k)|^2}{E\left[|D(\omega_k)|^2\right]} - 1\right)$$

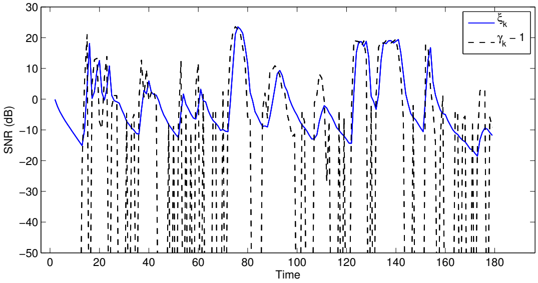
Effect of smoothed SNR

- Instantaneous SNR:

$$\hat{\xi}_k = \gamma(\omega_k) - 1 = \frac{|Y(\omega_k)|^2}{E\left[|D(\omega_k)|^2\right]} - 1$$

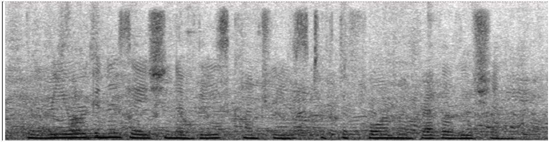
- Decision directed approach

$$\hat{\xi}_k(m) = a\frac{\hat{X}_k^2(m-1)}{E\left[|D(\omega_k, m-1)|^2\right]} + (1-a)\left(\frac{|Y(\omega_k)|^2}{E\left[|D(\omega_k)|^2\right]} - 1\right)$$

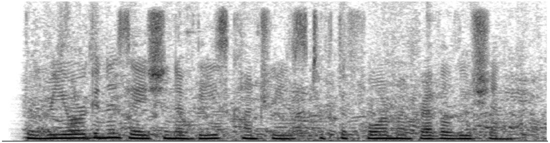


Examples with decision directed a-priori SNR estimation

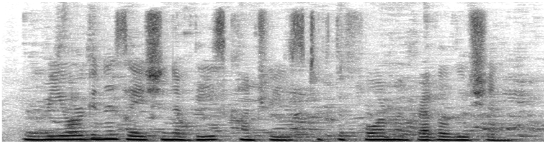
Noisy speech



MMSE-STSA (Instantaneous SNR)

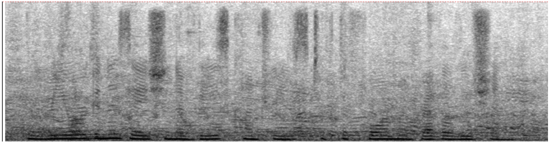


MMSE-STSA (Decision Directed SNR)

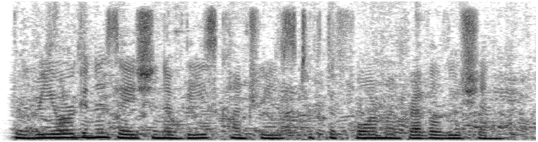


Examples with decision directed a-priori SNR estimation

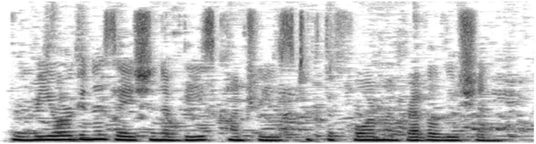
Noisy speech



Wiener Filter (Instantaneous SNR)



Wiener Filter (Decision Directed SNR)



Summary

- Model noise when speech is absent. $|\hat{D}(\omega)| = E\left[|D(\omega)|\right]$
- Separate speech by applying gain on the noisy spectrum.
  1. Spectral subtraction: $|\hat{X}(\omega)| = |Y(\omega)| - |\hat{D}(\omega)|$
  2. Wiener filter: $\hat{X}(\omega) = \frac{E\left[|X(\omega)|^2\right]}{E[|X(\omega)|^2] + E[|D(\omega)|^2]} Y(\omega)$
  3. STSA-MSME: $\hat{X}(\omega) = G(\xi(\omega), \gamma(\omega))Y(\omega)$

References I

- Steven Boll, *Suppression of acoustic noise in speech using spectral subtraction*, Acoustics, Speech and Signal Processing, IEEE Transactions on **27** (1979), no. 2, 113–120.
- M. Berouti, M. Schwartz, and J. Makhoul, *Enhancement of speech corrupted by acoustic noise*, IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP '79), 1979, pp. 208–211.
- Y. Ephraim and D. Malah, *Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator*, IEEE Transactions on Acoustics, Speech and Signal Processing (1984), 1109–1121.
- Jae S Lim and Alan V Oppenheim, *Enhancement and bandwidth compression of noisy speech*, Proceedings of the IEEE **67** (1979), no. 12, 1586–1604.