Source Separation Tutorial Mini-Series I
Speech enhancement algorithms

Eunjoon Cho
Stanford University, EE
April 2nd, 2013
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.

Eunjoon Cho (Stanford University, EE)
Source Separation Tutorial Mini-Series I
April 2nd, 2013
Speech enhancement: The application perspective

- Goal: Increase the intelligibility/quality of noisy speech.
- Practical constraints.
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.
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]
Discuss approaches in the frequency domain: \( Y(\omega) = X(\omega) + D(\omega) \)

Overall flow of STFT processing

\[
\begin{align*}
y(n) &\quad \rightarrow \quad w(n - mR) \quad \rightarrow \quad y_m(n) \quad \rightarrow \quad \text{FT} \quad \rightarrow \quad Y_m(\omega) \\
\hat{x}(n) &\quad \leftarrow \quad \text{Overlap add} \quad \leftarrow \quad \hat{x}_m(n) \quad \leftarrow \quad \text{IFT} \quad \leftarrow \quad \hat{X}_m(\omega)
\end{align*}
\]
Spectral subtraction

\[ |Y_m(\omega)| \]
Spectral subtraction

\[ |Y_m(\omega)|, \; |\hat{D}_m(\omega)| \]
Spectral subtraction

\[ |Y_m(\omega)|, E[|D_m(\omega)|] \]
Spectral subtraction

\[ |\hat{X}_m(\omega)| = |Y_m(\omega)| - E[|D_m(\omega)|] \]
Spectral subtraction

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

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

\[
\angle \hat{X}_m(\omega) = \angle Y_m(\omega)
\]
Power spectral subtraction

\[ |\hat{X}_{m}(\omega)|^{\alpha} = \max\{|Y_{m}(\omega)|^{\alpha} - E[|D_{m}(\omega)|^{\alpha}], 0\} \]
Power spectral subtraction

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

- The power spectral subtraction: \( \alpha = 2 \)

\[ |\hat{X}_m(\omega)|^2 = \max\{|Y_m(\omega)|^2 - E[|D_m(\omega)|^2], 0\} \]
Gain for power spectral subtraction

From the power spectral subtraction,

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

the gain can be expressed as follows

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

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

- Different gain functions for various \( \alpha \)

\[ H_m(\omega) = \frac{|\hat{X}_m(\omega)|}{|Y_m(\omega)|} = \left( \frac{\gamma(\omega)^{\alpha/2} - 1}{\gamma(\omega)^{\alpha/2}} \right)^{1/\alpha} \]
Example of spectral subtraction

Noisy speech

Power spectral subtraction

Magnitude spectral subtraction
Q. Why do we get musical noise?
Q. Why do we get musical noise?

A1. Inaccurate estimate of unknown variables.

From $Y_m(\omega) = X_m(\omega) + D_m(\omega)$,

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

$$- (X_m(\omega)D_m^*(\omega) + X_m^*(\omega)D_m(\omega))$$

$$|D_m(\omega)|^2 \approx E[|D_m(\omega)|^2]$$

$$2Re\{X_m(\omega)D_m^*(\omega)\} \approx E[2Re\{X_m(\omega)D_m^*(\omega)\}] = 0$$
Noise estimation

- Simple method is to average power spectra when there is no speech activity $|\hat{D}_m(\omega)|^2 = E[|Y_m(\omega)|^2] = E[|D_m(\omega)|^2]$
  
  - Need an accurate voice activity detector (VAD)
  - Issues with non-stationary noise (babble noise) conditions

- Issues with $|\hat{D}_m(\omega)|^2 = E[|D_m(\omega)|^2]$
  
  - The noise power spectrum (periodogram) has high variance with respect to the underlying power spectral density
Cross term spectrum

- $2\text{Re}\{X_m(\omega)D_m^*(\omega)\}$ requires the phase information which is difficult to estimate
Q. Why do we get musical noise?
A1. Inaccurate estimate of unknown variables.
A2. How we engineer situations when we have a bad estimate.
   - Half rectify negative values.

\[ |\hat{X}_m(\omega)|^2 = \max\{|Y_m(\omega)|^2 - \mathbb{E}[|D_m(\omega)|^2], 0\} \]
Artifacts from half-wave rectifying

\[1\] 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

\[ \alpha = 1, \text{Frame 25} \]

\[ \alpha = 2, \text{Frame 25} \]

\[ \alpha = 3, \text{Frame 25} \]
Oversubtraction [BSM79]

- Fill in valleys at frequencies to mask residue noise

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

\[\alpha = 2, \beta = 0.001, \text{Frame 25}\]

\[\alpha = 2, \beta = 0.5, \text{Frame 25}\]
Oversubtraction [BSM79]

- $\alpha$ should be dependent on the frame segmental SNR ($\gamma$)
- Less attenuation (small $\alpha$) for high SNR, and more attenuation (large $\alpha$) for low SNR

![Graph showing the relationship between SNR (dB) and $\alpha$]
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 \( h^\star \) that minimizes \( E[e^2(n)] \) by solving \( \frac{\partial E[e^2(n)]}{\partial h} = 0 \)

\[ h^\star = R_{yy}^{-1} r_{yx} = (R_{xx} + R_{dd})^{-1} r_{xx} \]
Wiener filter in frequency domain

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

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

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

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

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

More generally,

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

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

- \( E[|X(\omega)|^2] \) is unknown
Connection with spectral subtraction

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

- \( E[|X(\omega)|^2] \) is unknown
- If \( \alpha = 1, \beta = 1/2, \) and \( E[|X(\omega)|^2] = |\hat{X}(\omega)|^2 \) then...
Connection with spectral subtraction

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

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

\[ |\hat{X}(\omega)| = H(\omega)|Y(\omega)| = \sqrt{\frac{|\hat{X}(\omega)|^2}{|\hat{X}(\omega)|^2 + E[|D(\omega)|^2]}|Y(\omega)|} \]
Connection with spectral subtraction

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

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

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

\[ |\hat{X}(\omega)|^2(|\hat{X}(\omega)|^2 + E[|D(\omega)|^2]) = |\hat{X}(\omega)|^2|Y(\omega)|^2 \]
Connection with spectral subtraction

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

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

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

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

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

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

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

, 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

Gain (dB), $20\log(H(\omega))$

A-posteriori SNR (dB), $10\log(\gamma(\omega))$

Wiener filter
Square root wiener filter
(i.e., Power spectral subtraction)
Connection with spectral subtraction

- If $\alpha \neq 1$, using the same method
  \[
  |\hat{X}(\omega)|^2 = |\hat{Y}(\omega)|^2 - \alpha E[|D(\omega)|^2]
  \]
  which is the spectral over subtraction method.

- Note the Wiener filter is zero-phase, and thus $\angle \hat{X}(\omega) = \angle \hat{Y}(\omega)$, 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 \( p(Y(\omega_k), X_k) \)
\[ \min E \left[ (X_k - \hat{X}_k)^2 \right] \]

- From Bayesian statistics the optimal MMSE estimator is,

\[
\hat{X}_k = E [X_k | Y(\omega_k)] \\
= \int_0^{\infty} x_k p(x_k | Y(\omega_k)) \, dx_k \\
= \int_0^{\infty} x_k p(Y(\omega_k) | x_k) p(x_k) \, dx_k \\
= \frac{\int_0^{\infty} x_k p(Y(\omega_k) | x_k) p(x_k) \, dx_k}{p(Y(\omega_k))}
\]

- 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

- $X(w_k) \sim \mathcal{N}(0, E[|X(w_k)|^2])$
- $D(w_k) \sim \mathcal{N}(0, E[|D(w_k)|^2])$
- $Y(w_k) \sim \mathcal{N}(0, E[|X(w_k)|^2] + E[|D(w_k)|^2])$
- $X_k \sim \text{Rayleigh}(\sigma)$, with $\sigma = \sqrt{E[|X(w_k)|^2]/2}$
Spectral gain of MMSE-STSA estimator

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

- Using a temporary variable, $\nu_k = \frac{\xi_k}{1+\xi_k} \gamma_k$,

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

\[
= G(\xi_k, \gamma_k) Y_k
\]
Fig. 6. Gain curves describing (a) MMSE gain function $G_{\text{MMSE}}(\xi_k, \gamma_k)$ 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 $\xi_k = \gamma_k - 1$. 

Gain as a function of a-priori SNR
Estimating the a-priori SNR

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

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

- Decision directed approach

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

- **Instantaneous SNR:**
  \[ \hat{\xi}_k = \gamma(\omega_k) - 1 = \frac{|Y(\omega_k)|^2}{E[|D(\omega_k)|^2]} - 1 \]

- **Decision directed approach**
  \[ \hat{\xi}_k(m) = a \frac{\hat{X}_k^2(m-1)}{E[|D(\omega_k, m-1)|^2]} + (1-a) \left( \frac{|Y(\omega_k)|^2}{E[|D(\omega_k)|^2]} - 1 \right) \]
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)
Model noise when speech is absent.  \[ |\hat{D}(\omega)| = E[|D(\omega)|] \]
Separate speech by applying gain on the noisy spectrum.

1. Spectral subtraction:  \[ |\hat{X}(\omega)| = |Y(\omega)| - |\hat{D}(\omega)| \]
2. Wiener filter:  \[ \hat{X}(\omega) = \frac{E[|X(\omega)|^2]}{E[|X(\omega)|^2] + E[|D(\omega)|^2]} Y(\omega) \]
3. STSA-MSME:  \[ \hat{X}(\omega) = G(\xi(\omega), \gamma(\omega))Y(\omega) \]
References I


