Dealing with Acoustic Noise

Part 2: Beamforming

Mark Hasegawa-Johnson
University of Illinois
Lectures at CLSP WS06
July 25, 2006
AVICAR Recording Hardware

4 Cameras, Glare Shields, Adjustable Mounting

Best Place= Dashboard

8 Mics, Pre-amps, Wooden Baffle.

Best Place= Sunvisor.

System is not permanently installed; mounting requires 10 minutes.
AVICAR Database

- 100 Talkers
- 4 Cameras, 7 Microphones
- 5 noise conditions: Engine idling, 35mph, 35mph with windows open, 55mph, 55mph with windows open
- Three types of utterances:
  - Digits & Phone numbers, for training and testing phone-number recognizers
  - Phonetically balanced sentences, for training and testing large vocabulary speech recognition
  - Isolated letters, to see how video can help with an acoustically hard problem
- Open-IP public release to 15 institutions, 5 countries
Noise and Lombard Effect
Beamforming Configuration

- Microphone Array
- Closed Window (Acoustic Reflector)
- Open Window (Noise Source)
- Talker (Automobile Passenger)
**Frequency-Domain Expression**

\[ \hat{S} = W^T X, \quad X = HS + V \]

- \( X_{mk} \) is the measured signal at microphone \( m \) in frequency band \( k \).
- Assume that \( X_{mk} \) was created by filtering the speech signal \( S_k \) through the room response filter \( H_{mk} \) and adding noise \( V_{mk} \).
- Beamforming estimates an “inverse filter” \( W_{mk} \).

\[
W = \begin{bmatrix}
W_{1k} \\
W_{2k} \\
\vdots \\
W_{Mk}
\end{bmatrix}, \quad X = \begin{bmatrix}
X_{1k} \\
X_{2k} \\
\vdots \\
X_{Mk}
\end{bmatrix}
\]

\[
H = \begin{bmatrix}
H_{1k} \\
H_{2k} \\
\vdots \\
H_{Mk}
\end{bmatrix}, \quad V = \begin{bmatrix}
V_{1k} \\
V_{2k} \\
\vdots \\
V_{Mk}
\end{bmatrix}
\]
Time-Domain Approximation

\[ \hat{S} = W^T X, \quad X = HS + V \]

\[ \hat{S} = \hat{s}[n], \quad W = \begin{bmatrix} w_1[0] \\
                          w_1[1] \\
                          \vdots \\
                          w_M[L] \end{bmatrix}, \quad X = \begin{bmatrix} x_1[n] \\
                                              x_1[n-1] \\
                                              \vdots \\
                                              x_M[n-L] \end{bmatrix}, \quad V = \begin{bmatrix} v_1[n] \\
                                                v_1[n-1] \\
                                                \vdots \\
                                                v_M[n-L] \end{bmatrix} \]

\[ H = \begin{bmatrix} h_1[0] & h_1[1] & \cdots & h_1[L] & 0 & \cdots & 0 \\
                        0 & h_1[0] & \cdots & h_1[L-1] & h_1[L] & \cdots & 0 \\
                        \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
                        0 & 0 & \cdots & h_M[0] & h_M[1] & \cdots & h_M[L] \end{bmatrix}, \quad S = \begin{bmatrix} s[n] \\
                                             s[n-1] \\
                                             \vdots \\
                                             s[n-2L] \end{bmatrix} \]
Optimality Criteria for Beamforming

• $\hat{s}[n] = W^T X$ is the estimated speech signal

• $W$ is chosen for
  – Distortionless response:
    
    $W^T H S = S$
    
    Time Domain: $W^T H = [1,0,\ldots,0]$
    
    Freq Domain: $W^H H = 1$
  
  – Minimum variance:
    
    $W = \text{argmin}(W^T R W), \ R = E[V V^T]$
  
  – Multichannel spectral estimation:
    
    $E[f(S)|X] = E[f(S)|\hat{S}]$
Delay-and-Sum Beamformer

\[ \hat{s}[n] = \frac{1}{M} \sum_{m} x[n - n_m] \]

\[ \hat{S} = W^T X = \frac{1}{M} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}^T \begin{bmatrix} e^{-j2\pi kn_1/N} X_{1k} \\ \vdots \\ e^{-j2\pi kn_M/N} X_{Mk} \end{bmatrix} \]

...or...

\[ \hat{S} = W^T X = \frac{1}{M} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 1 \end{bmatrix}^T \begin{bmatrix} x_1[n - n_1] \\ x_1[n - n_1 - 1] \\ \vdots \\ x_M[n - n_M] \\ x_M[n - n_M - L] \end{bmatrix} \]

Suppose that the speech signal at microphone \( m \) arrives earlier than at some reference microphone, by \( n_m \) samples. Then we can estimate \( S \) by delaying each microphone \( n_m \) samples, and adding:
What are the Delays?

Far-field approximation: inter-microphone delay \( \tau = \frac{d \sin \theta}{c} \)

Near-field (talker closer than \(~10d\)): formulas exist
Delay-and-Sum Beamformer has Distortionless Response

\[ \hat{S} = W^T X = W^T HS + W^T V \]

In the noise-free, echo-free case, we get...

\[
\hat{s}[n] = W^T HS = \frac{1}{M} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix}^T \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \end{bmatrix} \begin{bmatrix} s[n] \\ \vdots \\ s[n-L] \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}^T \begin{bmatrix} s[n] \\ \vdots \\ s[n-L] \end{bmatrix} = s[n]
\]
Delay-and-Sum Beamformer with Non-Integer Delays

\[ x_m(t) = s(t + \tau_m) + v_m(t) \]

\[ x_m[n] = h_m[n] * s[n] + v_m[n] \]

\[ h_m[n] = \text{sinc}(\pi(n + F_s \tau_m)) \]

\[
\begin{bmatrix}
  h_1[0] & h_1[1] & \cdots & h_1[L] & 0 & \cdots & 0 \\
  0 & h_1[0] & \cdots & h_1[L-1] & h_1[L] & \cdots & 0 \\
  \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
  0 & 0 & \cdots & h_1[0] & h_1[1] & \cdots & h_1[L] \\
  h_2[0] & h_2[1] & \cdots & h_2[L] & 0 & \cdots & 0 \\
  \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
  0 & 0 & \cdots & h_M[0] & h_M[1] & \cdots & h_M[L]
\end{bmatrix}
\approx
\begin{bmatrix}
  s[n] \\
  s[n-1] \\
  \vdots \\
  s[n-2L]
\end{bmatrix}
Distortionless Response for Channels with Non-Integer Delays or Echoes

\[ \hat{S} = W^T H S = [1, 0, \cdots, 0] \begin{bmatrix} \vdots \\ s[n] \\ s[n-L] \end{bmatrix} = s[n] \]

So, we need to find any \( W \) such that...

\[ W^T H = [1, 0, \cdots, 0] \equiv \delta^T \]

Here is one way to write the solution:

\[ W^T = \delta^T (H^T A H)^{-1} H^T A \text{ for any square } A \]

Here is another way to write it:

\[ W^T = \delta^T (H^T H)^{-1} H^T + W_B^T B^T \text{ for any vector } W_B \]

... where \( B \) is the “null space” of matrix \( H \), i.e., ...

\[ B^T H = 0 \]
Beam Patterns

- Seven microphones, spaced 4cm apart
- Delay-and-Sum Beamformer, Steered to $\theta=0$
Minimum Variance Distortionless Response

(Frost, 1972)

\[ \hat{s}[n] = W^T(HS + V) \]

• Define an error signal, e[n]:

\[ e[n] = \hat{s}[n] - W^T HS = W^TV \]

\[ E[e^2[n]] = W^T E[VV^T] W = W^T R_V W \]

• Goal: minimize the error power, subject to the constraint of distortionless response:

\[ W^* = \arg \min_W W^T R_V W \text{ subject to } W^T H = \delta^T \]
Minimum Variance Distortionless Response: The Solution

\[ W^* = \arg\min_W W^T R_V W \text{ subject to } W^T H = \delta^T \]

- The closed-form solution (for stationary noise):
  \[ W^{*T} = \delta^T \left( H^T R_V^{-1} H \right)^{-1} H^T R_V^{-1} \]
  \[ \hat{s}[n] = \delta^T \left( H^T R_V^{-1} H \right)^{-1} H^T R_V^{-1} X \]

- The adaptive solution: adapt \( W_B \) so that…
  \[ W^{*T} = \delta^T \left( H^T H \right)^{-1} H^T + W_B^T B^T \]
  \[ W_B = \arg\min E \left[ |S - W^T X|^2 \right] \]
Beam Patterns, MVDR

- MVDR beamformer, tuned to cancel a noise source distributed over $60 < \theta < 90$ deg.
Multi-Channel Spectral Estimation

• We want to estimate some function $f(S)$, given a multi-channel, noisy, reverberant measurement $X$:

\[
E[f(S)|X] = \frac{\int f(S)p(S,X)dS}{\int p(S,X)dS}
\]

• Assume that $S$ and $X$ are jointly Gaussian, thus:

\[
p(X,S) = p(X|S)p(S)
\]

\[
p(X|S) = |2\pi R_V|^{-1/2} e^{-\frac{1}{2} (X-HS)^T R_V^{-1} (X-HS)}
\]

\[
p(S) = |2\pi R_S|^{-1/2} e^{-\frac{1}{2} S^T R_S^{-1} S}
\]
p(X|S) Has Two Factors

\[ p(X|S) = p(X_{\perp \hat{S}})p(\hat{S}|S) \]

\[ p(X_{\perp \hat{S}}) = |2\pi R_{\perp \hat{S}}|^{-1/2}e^{-\frac{1}{2}X^TR_{\perp \hat{S}}X} \]

\[ p(\hat{S}|S) = |2\pi R_{\hat{S}}|^{-1/2}e^{-\frac{1}{2}(\hat{S}-S)^TR_{\hat{S}}^{-1}(\hat{S}-S)} \]

- Where \( \hat{S} \) is (somehow, by magic) the MVDR beamformed signal:

\[ \hat{S} = (H^TR^{-1}H)^{-1}H^TR_{V}^{-1}X \]

... and the covariance matrices of \( \hat{S} \) and its orthogonal complement are:

\[ R_{\hat{S}}^{-1} = H^TR_{V}^{-1}H, \quad R_{\perp \hat{S}} = R_{V}(R_{V} - HR_{\hat{S}}H^T)^{-1}R_{V} \]
Sufficient Statistics for Multichannel Estimation

(Balan and Rosca, SAMSP 2002)

\[
E [f(S)|X] = \frac{\int f(S)p(X|S)p(S)\,dS}{\int p(X|S)p(S)\,dS}
\]

\[
E [f(S)|X] = \frac{\int f(S)p(\hat{S}|S)p(S)\,dS}{\int p(\hat{S}|S)p(S)\,dS}
\]

\[
E [f(S)|X] = E \left[f(S)|\hat{S}\right]
\]
Multi-Channel Estimation: Estimating the Noise

- Independent Noise Assumption
- Measured Noise Covariance
  - Problem: $R_v$ may be singular, especially if estimated from a small number of frames
- Isotropic Noise (Kim and Hasegawa-Johnson, AES 2005)
  - Isotropic = Coming from every direction with equal power
  - The form of $R_v$ has been solved analytically, and is guaranteed to never be exactly singular
  - Problem: noise is not really isotropic, e.g., it comes primarily from the window and the radio
Isotropic Noise: Not Independent

Coherence of Isotropic Noise at 4cm

Coherence, $R_{y}(1,2)/R_{y}(1,1)$ vs Frequency (Hz)
Adaptive Filtering for Multi-channel Estimation

(Kim, Hasegawa-Johnson, and Sung, ICASSP 2006)
MVDR with Correct Assumed Channel

MVDR eliminates high-frequency noise,
MMSE-logSA eliminates low-frequency noise
MMSE-logSA adds reverberation at low frequencies;
reverberation seems to not effect speech recognition accuracy
MVDR with Incorrect Assumed Channel

- One Mic, No Processing
- One Mic, MMSE-LSA
- Delay-Sum, No Postfilter
- Delay-Sum, MMSE-LSA
- MVDR, No Postfilter
- MVDR, MMSE-LSA

Time (sec)
Channel Estimation

- Measured signal $x[n]$ is the sum of noise ($v[n]$), a “direct sound” ($a_0s[n-n_0]$), and an infinite number of echoes:

$$x[n] = v[n] + \sum_{i=0}^{\infty} a_i s[n - n_i]$$

- The process of adding echoes can be written as convolution:

$$x[n] = v[n] + h[n] * s[n]$$

$$h[n] = \sum_{i=0}^{\infty} a_i \delta[n - n_i]$$
Channel Estimation

- If you know enough about $R_s[m]$, then the echo times can be estimated from $R_x[m]$:

\[
R_x[m] = \mathbb{E}[x[n]x[n - m]]
\]

\[
R_v[m] = \mathbb{E}[v[n]v[n - m]] \approx \delta[m]
\]

\[
R_s[m] = \mathbb{E}[x[n]x[n - m]]
\]

\[
R_x[m] = R_v[m] + \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} a_i a_j R_s[m - (n_i - n_j)]
\]
Channel Estimation

• For example, if the “speech signal” is actually white noise, then $R_x[m]$ has peaks at every inter-echo-arrival time:

$$
R_x[m] = \delta[m] + \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} a_i a_j \delta[n_i - n_j]
$$
Channel Estimation

- Unfortunately, $R_s[m]$ is usually not sufficiently well known to allow us to infer $n_i$ and $n_j$.
Channel Estimation

• Seeded methods, e.g., maximum-length sequence pseudo-noise signals, or chirp signals
  – Problem: channel response from loudspeaker to microphone not same as from lips to microphone

• Independent components analysis
  – Problem: doesn’t work if speech and noise are both nearly Gaussian
Channel Response as a Random Variable: EM Beamforming

(Kim, Hasegawa-Johnson, and Sung, in preparation)

\[
\hat{S}_{ML} = \arg \max \log p(X|S)
\]

\[
\log p(X|S) = \log \int p(X|S, H)p(H)dH
\]

\[
\geq E_H [\log p(X|S, H)]
\]
WER Results, AVICAR

Ten-digit phone numbers; trained and tested with 50/50 mix of quiet (engine idling) and very noisy (55mph, windows open)