Linear and Parametric Microphone Array Processing

Part II: Linear Spatial Processing

Emanuël A.P. Habets¹ and Sharon Gannot²

¹International Audio Laboratories Erlangen, Germany
²Faculty of Engineering, Bar-Ilan University, Israel

ICASSP 2013, Vancouver, Canada
Linear Spatial Noise Reduction Techniques I

Families of Methods

1. **Fixed beamforming** Combine the microphone signals using a time-invariant filter-and-sum operation (data-independent)
   
   [Jan and Flanagan, 1996]; [Doclo and Moonen, 2003].

2. **Blind Source Separation (BSS)** Considers the received signals at the microphones as a mixture of all sound sources filtered by the RIRs. Utilizes Independent Component Analysis (ICA) techniques
   
   [Makino et al., 2007]; TRINICON, [Buchner et al., 2004].

3. **Adaptive Beamforming** Combine the spatial focusing of fixed beamformers with adaptive suppression of (spectrally and spatially time-varying) background noise

   **General reading:** [Cox et al., 1987]; [Van Veen and Buckley, 1988]; [Van Trees, 2002].
Some Criteria


3. **Minimum mean square error (MMSE) - GSVD based spatial Wiener filter** [Doclo and Moonen, 2002a].

4. **Speech distortion weighted multichannel Wiener filter (SDW-MWF)** [Doclo and Moonen, 2002b]; [Spriet et al., 2004]; [Doclo et al., 2005].

5. **Maximum signal to noise ratio (SNR)** [Warsitz and Haeb-Umbach, 2007].

6. **Linearly constrained minimum variance (LCMV)** [Markovich et al., 2009].
Linear Spatial Noise Reduction Techniques III

Some Books

1. Acoustic signal processing for telecommunication [Gay and Benesty, 2000].
3. Speech Enhancement [Benesty et al., 2005].
4. Blind speech separation [Makino et al., 2007].
5. Microphone Array Signal Processing [Benesty et al., 2008a].
6. Springer handbook of speech processing [Benesty et al., 2008b].
8. Speech processing in modern communication: Challenges and perspectives [Cohen et al., 2010].
Spatial Filters

Beamforming: Filter and Sum

\[ y(t) = w^H(t)z(t). \]

**w**: \( M \times 1 \) beamforming vector of filters (or just gains).
Far-Field Wave $s(t) = e^{j\omega t}$

Narrow-band Signal

\[
y(t) = \sum_{m=0}^{M-1} w_m^* e^{j\omega_0 (t - \tau_m)} \\
= e^{j\omega_0 t} \sum_{m=0}^{M-1} w_m^* e^{-j\omega_0 \left(\frac{d \cos(\theta)}{c}\right)m} \\
= e^{j\omega_0 t} \sum_{m=0}^{M-1} w_m^* e^{-j2\pi \frac{d}{\lambda_0} \cos(\theta)m}
\]

Beampattern is the DTFT of the weights

\[
y(t) = e^{j\omega_0 t} \mathcal{W} \left( \frac{d}{\lambda_0}; \cos(\theta) \right)
\]
The Delay & Sum Beamformer

Uniform Linear Array (ULA)

- \( w_m = \frac{1}{M}; \ m = 0, \ldots, M - 1. \)
- For simplicity, assume symmetric array.
- Steered to \( \cos(\theta_0). \)
- Beampattern:

\[
B(\theta) = \frac{1}{M} \cdot \frac{\sin \left( \frac{M}{2} \frac{2\pi}{\lambda_0} \left( \cos(\theta) - \cos(\theta_0) \right) \right)}{\sin \left( \frac{1}{2} \frac{2\pi}{\lambda_0} \left( \cos(\theta) - \cos(\theta_0) \right) \right)}
\]

Beamformers

- Discriminate between angles.
- Can be steered by setting \( \mathbf{w}. \)
- Depends on the ratio \( \frac{d}{\lambda_0}. \)
Beampattern

(a) $\theta_0 = 90^\circ; \frac{d}{\lambda_0} = \frac{1}{2}$
(b) $\theta_0 = 0^\circ; \frac{d}{\lambda_0} = \frac{1}{2}$
(c) $\theta_0 = 40^\circ; \frac{d}{\lambda_0} = \frac{1}{2}$

(d) $\theta_0 = 90^\circ; \frac{d}{\lambda_0} = \frac{1}{32}$
(e) $\theta_0 = 90^\circ; \frac{d}{\lambda_0} = \frac{4}{1}$
10 microphone uniform linear array.

2 Desired sources in green and 2 interfering sources in red.

Can be obtained by applying the LCMV criterion.
Directivity and White Noise Gain (WNG) [Van Trees, 2002]

Definitions

- Propagation vector: \( \mathbf{u} = [\sin(\theta) \cos(\phi) \sin(\theta) \sin(\phi) \cos(\theta)]^T \).
- Beampattern: \( B(\phi, \theta) \).
- Beampower: \( P(\phi, \theta) = |B(\phi, \theta)|^2 \).

Directivity

- Assume that desired response is normalized: \( P(\phi_0, \theta_0) = 1 \).
- \( D = \left( \frac{1}{4\pi} \int_0^\pi \int_0^{2\pi} \sin(\theta) P(\phi, \theta) d\phi d\theta \right)^{-1} \).
- Directivity Index: \( DL = 10 \log_{10}(D) [\text{dB}] \).
- Maximum Directivity for ULA with \( d = \frac{\lambda}{2} \) is \( M \). It is achieved by the delay & sum beamformer.
Directivity and White Noise Gain (WNG) [Van Trees, 2002]

### White Noise Gain
- SNR improvement for spatially white input: \( A_w = \frac{\text{SNR}_{\text{out}}}{\text{SNR}_{\text{in}}} = \| w \|^{-2} \).
- Sensitivity to array weight imperfections and sensor misalignment is \( T_{se} = \frac{1}{A_w} = \| w \|^2 \) (hence, large WNG is better).

### Maximum Directivity [Parsons, 1987]
- MVDR criterion for diffuse noise field: super-directive beamformer.
- Obtained for linear endfire array with vanishingly small inter-sensor distance \( (d \to 0)! \)
- Maximum achievable directivity is \( M^2 \).
- In that case \( T_{se} \to \infty \) [Gilbert and Morgan, 1955] (see extension AASP-L4, Levin, Gannot and Habets).
- Robust design limiting the sensitivity exists [Cox et al., 1986].
- Forms the basis of differential microphone arrays [Elko, 1996].
Array Design for Speech Propagating in Acoustic Environments

- **Beampatterns:** Array response as a function of the angle of arrival (AoA).
- In reverberant environments (especially for low DRR), sound propagation is more involved than merely the AoA.
- The steering vector (comprised of the AoA) generalizes to **acoustic transfer function (ATF)**.
- The ATF summarizes all arrivals of the speech signals.
- The vector of received signals is treated as a vector in an **abstract linear space**.
- **Linear Algebra** methods are utilized to construct beamformers.
- AoA becomes less prominent.
A Noisy Example
Problem Formulation

Multiple Wideband Signals (e.g. Speech)

Short-Time Fourier Transform (STFT) -
Multiplicative Transfer Function (MTF) Approximation

\[ t \xrightarrow{\text{STFT}} \{\ell, k\} ; \text{Convolution} \xrightarrow{\text{STFT}} \text{Multiplication (for long enough frames)}. \]

Microphone Signals \((m = 0, \ldots, M - 1)\):

\[
z_m(\ell, k) = \sum_{j=1}^{P_d} s_d^j h_{jm} + \sum_{j=1}^{P_i} s_i^j h_{jm} + \sum_{j=1}^{P_n} s_n^j h_{jm} + n_m
\]

Vector Formulation

\[
z(\ell, k) = H^d s^d + H^i s^i + H^n s^n + n \triangleq Hs + n.
\]

\[ P = P_d + P_i + P_n \leq M \]

Beamforming in the STFT Domain

Apply filter & sum beamforming independently for each frequency bin.
Problem Formulation

Power Spectral Density (PSD)

Microphone Signals

\[ z(\ell, k) = H^d s^d + H^i s^i + H^n s^n + n \triangleq Hs + n \]

The PSD of the Various Components:

- **Stationary Sources:** \( \Phi_{zz}^{stat} = H^n \Phi_{s^n s^n} (H^n)^H + \Phi_{nn} \).
- **Constraints Sources:**
  \[
  H \Phi_{ss} H^H \triangleq H^d \Phi_{s^d s^d} (H^d)^H + H^i \Phi_{s^i s^i} (H^i)^H + H^n \Phi_{s^n s^n} (H^n)^H .
  \]
- **Microphone Signals:** \( \Phi_{zz}(\ell, k) = H \Phi_{ss} H^H + \Phi_{nn} \).
- **Noise+Interference Sources:**
  \[
  \Phi_{vv}(\ell, k) \triangleq H^i \Phi_{s^i s^i} (H^i)^H + H^n \Phi_{s^n s^n} (H^n)^H + \Phi_{nn}
  \]
Linearly Constrained Minimum Variance Beamformer

[Er and Cantoni, 1983]; [Van Veen and Buckley, 1988]

**LCMV Criterion**

- \( y(\ell, k) = w^H(\ell, k)z(\ell, k) \).
- Let \( \Phi_{nn} = E\{nn^H\} \) be the \( M \times M \) correlation matrix of the unconstraint sources.
- **Minimize** noise power \( w^H \Phi_{nn} w \)
  
  Such that a **linear** constraint set is satisfied: \( C^H w = g \).
- \( C : M \times P \) constraints matrix.
- \( g : P \times 1 \) response vector.

**Closed-form Solution**

\[
w(\ell, k) = \Phi_{nn}^{-1} C \left( C^H \Phi_{nn}^{-1} C \right)^{-1} g
\]
Linearly Constrained Minimum Power (LCMP) Beamformer

[Van Trees, 2002]

**LCMV vs. LCMP**

- Assume $\mathbf{C} = \mathbf{H}$ (all directional signals constrained).

\[
\mathbf{w}_{\text{LCMP}} = \underset{\mathbf{w}}{\text{argmin}} \left\{ \mathbf{w}^H \Phi_{zz} \mathbf{w} \right\} \text{ s.t. } \mathbf{H}^H \mathbf{w} = \mathbf{g} \\
= \underset{\mathbf{w}}{\text{argmin}} \left\{ \mathbf{w}^H (\mathbf{H} \Phi_{ss} \mathbf{H}^H + \Phi_{nn}) \mathbf{w} \right\} \text{ s.t. } \mathbf{H}^H \mathbf{w} = \mathbf{g} \\
= \underset{\mathbf{w}}{\text{argmin}} \left\{ \mathbf{g}^H \Phi_{ss} \mathbf{g} + \mathbf{w}^H \Phi_{nn} \mathbf{w} \right\} \text{ s.t. } \mathbf{H}^H \mathbf{w} = \mathbf{g} \\
= \underset{\mathbf{w}}{\text{argmin}} \left\{ \mathbf{w}^H \Phi_{nn} \mathbf{w} \right\} \text{ s.t. } \mathbf{H}^H \mathbf{w} = \mathbf{g} = \mathbf{w}_{\text{LCMV}}
\]

- If $\mathbf{H}$ is not accurately estimated, the LCMP beamformer exhibits self-cancellation and hence severe speech distortion.

- It is quite common in the literature to use only the term LCMV for both beamformers.
LCMV Minimization

Graphical Interpretation [Frost III, 1972]
The Minimum Variance Distortionless Beamformer

[Affes and Grenier, 1997]; [Hoshuyama et al., 1999]; [Gannot et al., 2001]

Beamformer Design:
- One desired signal $\Rightarrow$ Single constraint $(P = 1)$.
- “Steer a beam” to desired source and minimize other directions.
- $C = h^d; \quad g = 1.$

Closed-form Solution (MPDR eq. MVDR):

$$w(\ell, k) = \frac{\Phi_{zz}^{-1}h_d}{(h_d)^H \Phi_{zz}^{-1}h_d} = \frac{\Phi_{nn}^{-1}h_d}{(h_d)^H \Phi_{nn}^{-1}h_d}$$

Output signal:

$$y = s^d + \text{residual noise and interference signals}$$
## Multiple Speech Distortion Weighted Multichannel Wiener Filter (MSDW-MWF)

[Markovich-Golan et al., 2012b]

### Notation (Reminder)
- Received signals: $\mathbf{z}(\ell, k) = \mathbf{H}\mathbf{s} + \mathbf{n}$.
- $P < M$ constrained sources: $\mathbf{s}(\ell, k) \triangleq [s_1 \cdots s_P]^T$ and respective ATFs: $\mathbf{H}(\ell, k) \triangleq [h_1 \cdots h_P]$.
- Sources covariance matrix: $\Phi_{ss} = \text{diag}\{\phi_{s_1s_1}, \ldots, \phi_{s_Ps_P}\}$.
- Microphones covariance matrix: $\Phi_{zz} \triangleq \mathbf{H}\Phi_{ss}\mathbf{H}^\dagger + \Phi_{nn}$.

### MSDW-MWF
- Control the distortion of each individual source.
- Minimize the weighted mean square error (MSE).
- Desired response for all constrained signals: $d(\ell, k) \triangleq \mathbf{g}^H\mathbf{s}(\ell, k)$.
- The beamformer output: $y(\ell, k) = \mathbf{w}^H\mathbf{z}(\ell, k)$.
- MSE: $E\{|d(\ell) - y(\ell)|^2\}$. 
Speech enhancement with a Single Source I

Speech Distortion Weighted Multichannel Wiener Filter (SDW-MWF)

[Doclo and Moonen, 2002b]; [Spriet et al., 2004]; [Doclo et al., 2005]
Speech enhancement with a Single Source II

Speech Distortion Weighted Multichannel Wiener Filter (SDW-MWF)

[Doclo and Moonen, 2002b]; [Spriet et al., 2004]; [Doclo et al., 2005]

### The Multichannel Wiener Filter (MWF) Criterion

\[
J_w \triangleq E \left\{ |d(\ell) - y(\ell)|^2 \right\} = \left| g - (h^d)^H w \right|^2 \phi_{sd}^{-1} \phi_{dd} + w^H \Phi_{nn} w
\]

### The Speech Distortion Weighted (SDW)-MWF Criterion

\[
J_{SDW-MWF} = \left| g - (h^d)^H w \right|^2 \phi_{sd}^{-1} \phi_{dd} + \mu w^H \Phi_{nn} w
\]

### The Speech Distortion Weighted (SDW)-MWF Solution

\[
w = \frac{\phi_{sd}^{-1} \Phi_{nn}^{-1} h^d}{\mu + \phi_{sd} (h^d)^H \phi_{nn}^{-1} h^d} g
\]
Speech Enhancement with Multiple Sources I

[Markovich-Golan et al., 2012b]
Speech Enhancement with Multiple Sources II

[Markovich-Golan et al., 2012b]

**The MSDW-MWF Criterion**

\[
J_{\text{MSDW-MWF}} \triangleq (g - H^H w)^H \Lambda \Phi_{ss} (g - H^H w) + w^H \Phi_{nn} w
\]

- Diagonal weights matrix: \( \Lambda \triangleq \text{diag}\{\lambda_1, .., \lambda_P\} \).

**MSDW-MWF Beamformer**

\[
w \triangleq \left( H\Lambda\Phi_{ss}H^H + \Phi_{nn} \right)^{-1} H\Lambda\Phi_{ss}g
\]
Special Cases of $\Lambda$

### MWF

- $\Lambda = I$.
- $w = \Phi_{zz}^{-1} H \Phi_{ss} g$.

### SDW-MWF (Reminder: Single Source of Interest)

- $\Lambda = \mu^{-1}$.
- $w = \left( h^d \phi_{sd} s^d (h^d)^H + \mu \Phi_{nn} \right)^{-1} h^d \phi_{sd} s^d g$.
- $\lim_{\mu \to 0} w = \frac{\Phi_{nn}^{-1} h^d}{(h^d)^H \Phi_{nn}^{-1} h^d} g$ (MVDR eq. MPDR).

### LCMV

- $\Lambda = \mu^{-1} \Phi_{ss}^{-1}$.
- $\lim_{\mu \to 0} w = \Phi_{nn}^{-1} H \left( H^H \Phi_{nn}^{-1} H \right)^{-1} g$ (LCMV eq. LCMP).
The Generalized Sidelobe Canceller Implementation

For Constrained Minimization [Griffiths and Jim, 1982]

Split the Beamformer

- \( w = w_0 - w_n \).
- Constraints Subspace: \( w_0 \in \text{Span}\{C\} \).
- Null Subspace: \( w_n \in \mathcal{N}\{C\} \).
- \( w_n \triangleq Bq \).
- \( B \): \( M \times (M - P) \) matrix. Spans the Null Subspace.
- \( q \): vector of \( M - P \) filters.
- \( \Rightarrow w = w_0 - Bq \).
The Generalized Sidelobe Canceller Implementation

GSC Output

\[ y = w_0^H z - q^H B^H z \]

Constraints Subspace (\( w_0 \in \text{Span}\{C\} \)):

\[ w_0(\ell, k) \triangleq C(C^H C)^{-1} g \]

Null Subspace (columns of \( B \) span \( \mathcal{N}\{C\} \)):

\[ B(\ell, k) \triangleq I_{M \times M} - C(C^H C)^{-1} C^H; \quad \text{(verify} \ B^H C = 0\). \]

Noise Cancelling Filters (orthogonality principle):

\[ E \left\{ u \left( z^H w_0 - u^H q \right) \right\} \Rightarrow q(\ell, k) = \left( B^H \Phi_{zz} B \right)^{-1} B^H \Phi_{zz} w_0 \]
The GSC Implementation

The GSC Structure [Griffiths and Jim, 1982]

**GSC Blocks**

- **Fixed beamformer (FBF)** - satisfies the constraints \( \mathbf{w}_0 \).
- **Blocking matrix (BM)** - generates \( M - P \) unconstrained signals \( \mathbf{B} \).
- **Noise canceller (ANC)** - adaptively (LMS) suppresses the residual noise utilizing \( M - P \) degrees of freedom (DoF) \( \mathbf{q} \) [Widrow et al., 1975]; [Shynk, 1992].
GSC Implementation of the MVDR Beamformer

**Blocks** [Griffiths and Jim, 1982]:

\[
\mathbf{w}_0(\ell, k) = \frac{\mathbf{h}_d}{\|\mathbf{h}_d\|^2}
\]

\[
\mathbf{B}(\ell, k) \triangleq \mathbf{I}_{M \times M} - \frac{\mathbf{h}_d (\mathbf{h}_d^H)^H}{\|\mathbf{h}_d\|^2}
\]

\[
\mathbf{q}(\ell, k) = \left( \mathbf{B}^H \mathbf{\Phi}_{zz} \mathbf{B} \right)^{-1} \mathbf{B}^H \mathbf{\Phi}_{zz} \mathbf{w}_0
\]

\(\mathbf{q}(\ell, k)\) can be recursively updated using the LMS algorithm [Shynk, 1992].
The Relative Transfer Function GSC (TF-GSC)

Relax Dereverberation Requirement [Gannot et al., 2001]

Modified Constraint Set:

\[ C(\ell, k) = h^d(\ell, k); \quad \tilde{g}(\ell, k) = (h^d_0(\ell, k))^* \]

\[ \Rightarrow (h^d(\ell, k))^H w = (h^d_0(\ell, k))^* \]

Equivalent to:

\[ \tilde{C}(\ell, k) = \tilde{h}^d(\ell, k) \triangleq \frac{h^d}{h^d_0} = \begin{bmatrix} 1 & \frac{h^d_1}{h^d_0} & \ldots & \frac{h^d_{M-1}}{h^d_0} \end{bmatrix}^T \]

\[ g(\ell, k) = 1. \]

The Relative Transfer Function

\[ \tilde{h}^d(\ell, k) - \text{The ratio of all ATFs to the reference ATF (\#0 in this case).} \]
The GSC Implementation

Relative Transfer Function GSC

The Transfer Function GSC utilizing RTF I

[Gannot et al., 2001]

**FBF:**

\[ \mathbf{w}_0(\ell, k) = \frac{\tilde{h}^d}{\|\tilde{h}^d\|^2} \]

**Blocking matrix**

- Noise reference signals: \( \mathbf{u} = \mathbf{B}^H \mathbf{z} \).
- Efficient implementation of the BM with \( M - 1 \) filters exists.

\[
\mathbf{B}(\ell, k) = \begin{bmatrix}
-(\tilde{h}_1^d)^* & -(\tilde{h}_2^d)^* & \cdots & -(\tilde{h}_{M-1}^d)^* \\
1 & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 \\
\vdots & \ddots & \ddots & \ddots \\
0 & 0 & \cdots & 1
\end{bmatrix}
\]

Compactly, \( u_0 = 0; \ u_m = z_m - \tilde{h}_m^d z_0, \ m \neq 0. \)
The Transfer Function GSC utilizing RTF II

[Gannot et al., 2001]

Output signal:

\[ y(\ell, k) = h_d^d s_d^d + \text{residual noise and interference signals} \]

Tradeoff:

Noise reduction is sacrificed if dereverberation is required [Habets et al., 2010].
Multi-Constraint Beamformer
Based on LCMV Beamforming [Markovich et al., 2009]

Applications:
- Conference call scenario with multiple participants.
- Hands-free cellular phone conversation in a car environment with several passengers.
- Cocktail Party scenario, in which desired conversation blend with many simultaneous conversations.

Problem Formulation (Reminder):
\[ z = H^d s^d + H^i s^i + H^n s^n + n \]
GSC Formulation

GSC Implementation of the LCMV (exists [Breed and Strauss, 2002])

\[ w = w_0 - Bq \]

Fixed Beamformer (in Constraints Subspace)

\[ w_0 = C \left( C^H C \right)^{-1} g \]

Blocking Matrix (in Constraints Null Subspace)

\[ B = I_{M \times M} - C \left( C^H C \right)^{-1} C^H \]

Can be efficiently implemented: \((M - P) \times P\) filters [Markovich-Golan et al., 2012a].

Noise Canceler

\[ q = \left( B^H \Phi_{zz} B \right)^{-1} B^H \Phi_{zz}(\ell, k)w_0 \]
The Constraints Set

Original

\[ C \triangleq H = \begin{bmatrix} H^d & H^i & H^n \end{bmatrix} \]
\[ g \triangleq \begin{bmatrix} 1 & \ldots & 1 & 0 & \ldots & 0 \\ P_d & & P-P_d \end{bmatrix}^T \]

LCMV output

Since all directional signals are constrained, \( q = 0 \) if \( \Phi_{nn} \) is spatially-white.

\[ y = \sum_{j=1}^{P_d} s^d_j + \text{noise components} \]
An Equivalent Constraints Set

An orthonormal basis \( Q \):

- Noise+Interference Sources PSD (no desired sources):
  \[
  \Phi_{vv}(\ell, k) \triangleq H^i \Phi_{si si} (H^i)^H + H^n \Phi_{sn sn} (H^n)^H + \Phi_{nn}
  \]

- Eigenvalue decomposition: \( \Phi_{vv}(\ell, k) = E \Lambda E^H \).

- Replace \([H^i \ H^n]\) with \( Q \), comprised of the eigenvectors that correspond to the significant eigenvalues (# of significant eigenvalues is, hopefully, \( P_i + P_n \)).

\[
\dot{C}^H \mathbf{w} = \mathbf{g}
\]

\[
\dot{C} \triangleq \begin{bmatrix} H^d & Q \end{bmatrix}
\]
A Modified Constraints Set

Relax the dereverberation requirements using RTFs:

\[ \tilde{g} \triangleq \left[ (h_{10}^d)^* \ldots (h_{P_d0}^d)^* \frac{0 \ldots 0}{P_d \ P_{-P_d}} \right]^T \]

\[ \Rightarrow \tilde{h}_j^d \triangleq h_j^d / h_{j0}^d; \quad g \triangleq \left[ \begin{array}{c} 1 \ldots 1 \frac{0 \ldots 0}{P_d \ P_{-P_d}} \end{array} \right]^T \]

Hence, a modified constraints set: \( \tilde{C} \triangleq [\tilde{H}^d \ Q] \).

LCMV output:

\[ y = \sum_{j=1}^{P_d} h_{j0}^d s_j^d + \text{noise components} \]
Features & Drawbacks of the Proposed Beamformers

+ No need for sensor position calibration.
+ Beamformer components estimated from the received signals.
+ High amount of noise and interference reduction.
+ Low speech distortion.
- Number of filter coefficients to be estimated tends to be very large.
- Hence frame length tends to be large as well (can be mitigated at the expense of increased complexity. See CTF approximation).
- Limited performance in diffuse noise fields (can be mitigated by using postfiltering).

Performance Analysis

Theoretical and practical comparison of MVDR and LCMV beamformers can be found in [Markovich et al., 2008]; [Habets et al., 2009].
## Objective Performance Measures

Desired > nonstationary by 6dB; Desired > stationary by 13dB

<table>
<thead>
<tr>
<th>$T_{60}$</th>
<th>Source</th>
<th>FBF SIR</th>
<th>Total SIR</th>
<th>SSNR</th>
<th>LSD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$s_1^i$</td>
<td>$s_2^i$</td>
<td>$s_1^n$</td>
<td>$s_1^i$</td>
</tr>
<tr>
<td>150ms</td>
<td>$s_1^d$</td>
<td>18.8</td>
<td>22.4</td>
<td>19.1</td>
<td>18.5</td>
</tr>
<tr>
<td></td>
<td>$s_2^d$</td>
<td>18.7</td>
<td>22.3</td>
<td>19.1</td>
<td>18.7</td>
</tr>
<tr>
<td>200ms</td>
<td>$s_1^d$</td>
<td>18.1</td>
<td>20.6</td>
<td>19.5</td>
<td>18.3</td>
</tr>
<tr>
<td></td>
<td>$s_2^d$</td>
<td>18.1</td>
<td>20.7</td>
<td>19.6</td>
<td>18.9</td>
</tr>
<tr>
<td>250ms</td>
<td>$s_1^d$</td>
<td>18.5</td>
<td>19.8</td>
<td>19.9</td>
<td>18.4</td>
</tr>
<tr>
<td></td>
<td>$s_2^d$</td>
<td>18.5</td>
<td>19.8</td>
<td>19.9</td>
<td>19.4</td>
</tr>
<tr>
<td>300ms</td>
<td>$s_1^d$</td>
<td>17.6</td>
<td>17.6</td>
<td>19.5</td>
<td>18.3</td>
</tr>
<tr>
<td></td>
<td>$s_2^d$</td>
<td>17.4</td>
<td>17.5</td>
<td>19.3</td>
<td>18.6</td>
</tr>
</tbody>
</table>

**Table:** 2 desired sources, 2 competing speakers, 1 stationary noise source. The desired signal at the input is larger than the competing signal by 6dB and larger than the stationary noise by 13dB. 10 microphones simulated environment. LSD & SSNR are the distortion measures between desired signal components at the output and at the input microphone #1.
Single Desired Speaker
Directional Noise Field

Figure: Female (desired) and male (interference) with Directional noise. 8 microphones recorded at BIU acoustic lab set to $T_{60} = 300$ms.
Single Desired Speaker
Pseudo-Babble Noise Field

(a) Noisy at mic. #1
(b) Enhanced signal

Figure: Male (desired) and Female (interference) contaminated by pseudo-babble noise. 8 microphones recorded at BIU acoustic lab set to $T_{60} = 300 \text{ms}$. 

E.A.P. Habets (FAU) and S. Gannot (BIU) Linear and Parametric Mic. Array Proc. ICASSP 2013 42 / 113
Multi-Speaker

Figure: 1 desired source and 3 competing speakers. 8 microphones recorded at BIU acoustic lab set to $T_{60} = 300\text{ms}$. Approximately 20dB SIR and SNR improvement.
The Importance of the RTF

Features

- Generalizes the time difference of arrival (TDOA) to ratio of ATFs.
- Usually exhibits “better behaviour” than the ATF.
- RTF is equivalent to Interaural Transfer Function (ITF).
- **Drawback**: Non-causal (in severe cases can cause “pre-echo”).
Relative Transfer Function Estimation
Single Desired Source with Stationary Noise

System Perspective:

\[ z_m(\ell, k) = \tilde{h}_m^d(\ell, k)(\ell, k)z_0(\ell, k) + u_m(\ell, k) \]

System Identification:

\[ \hat{\Phi}_{zmz_0}(\ell, k) = \tilde{h}_m^d(\ell, k)\hat{\Phi}_{z_0z_0}(\ell, k) + \Phi_{umz_0}(\ell, k) + \varepsilon_m(\ell, k) \]

Estimation is Biased:

\[ u_m(\ell, k) \text{ and } z_0(\ell, k) \text{ are correlated} \Rightarrow \text{Biased estimator for } \tilde{h}_m^d(\ell, k). \]
Relative Transfer Function Estimation I

Based on Speech Non-stationarity [Shalvi and Weinstein, 1996]; [Gannot et al., 2001]

Assumptions:
- System is Time-Invariant.
- Noise has only stationary components.
- Speech is non-stationary (use frames $\ell_i$, $i = 1, \ldots, I$).

\[
\begin{bmatrix}
\hat{\Phi}_{zmz_0}(\ell_1, k) \\
\hat{\Phi}_{zmz_0}(\ell_2, k) \\
\vdots \\
\hat{\Phi}_{zmz_0}(\ell_I, k)
\end{bmatrix}
= 
\begin{bmatrix}
\hat{\Phi}_{z_0z_0}(\ell_1, k) & 1 \\
\hat{\Phi}_{z_0z_0}(\ell_2, k) & 1 \\
\vdots & \\
\hat{\Phi}_{z_0z_0}(\ell_I, k) & 1
\end{bmatrix}
\begin{bmatrix}
\tilde{h}_m^d(k) \\
\Phi_{umz_0}(k)
\end{bmatrix}
+ 
\begin{bmatrix}
\varepsilon_m(\ell_1, k) \\
\varepsilon_m(\ell_2, k) \\
\vdots \\
\varepsilon_m(\ell_I, k)
\end{bmatrix}
\]
Relative Transfer Function Estimation II
Based on Speech Non-stationarity [Shalvi and Weinstein, 1996]; [Gannot et al., 2001]

Solution
For \( m = 1, \ldots, M - 1 \):

\[
\hat{h}_m^d(k) = \frac{\langle \hat{\Phi}_{z_mz_0} \hat{\Phi}_{z_0z_0} \rangle (k) - \langle \hat{\Phi}_{z_mz_0} \rangle (k) \langle \hat{\Phi}_{z_0z_0} \rangle (k)}{\langle \hat{\Phi}_{z_0z_0} \rangle (k) - \langle \hat{\Phi}_{z_0z_0} \rangle^2 (k)}
\]

where, \( T_i \) the length of segment \( T_i \) and

\[
\langle \Psi \rangle (k) = \frac{\sum_{i=1}^{l} T_i \Psi(\ell_i, k)}{\sum_{i=1}^{l} T_i}.
\]

An extension to two nonstationary sources in stationary noise exists
[Reuven et al., 2008].
Alternative Estimation Procedures

- Assume direct-path model for the RIR and use TDOA estimation.
- Use speech presence probability and spectral subtraction [Cohen, 2004].
- ...
Multi-Sources Case [Markovich et al., 2009]

Implementing the GSC Necessitates:

- Desired sources RTFs, $\tilde{H}^d(\ell, k)$.
- Interferences subspace basis, $Q(\ell, k)$.

Assumptions and Observations

- The ATFs are slowly-time varying.
- Segments with non-overlapping activity of desired and interference speakers are available.
- Double-talk within the group is allowed.
- Stationary sources are always active.
Interferences Subspace Estimation

Step 1

**EVD and Pruning**

- Estimate the signals subspace at each time segment without any desired sources active

\[
\hat{\Phi}_{zz}(\ell_i, k) = \tilde{E}_i \Lambda_i \tilde{E}_i^H
\]

- All eigenvectors corresponding to “weak” eigenvalues are discarded
Interferences Subspace Estimation

Step 2

Union of Estimates

- Straightforward:
  \[ \mathbf{E}(k) \triangleq \bigcup_{i=1}^{N_{\text{seg}}} \bar{\mathbf{E}}_i(k) \]

- Practical use QRD

\[
\begin{bmatrix}
\bar{\mathbf{E}}_1(k) \bar{\Lambda}_{1}^{\frac{1}{2}}(k) & \cdots & \bar{\mathbf{E}}_{N_{\text{seg}}}(k) \bar{\Lambda}_{N_{\text{seg}}}^{\frac{1}{2}}(k)
\end{bmatrix}
\mathbf{P}(k) = \mathbf{Q}(k)\mathbf{R}(k)
\]

- Discard vectors from the basis \( \mathbf{Q}(k) \) that correspond to “weak” coefficients in \( \mathbf{R}(k) \).
EVD per Frame - Graphical Interpretation
Frame 1, strong eigenvectors
EVD per Frame - Graphical Interpretation

Frame 2, strong eigenvectors
EVD per Frame - Graphical Interpretation

Frame 3, strong eigenvectors
QRD Calculation

Graphical Interpretation
QRD Pruning
Graphical Interpretation
Desired Sources RTF Estimation
One Concurrent Desired Speaker

PSD Estimation

- Stationary noise PSD:

\[ \Phi_{zz}^{\text{stat}} = H^n \Phi_{ss}^n (H^n)^H + \Phi_{nn} \]

- One desired source \( (i_0) \), no non-stationary source:

\[ \hat{\Phi}_{zz}^{d,i_0} \approx \phi_{i_0}^{d} \hat{h}_{i_0}^{d} (\hat{h}_{i_0}^{d})^H + \Phi_{zz}^{\text{stat}} \]

Largest Generalized Eigenvector

\[ \hat{\Phi}_{zz}^{d,i_0} f_{i_0} = \lambda_{i_0} \Phi_{zz}^{\text{stat}} f_{i_0} \Rightarrow \hat{h}_{i_0}^{d} \triangleq \frac{\Phi_{zz}^{\text{stat}} f_{i_0}}{(\Phi_{zz}^{\text{stat}} f_{i_0})_0} \]
Multichannel Post-filtering (for single desired source)

Using matrix inversion lemma [Simmer et al., 2001]; [Doclo et al., 2010]

Why Postfiltering?

- In diffuse noise field multichannel processing is not enough!
- For nonstationary signals advanced single microphone spectral enhancement methods are beneficial [Cohen and Gannot, 2008].

MWF for estimating speech component at reference microphone (#0)

\[
\mathbf{w}_{\text{SDW-MWF}} = \frac{\phi_{d} \phi_{nn}^{-1} h_{d}^{*}}{\mu + \phi_{d} (h_{d}^{H}) \Phi_{nn}^{-1} h_{d}^{*}}
\]

where, \( \phi_{y_{s}y_{s}} = |h_{0}^{d}|^2 \phi_{d} \phi_{n} \) is the desired speech component at the MVDR output and \( \phi_{y_{n}y_{n}} \) is the respective noise output.
Zelinski Postfilter [Zelinski, 1988]

Assumptions

- Distortionless beamformer $\phi_{y_sy_s} = \phi_{s^d s^d}$.
- Spatially white noise field, $\Phi_{nn} = \phi_{nn}I$ (no other interference sources).
- Hence, $\phi_{z_iz_j} = \phi_{s^d s^d}$; $i \neq j$ & $\phi_{z_iz_i} = \phi_{s^d s^d} + \phi_{nn}$.

Estimated Wiener Postfilter

- Recursive estimation of the auto- and cross-spectra:
  $\hat{\phi}_{z_iz_j}(\ell) = \alpha \hat{\phi}_{z_iz_j}(\ell - 1) + (1 - \alpha) z_i(\ell) z_j^*(\ell)$.
- Zelinski’s postfilter:
  $$w_{Zel}(\ell, k) = \frac{2}{M(M-1)} \sum_{i=0}^{M-2} \sum_{j=i+1}^{M-1} \Re(\hat{\phi}_{z_iz_j}(\ell, k))$$
  $$\frac{1}{M} \sum_{i=0}^{M-1} \hat{\phi}_{z_iz_i}(\ell, k)$$

- Combined with Spectral Subtraction [Meyer and Simmer, 1997].
- Further developed and analyzed [Marro et al., 1998].
McCowan & Bourlard Postfilter  [McCowan and Bourlard, 2003]

Further Assumptions

- Noise field with known and isotropic coherence function, \( \phi_{n_i n_j} = \phi_{nn} \Gamma_{n_i n_j} \) (no other interference sources).
- Hence, \( \phi_{z_i z_j} = \phi_{s^d s^d} + \phi_{nn} \Gamma_{n_i n_j} \); \( i \neq j \) & \( \phi_{z_i z_i} = \phi_{z_j z_j} = \phi_{s^d s^d} + \phi_{nn} \).
- Diffuse noise field is usually assumed \( \Gamma_{n_i n_j}(\omega) = \text{Sinc}(\frac{\omega_{d_{ij}}}{c}) \).

Estimated Wiener Postfilter

- McCowan & Bourlard postfilter:

\[
\hat{\phi}_{s^d s^d}(\ell, k) = \frac{\Re(\hat{\phi}_{z_i z_j}) - 0.5\Re(\Gamma_{n_i n_j})(\hat{\phi}_{z_i z_i} + \hat{\phi}_{z_j z_j})}{1 - \Re(\Gamma_{n_i n_j})}
\]

\[
w_{MB}(\ell, k) = \frac{2}{M(M-1)} \sum_{i=0}^{M-2} \sum_{j=i+1}^{M-1} \hat{\phi}_{s^d s^d}
\]

\[
\triangleq \frac{1}{M} \sum_{i=0}^{M-1} \hat{\phi}_{z_i z_i}
\]
Improved Noise PSD Estimation

**Noise Over-estimation**
Both postfilters [Zelinski, 1988] and [McCowen and Bourlard, 2003] use over-estimated noise PSD, since they use the input signals rather than the beamformer output.

**Noise PSD at beamformer output [Leukimmitatis et al., 2006]**
Replace the denominator by:

\[
\hat{\phi}_{n_i n_j}(\ell, k) = \frac{0.5(\hat{\phi}_{z_i z_i} + \hat{\phi}_{z_j z_j}) - \Re(\hat{\phi}_{z_i z_j})}{1 - \Re(\Gamma_{n_i n_j})}
\]

\[
\hat{\phi}_{nn}(\ell, k) = \frac{2}{M(M-1)} \sum_{i=0}^{M-2} \sum_{j=i+1}^{M-1} \hat{\phi}_{n_i n_j}(\ell, k)
\]

\[
w_{\text{Leuk}}(\ell, k) = \frac{\hat{\phi}_{s_d s_d}^H}{\hat{\phi}_{s_d s_d} + \hat{\phi}_{nn} w_{\text{MVDR}}^H \Gamma_{nn} w_{\text{MVDR}}}
\]
Nonlinear Postfilter [Balan and Rosca, 2002]

Motivation
- Nonlinear processing has many advantages in speech enhancement.
- A plethora of nonlinear algorithms for single microphone speech enhancement exist.
- An extension to the multichannel case can be derived.

Sufficient Statistics
- Conditional p.d.f.:
  \[
  P_r(z|s^d; \phi_{s^d s^d}, \Phi_{nn}, h^d) = \frac{1}{\pi \Phi_{nn}} \exp\left\{ -(z - h^d s^d)^H \Phi_{nn}^{-1} (z - h^d s^d) \right\}
  \]
- MVDR output is sufficient statistics for \( s_d \): 
  \[
  T(z) = \frac{(h^d)^H \Phi_{nn}^{-1} z}{(h^d)^H \Phi_{nn}^{-1} h^d}
  \]
- \( P_r(\rho(s^d)|z) = P_r(\rho(s^d)|T(z)) \)
Nonlinear Postfilter \cite{BalanRosca2002}  

**Log Spectral Amplitude Estimator** extending \cite{EphraimMalah1985}

- Beamformer output: \( y = s + \frac{(h^d)^H \Phi_{nn}^{-1} h}{(h^d)^H \Phi_{nn}^{-1} h^d} n \).
- LSA criterion:
  \[
  |\hat{s}^d| = \exp\{E\{\log(|s^d|)|z\}\} = \exp\{E\{\log(|s^d|)|T(z)\}\}
  \]
- Estimator:
  \[
  |\hat{s}^d| = \frac{\xi}{1 + \xi} \exp \left\{ \frac{1}{2} \int_{\nu}^{\infty} \frac{e^{-t}}{t} dt \right\} |y|
  \]
  where \( \xi \equiv \phi_{s^d s^d}(h^d)^H \Phi_{nn}^{-1} h^d \) is the a priori SNR,
  \( \gamma \equiv |y|^2 (h^d)^H \Phi_{nn}^{-1} h^d \) is the a posteriori SNR and \( \nu = \frac{\xi \gamma}{1 + \xi} \).
- Final estimator is obtained by \( \hat{s}^d = |\hat{s}^d| e^{\zeta}(y) \).
- Gives motivation to the algorithm presented next.
GSC & Speech Presence Probability based Postfiltering

[Cohen et al., 2003]; [Gannot and Cohen, 2004]

- Use main output and reference noise signals to update the speech presence probability.
- Feed backward the decision to update GSC parameters.
- Use the speech presence probability to update the OM-LSA [Cohen and Berdugo, 2001] algorithm for residual noise reduction.
Hypothesis Test

\[ \Lambda_Y(k, \ell) > \Lambda_0 \]

- \( \Lambda_Y \) - local non-stationarity at beamformer output.
- \( \Lambda_U \) - local non-stationarity at noise reference signals.
- \( \Omega \) - The transient beam-to-reference ratio (TBRR).
- \( \gamma_s \) - a posteriori SNR at the beamformer output.

\( \Omega(k, \ell) < \Omega_{\text{low}} \) or \( \gamma_s(k, \ell) < 1 \)

\( \Omega(k, \ell) > \Omega_{\text{high}} \) and \( \gamma_s(k, \ell) > \gamma_0 \)
Experimental Study I
Car Scenario

(a) Clean speech

(b) Noisy at mic. #1
Experimental Study II

Car Scenario

Figure: Speech utterance: “Dial: One, Two, Three, Four, Five, Six, Seven, Eight”. Car with open windows equipped with 4 microphones.
The Convolutive TF-GSC [Talmon et al., 2009a]

Motivation

The GSC [Griffiths and Jim, 1982]

Implemented in time-domain and assumes delay-only propagation. Hence speech distortion is expected.

The TF-GSC [Gannot et al., 2001]

- The RTFs are incorporated into the GSC beamformer.
- Adaptation to reverberant environments obtained by time-frequency implementation.
- For high $T_{60}$:
  - The RIRs and the respective relative RIRs become very long.
  - Multiplication in frequency-domain (MTF approximation) is only valid if the time frames are significantly larger than the relative RIR.
  - In practice, short frames are used, resulting in inaccurate representation of the RTF and hence performance degradation.
The Convolutive TF-GSC \cite{Talmon et al., 2009a} II

Motivation

**Time-Domain MVDR** \cite{Chen et al., 2008}
- Full relative RIR is taken into account.
- Theoretically, optimal MVDR in reverberant environment.
- The full-length RTF estimation requires:
  - Very long observations, limiting the ability to work in dynamic environments and to track time-variations.
  - Large computational complexity.
- In practice, the speech source RIRs are modelled as shorter filters.

**STFT Implementation** \cite{Talmon et al., 2009a} Enables:
- Short frames.
- Long relative RIRs.
In the STFT Domain:

- Formulate the problem using system representation in the STFT domain [Avargel and Cohen, 2007].

- Build a GSC scheme (a TF-GSC extension).

- Suggest practical solutions using approximations. Specifically, show solutions under the MTF and CTF approximations.

- Incorporate the RTF identification based on the CTF model [Talmon et al., 2009b] and compare experimental results with the TF-GSC.

- Currently, applicable only to single desired source.
Signal Model I

**Time Domain**

\[ z_m(t) = s^d(t) * h^d_m(t) + n_m(t) = \tilde{s}^d(t) * \tilde{h}^d_m(t) + n_m(t) \]
- \( \tilde{s}^d(t) = s^d(t) * h^d_1(t) \) - Desired signal component at microphone #1.
- \( \tilde{h}^d_m(t) \) - relative RIR between microphone #1 and microphone #\( m \).

**STFT Domain**

\[ z_m(\ell, k) = \sum_{k'=0}^{N_{FFT}-1} \sum_{p'} \tilde{h}_m(\ell', k', k) \tilde{s}_d(\ell - \ell', k') + n_m(\ell, k) \]

Concatenating successive signal frames:

\[ z_m(k) = \sum_{k'=0}^{N_{FFT}-1} \tilde{H}_m(k', k) \tilde{s}_d(k') + n_m(k) \]

\[ CTF \approx \tilde{H}_m(k) \tilde{s}_d(k) + n_m(k) \]
CTF vs. MTF

Signal Model II

Beamforming in the STFT Domain

\[ \hat{s}_d(k) = \sum_{m=1}^{M} \sum_{k'=0}^{N_{\text{FFT}}-1} W_m^H(k', k) z_m(k') \approx \sum_{m=1}^{M} W_m^H(k) z_m(k) \]

MVDR & GSC

- Constrained power minimization (MVDR) can be defined.
- GSC structure exists (\# of constraints < \# of measurements).
- Similarly to the TF-GSC, \( \hat{H}_m(k', k) \) can be identified [Talmon et al., 2009b].
Setup

Comparing the proposed method to the TF-GSC:

- **Image method** ([Allen and Berkley, 1979], implemented by [Habets, 2006]).
- Array of 5 microphones.
- Reverberation time $T_{60} = 0.5\text{s}$.
- **TF-GSC**:
  - Frame length - $N = 512$.
  - RTF length - 500.
  - Noise Canceller length - 450.
- **CTF-GSC**:
  - In FBF and BM - $N = 512, 50\%$ overlap.
  - In adaptive NC - $N = 512, 75\%$ overlap.
  - RTF length - 5 frames.
Signal Blocking

The signal blocking factor (SBF) is defined by:

\[
SBF = 10 \log_{10} \frac{E \left\{ \left( \tilde{s}_d(t) \right)^2 \right\}}{\text{Mean}_m E \left\{ u_m^2(t) \right\}}
\]

where \( u_m(t); \ m = 2, \ldots, M \) are the blocking matrix outputs.

<table>
<thead>
<tr>
<th>The blocking ability [dB] (known RTF)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>TF</td>
</tr>
<tr>
<td>CTF</td>
</tr>
</tbody>
</table>
Known RTF, Input SNR=0dB I

(a) Reverberated speech at microphone #1.  
(b) Noisy signal at microphone #1.
CTF vs. MTF

Experimental Results

Known RTF, Input SNR=0dB II

(c) TF-GSC output.

(d) CTF-GSC output.
Output SNR and Noise Reduction [dB] for known RTF

<table>
<thead>
<tr>
<th>In SNR</th>
<th>SNR</th>
<th>NR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TF-GSC</td>
<td>CTF-GSC</td>
</tr>
<tr>
<td>-5</td>
<td>3.8</td>
<td>8.5</td>
</tr>
<tr>
<td>-2.5</td>
<td>5.2</td>
<td>10.0</td>
</tr>
<tr>
<td>0</td>
<td>6.2</td>
<td>11.2</td>
</tr>
<tr>
<td>2.5</td>
<td>7.0</td>
<td>12.0</td>
</tr>
<tr>
<td>5</td>
<td>7.9</td>
<td>12.5</td>
</tr>
</tbody>
</table>

Input SNR [dB] | Output SNR [dB] | TF-GSC | CTF-GSC

-5 | 0 | 5 | 0 | 5 | 0 | 5

E.A.P. Habets (FAU) and S. Gannot (BIU) Linear and Parametric Mic. Array Proc. ICASSP 2013 77 / 113
Estimated RTF

Signal Blocking

Figure: SBF curves obtained by the RTF identification method based on the MTF and CTF models.
Identified RTF, Input SNR=5dB I

(a) Reverberated speech at microphone #1.  (b) Noisy signal at microphone #1.
Identified RTF, Input SNR=5dB II

(c) TF-GSC output.

(d) CTF-GSC output.
Identified RTF, Input SNR=5dB III

(e) BM output TF-GSC.

(f) BM output CTF-GSC.
## Summary

### Output SNR and Noise Reduction [dB] for estimated RTF

<table>
<thead>
<tr>
<th>In SNR</th>
<th>SNR TF-GSC</th>
<th>SNR CTF-GSC</th>
<th>NR TF-GSC</th>
<th>NR CTF-GSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>-5</td>
<td>0.9</td>
<td>-0.4</td>
<td>-2.2</td>
<td>-4.7</td>
</tr>
<tr>
<td>-2.5</td>
<td>2.6</td>
<td>2.5</td>
<td>-2.9</td>
<td>-5.3</td>
</tr>
<tr>
<td>0</td>
<td>3.8</td>
<td>5.1</td>
<td>-3.2</td>
<td>-6.0</td>
</tr>
<tr>
<td>2.5</td>
<td>6.5</td>
<td>7.3</td>
<td>-4.0</td>
<td>-6.8</td>
</tr>
<tr>
<td>5</td>
<td>6.8</td>
<td>9.1</td>
<td>-4.3</td>
<td>-7.7</td>
</tr>
<tr>
<td>7.5</td>
<td>7.4</td>
<td>10.3</td>
<td>-4.8</td>
<td>-8.4</td>
</tr>
<tr>
<td>10</td>
<td>7.8</td>
<td>11.0</td>
<td>-5.5</td>
<td>-9.1</td>
</tr>
</tbody>
</table>

![Graph showing output SNR and noise reduction for different input SNRs for TF-GSC and CTF-GSC techniques.]
Dynamic Scenario [Markovich-Golan et al., 2010]

Subspace tracking of Multiple Sources

**Goal**

Extract desired moving speakers from a mixture of speakers using the LCMV beamformer.

**Working hypothesis**

- Activity indicator for desired speech signals is available.
- Availability of time segments with nonconcurrent desired and interfering speakers.
- “Stable” subspaces represent static speakers with high probability.

**Features**

- Tracking ability using projection approximation subspace tracking deflation (PASTd) [Yang, 1995].
- Double talk within group allowed during estimation.
- “Expiry time” for outdated basis vectors.
Dynamic Scenario

Tracking Procedure

**LCMV beamformer**

Definitions (Reminder)

\[
w = \Phi_{zz}^{-1} C \left( C^\dagger \Phi_{zz}^{-1} C \right)^{-1} g
\]

**Straightforward Constraints Set**

\[
C = \begin{bmatrix} H^d & H^i \end{bmatrix} \quad g = \begin{bmatrix} 1_{1 \times P_d} & 0_{1 \times P_i} \end{bmatrix}^T
\]

**Modified Constraints Set**

\[
\tilde{C} = \begin{bmatrix} Q^d & Q^i \end{bmatrix} \quad \tilde{g} = \begin{bmatrix} (Q^d_{1,1})^* & \cdots & (Q^d_{P_d,1})^* & 0_{1 \times P_i} \end{bmatrix}^T
\]

where \( Q^d, Q^i \) - bases for desired and interfering subspaces.

**Output**

\[
y(\ell, k) = \sum_{j=1}^{P_d} h^d_{j,1}(\ell, k) s^d_j(\ell, k) + \text{residual noise}
\]
Tracking Scheme I

Forgetting Factor Consideration

- Tracking $Q^d$ and $Q^i$ is a variant of the PASTd algorithm [Yang, 1995] with pre-whitening.
- Forgetting factor $\beta$ controls the adaptation, with $N_{\beta} = \frac{1}{1-\beta}$ the algorithm’s memory length.
- Standard PASTd suffers from contradicting requirements for $\beta$:
  - Fast adaptation $\Rightarrow$ small $\beta$.
  - Long memory $\Rightarrow$ large $\beta$.
- The contradicting requirements can be mitigated by combined tracking scheme.
Tracking Scheme II

Short & Long Memory

- Use short memory PASTd for fast adaptation of the instantaneous subspace of the $x$th group of signals, $\tilde{Q}^x(\ell, k)$.
- Declare stable subspaces, $Q^x(\ell, k)$, if the basis is valid for more than pre-defined number of frames.
  $I_{\text{stable}}^x(\ell)$ - Indicator for stable subspace of the $x$th group.
- Subspace union of the valid stable subspaces and the instantaneous subspace using QRD.
- Attribute an expiration time for each stable subspace.
Classification of Subspace Stability

- The energy of the projected signals onto the instantaneous subspace $\tilde{Q}_x(\ell, k)$ (integrated over past $N_\beta$ frames) consists of most of the signals' energy.

- $I^x_{\text{stable}}(\ell) = 1$ if the aggregated energy of the projected signals onto the instantaneous subspace (integrated over past $N_{\text{stable}} \gg N_\beta$ frames) consists of most of the signals' energy.
Tracking Example

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Stable subspace indicator</th>
<th>Instantaneous subspace rank</th>
<th>Union subspace rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>0 1 2 1 2 0 1 2</td>
<td>0 1 2 3 1 2 1 2</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Legend
- Static speaker
- Moving speaker
- Expiration time

subspace rank
Instantaneous rank
Union subspace rank

E.A.P. Habets (FAU) and S. Gannot (BIU)
Experimental Study

Results

Figure: 2 concurrent desired speakers and 2 competing speakers. 8 microphones recorded at BIU acoustic lab set to $T_{60} = 300\text{ms}$.

(a) Noisy at mic. #1

(b) Enhanced signal
Binaural LCMV Beamformer

Hadad, Gannot, Doclo, 2012

**Motivation**

- **Duplicate** the LCMV beamformer at both ears utilizing all microphones.
- The concept of RTF can be extended and used for preservation of binaural cues (ILD & ITD).
- Efficient implementation by block sharing.
Problem Formulation

Microphone Signals

\[ z = H^d s^d + H^i s^i + v \]

Left & Right Reference Microphones

\[ z_\ell = e^H_\ell z; \quad z_r = e^H_r z \]

where

\[ e_\ell = \begin{cases} 1 & m = m_\ell \\ 0 & \text{otherwise} \end{cases} \quad e_r = \begin{cases} 1 & m = m_r \\ 0 & \text{otherwise} \end{cases} \]

Binaural Spatial Filters

\[ y_\ell = w^H_\ell z; \quad y_r = w^H_r z. \]
Double LCMV Criterion

**Two BFs Utilizing All Microphones**

\[ \mathbf{w}_\ell = \text{LCMV}(\mathbf{z}; \mathbf{C}, \mathbf{g}_\ell); \quad \mathbf{w}_r = \text{LCMV}(\mathbf{z}; \mathbf{C}, \mathbf{g}_r) \]

**Orthonormal Basis for the ATFs**

\[ \{ \mathbf{H}_d = \mathbf{Q}_d \Theta_d; \quad \mathbf{H}_i = \mathbf{Q}_i \Theta_i \} \Rightarrow \mathbf{C} = \begin{bmatrix} \mathbf{Q}_d & \mathbf{Q}_i \end{bmatrix} \]

**Left & Right Response Vectors**

Apply dereverberation relaxation utilizing RTFs.

**Cue Gain Factors:**

Desired response \(0 < \eta \approx 1\); \quad \text{Interference response } 0 < \mu \ll 1
Interaural Signal Ratio (ISR)

**Input ISR**

\[
\text{ISR}^\text{in} = \frac{z_\ell}{z_r} = \frac{e_\ell^\dagger (H_d s_d + H_i s_i)}{e_r^\dagger (H_d s_d + H_i s_i)}.
\]

**Output ISR (in our implementation)**

\[
\text{ISR}^\text{out} = \frac{y_\ell}{y_r} = \frac{e_\ell^\dagger (\eta H_d s_d + \mu H_i s_i)}{e_r^\dagger (\eta H_d s_d + \mu H_i s_i)}.
\]
ISR vs. ITF

Properties

- Single source case: $\text{ISR}^{\text{out}} = \text{ISR}^{\text{in}}$ and ISR identifies with the ITF.
- Only one group is active $\Rightarrow$ spatial cues of the group maintained.
- Speech sparsity in STFT domain $\Rightarrow$ cues are preserved also for arbitrary activity pattern.
- Binaural cue preservation is only guaranteed for the constrained sources.
- Unconstrained stationary noise sources and residual (constrained) interference sources will “inherit” the input cues of the dominant source.
- $0 < \mu \ll 1$ will mask the artifacts resulting from leakage.
Block Diagram

\[ z \]

\[ M \]

\[ z \]

\[ M - N_d - N_i \]

\[ N_d + N_i \]

\[ g_r \]

\[ g_\ell \]

\[ q_r \]

\[ q_\ell \]

\[ y_r \]

\[ y_\ell \]
Setup

- **Hearing device:**
  - 2 hearing aid devices mounted on B&K HATS, with 2 microphones, 2cm inter-distance.
  - A $9 \times 5$ utility device with 4 mics. at the corners, average distance 3.5cm. The device placed on a table at a distance of 0.5m.

- **Signals:**
  - 1 desired speaker, $\theta_d = 30^\circ$, 1m (constrained).
  - 1 interference speaker at $\theta_i = -70^\circ$, 1m (constrained).
  - 1 directional stationary noise, $\theta_n = -40^\circ$, 2.5m (unconstrained).
  - SIR=0dB, SNR=14dB.

- **Acoustic lab:**
  - Dimensions $6 \times 6 \times 2.4$; Controllable reverb. time $T_{60} = 0.3s$.

- **STFT:**
  - Sampling frequency 8kHz, 4096 points, 75% overlap.

- **Algorithm Cue gain factors:**
  - Desired speech - $\eta = 1$.
  - Interference speech - $\mu = 0.1$ (20dB attenuation).
Sonograms

(a) Desired speaker, reference mic.

(b) Received reference microphones

(c) BLCMV outputs
ILD & ITD Preservation

(Faller and Merimaa, 2004)

(a) Noisy input

(b) Enhanced output

(c) Desired input

(d) Desired output
ILD & ITD Preservation  (Faller and Merimaa, 2004)

(e) Interference input

(f) Interference output

(g) Stationary input

(h) Stationary output
Audio Samples

Available at:

http://www.eng.biu.ac.il/gannot/speech-enhancement/
Features

- Controlled and acoustically isolated environment.
- 60 double-sided panels control the reverberation time.
- Equipped with microphone arrays, loudspeakers, measurement and acquisition equipment.
- Enables fast testing, implementation and verification of algorithms.
BIU Acoustics Lab: Picture Gallery
Thanks to my Collaborators

1. Shmulik Markovich-Golan
2. Prof. Israel Cohen
3. Prof. Ronen Talmon
4. David Levin
5. Prof. Emanuël Habets
6. Elior Hadad
7. Prof. Jacob Benesty
8. and many more...
References and Further Reading I

A signal subspace tracking algorithm for microphone array processing of speech.

Image method for efficiently simulating small-room acoustics.

On multiplicative transfer function approximation in the short-time Fourier transform domain.

Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase.
In *IEEE Workshop on Sensor Array and Multichannel Signal Processing*, pages 209–213, Rosslyn, Virginia, USA.

*Micophone array signal processing*.
Springer.

*Springer handbook of speech processing*.
Springer Verlag.

*Speech Enhancement*.
References and Further Reading II

*Microphone Arrays: Signal Processing Techniques and Applications.*
Springer-Verlag, Berlin.

A short proof of the equivalence of lcmv and gsc beamforming.

TRINICON: A versatile framework for multichannel blind signal processing.

A minimum distortion noise reduction algorithm with multiple microphones.

Relative transfer function identification using speech signals.

*Speech processing in modern communication: Challenges and perspectives.*
Topics in signal processing. Springer.

Speech enhancement for non-stationary noise environments.
*Springer Handbook of Speech Processing and Speech Communication*, chapter Spectral enhancement methods. In [Benesty et al., 2008b].


Design of far-field and near-field broadband beamformers using eigenfilters.

In [Benesty et al., 2005].

Microphone array systems for hands-free telecommunication.

Speech enhancement using a minimum mean-square error log-spectral amplitude estimator.

Derivative constraints for broad-band element space antenna array processors.

An algorithm for linearly constrained adaptive array processing.
*Proceedings of the IEEE, 60*(8):926–935.

Signal enhancement using beamforming and nonstationarity with applications to speech.
Speech enhancement based on the general transfer function GSC and postfiltering.

*Springer Handbook of Speech Processing and Speech Communication*, chapter Adaptive beamforming and postfiltering. In [Benesty et al., 2008b].


Optimum design of directive antenna arrays subject to random variations.

An alternative approach to linearly constrained adaptive beamforming.

New insights into the MVDR beamformer in room acoustics.

On the application of the LCMV beamformer to speech enhancement.
In *The IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPA)*, pages 141–144, New Paltz, New York, USA.
Room impulse response (RIR) generator.

Handbook on array processing and sensor networks, volume 63.

Sound capture for human/machine interfaces - Practical aspects of microphone array signal processing, volume 315 of Lecture Notes in Control and Information Sciences.
Springer, Heidelberg, Germany.

A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters.

Sound capture from spatial volumes: Matched-filter processing of microphone arrays having randomly-distributed sensors.
In IEEE Int. Conf. Acoust. Speech and Sig. Proc. (ICASSP), pages 917–920, Atlanta, Georgia, USA.

Adaptive microphone-array system for noise reduction.


A weighted multichannel Wiener filter for multiple sources scenarios.
In *The IEEE 27th Convention of IEEE Israel (IEEEI)*, Eilat, Israel.
best student paper award.

Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering.

Microphone array post-filter based on noise field coherence.

Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction.

Adaptive Array Noise Suppression of Handsfree Speaker Input in Cars.

Maximum directivity proof for three-dimensional arrays.

Dual-source transfer-function generalized sidelobe canceller.
System identification using nonstationary signals.  

Frequency-domain and multirate and adaptive filtering.  

*Post-Filtering Techniques*, chapter 3, pages 39–60.  
In [Brandstein and Ward, 2001].

Adaptive optimization of microphone arrays under a nonlinear constraint.  

Spatially pre-processed speech distortion weighted multi-channel wiener filtering for noise reduction.  

Convolutive transfer function generalized sidelobe canceler.  

Relative transfer function identification using convolutive transfer function approximation.  
Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings.
In Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pages 833–836, Albuquerque, New Mexico, USA. IEEE.

Wiley, New York.

Beamforming: A versatile approach to spatial filtering.

Blind acoustic beamforming based on generalized eigenvalue decomposition.

Adaptive noise cancelling: Principals and applications.
Proceeding of the IEEE, 63(12):1692–1716.

Projection Approximation Subspace Tracking.

A microphone array with adaptive post-filtering for noise reduction in reverberant rooms.
In IEEE Int. Conf. Acoust. Speech and Sig. Proc. (ICASSP), pages 2578–2581, New-York, USA.