# Multi-Microphone Speech Enhancement

## Using LCMV Beamformers

Sharon Gannot

Faculty of Engineering, Bar-Ilan University, Israel

IWAENC, 6.9.2012, Aachen, Germany

# Speech Enhancement and Source Extraction

## Utilizing Microphone Arrays

### Applications

1. Hands-free communications.
2. Teleconference.
3. Skype calls.
4. Hearing aids.
5. Eavesdropping.

### Why is it a Difficult Task?

1. Speech is a non-stationary signal that has high dynamic range.
2. Very long acoustic path (reverberation).
3. Time varying acoustic path.
4. Microphone arrays impose high computational burden.
5. Large (and distributed) arrays require large communication bandwidth.

# Room Acoustics Essentials

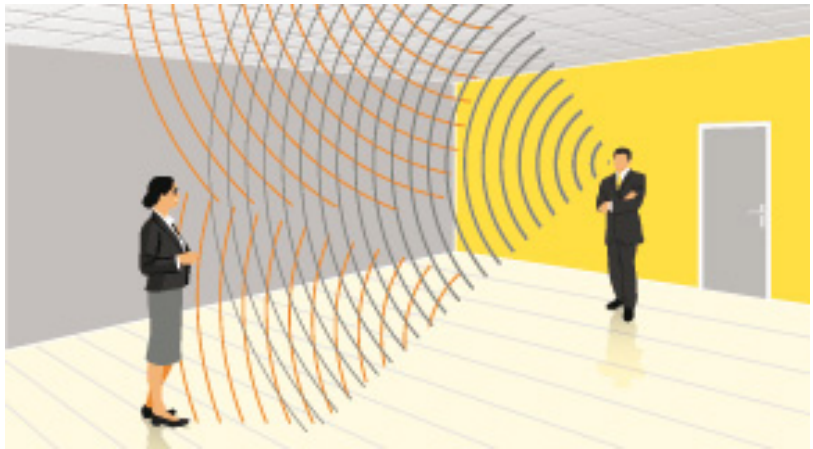
## Reverberation

- Late reflections tend to be diffused, hence performance degradation.
- Beamforming: higher processing burden and high latency in STFT implementation.
- Deteriorates intelligibility, quality and ASR performance.

## Sound Fields

- Directional Room impulse response relates source and microphones.
- Uncorrelated Signals on microphone are uncorrelated.
- Diffused Sound is coming from all directions.

Dal-Degan, Prati, 1988; Habets, Gannot, 2007



# Spatial Processing

## Design Criteria

### Tailored to Speech Applications

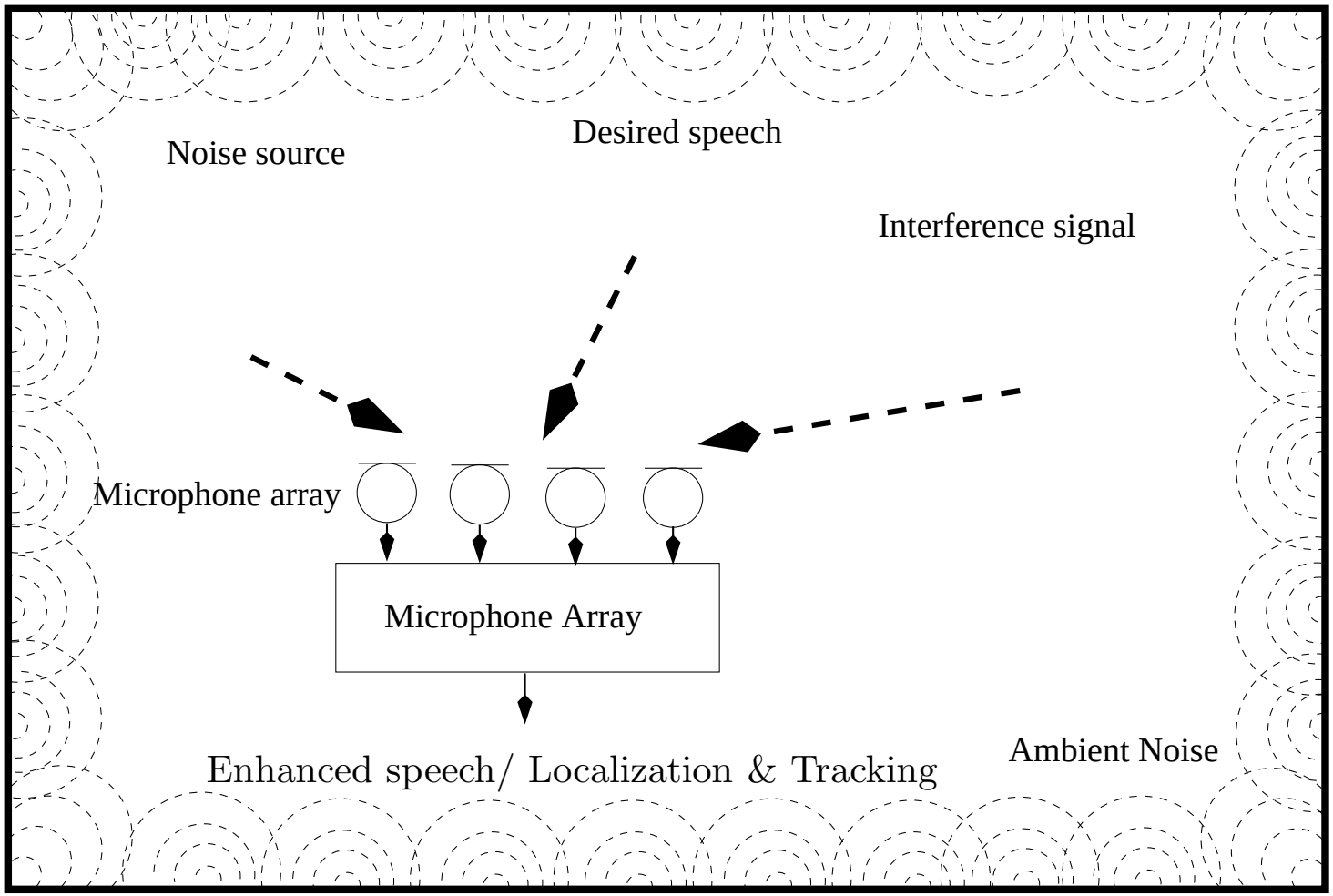
1. BSS, CASA. ▸ Other methods
2. Adaptive optimization Sondhi, Elko, 1986; Kaneda, Ohga, 1986.
3. Minimum variance distortionless response (MVDR) and GSC.

   Van Compernolle, 1990; Affes and Grenier, 1997; Nordholm et al., 1999; Hoshuyama et al., 1999 Gannot et al., 2001; Herbordt et al., 2005
4. Minimum mean square error (MMSE) - GSVD based spatial Wiener filter. Doclo, Moonen, 2002
5. Speech distortion weighted multichannel Wiener filter (SDW-MWF).

   Spriet et al., 2004
6. Maximum signal to noise ratio (SNR). Warsitz and Haeb-Umbach, 2005
7. Linearly constrained minimum variance (LCMV). Markovich et al., 2009

# Problem Formulation

## Extraction of Desired Speaker Signal(s) in Multi-Interference Reverberant Environment



# Talk Outline

- Spatial processors ("beamformers") based on the linearly constrained minimum variance (LCMV) criterion.
- Special cases:
  - Enhancing single desired source (MVDR).
  - Extracting desired source(s) in multiple competing speaker environment.
- Implementation: Generalized sidelobe canceller (GSC) in the short-time Fourier transform (STFT) domain.
- The relative transfer function (RTF): importance and estimation procedures.
- The applicability of the LCMV to binaural processing.

# Problem Formulation in the STFT Domain

## Microphone Signals ($m = 1, \ldots, M$):

$$z_m(\ell, k) = \sum_{j=1}^{N_d} s_j^d(\ell, k) h_{jm}^d(\ell, k)$$
$$+ \sum_{j=1}^{N_i} s_j^i(\ell, k) h_{jm}^i(\ell, k) + \sum_{j=1}^{N_n} s_j^n(\ell, k) h_{jm}^n(k) + n_m(\ell, k)$$

## Vector Formulation

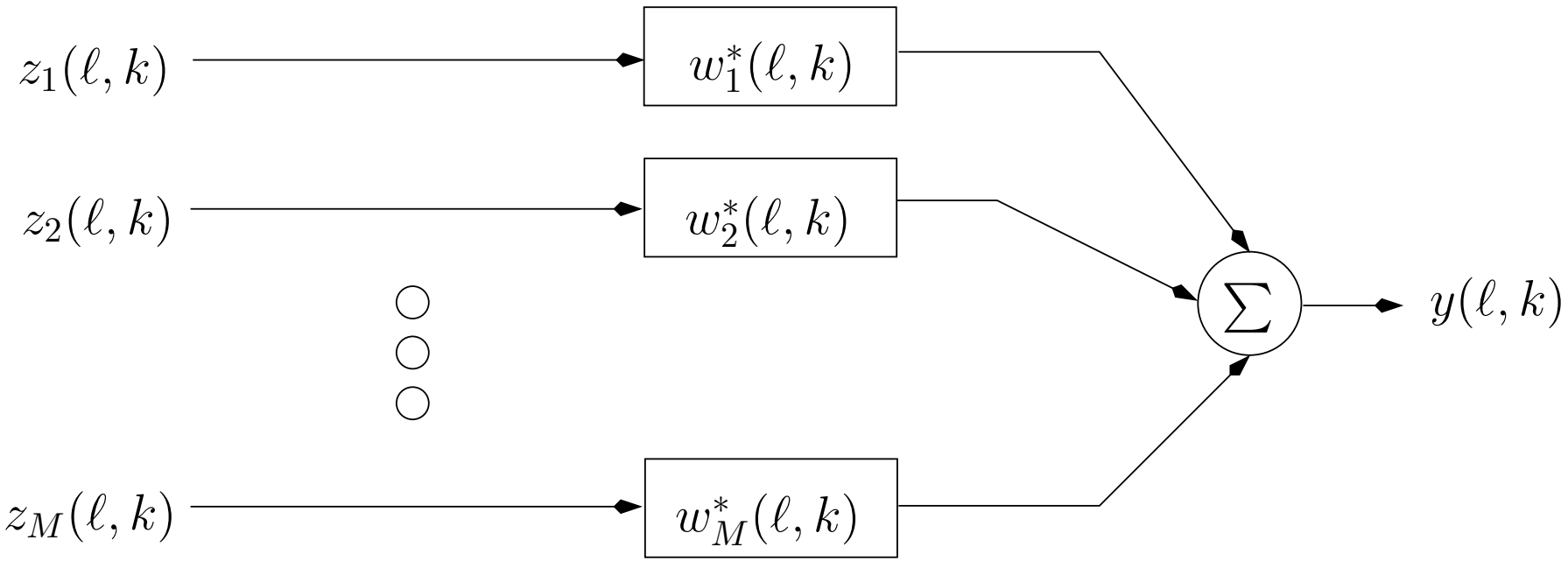
$$\mathbf{z}(\ell, k) = \mathbf{H}^d(\ell, k)\mathbf{s}^d(\ell, k) + \mathbf{H}^i(k)\mathbf{s}^i(\ell, k) + \underbrace{\mathbf{H}^n(\ell, k)\mathbf{s}^n(\ell, k) + \mathbf{n}(\ell, k)}_{\mathbf{v}(\ell,k)\text{; stationary component}}$$

$$N = N_d + N_i + N_n \leq M$$

# Spatial Filters

## Filter and Combine

$$y(\ell, k) = \mathbf{w}^H(\ell, k)\mathbf{z}(\ell, k).$$



# The Linearly Constrained Minimum Variance Beamformer

Er, Cantoni, 1983; Van Veen, Buckley, 1988

$$y(\ell, k) = \mathbf{w}^H(\ell, k)\mathbf{z}(\ell, k)$$

## LCMV Criterion

- Let $\mathbf{\Phi}_{zz}(\ell, k)$ be the received signals correlation matrix.
- Minimize output power

$$\mathbf{w}^H(\ell, k)\mathbf{\Phi}_{zz}(\ell, k)\mathbf{w}(\ell, k),$$

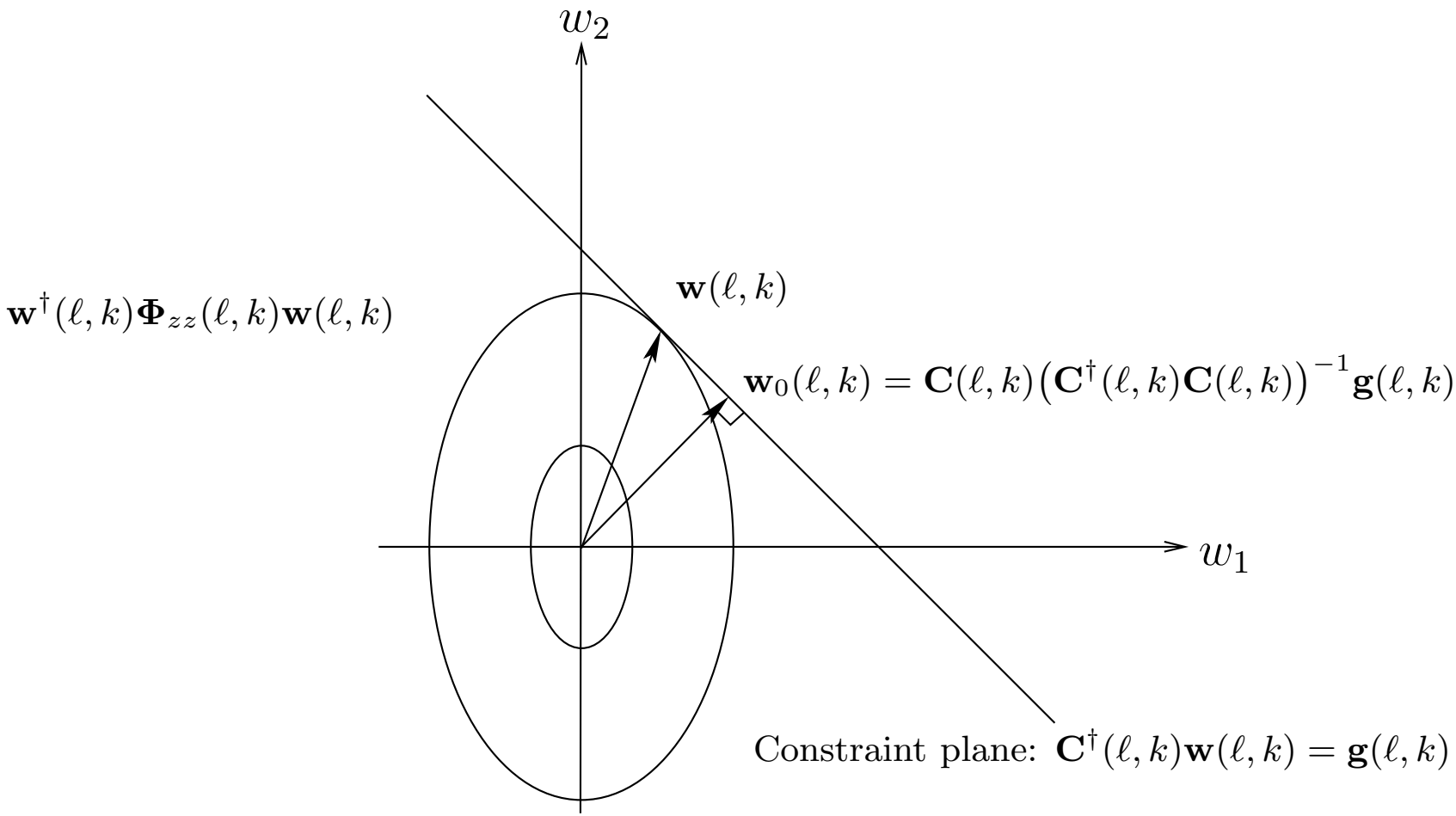
such that the linear constraint set is satisfied:

$$\mathbf{C}^H(\ell, k)\mathbf{w}(\ell, k) = \mathbf{g}(\ell, k).$$

Closed-form solution exists (but we hardly use them)

# LCMV Minimization

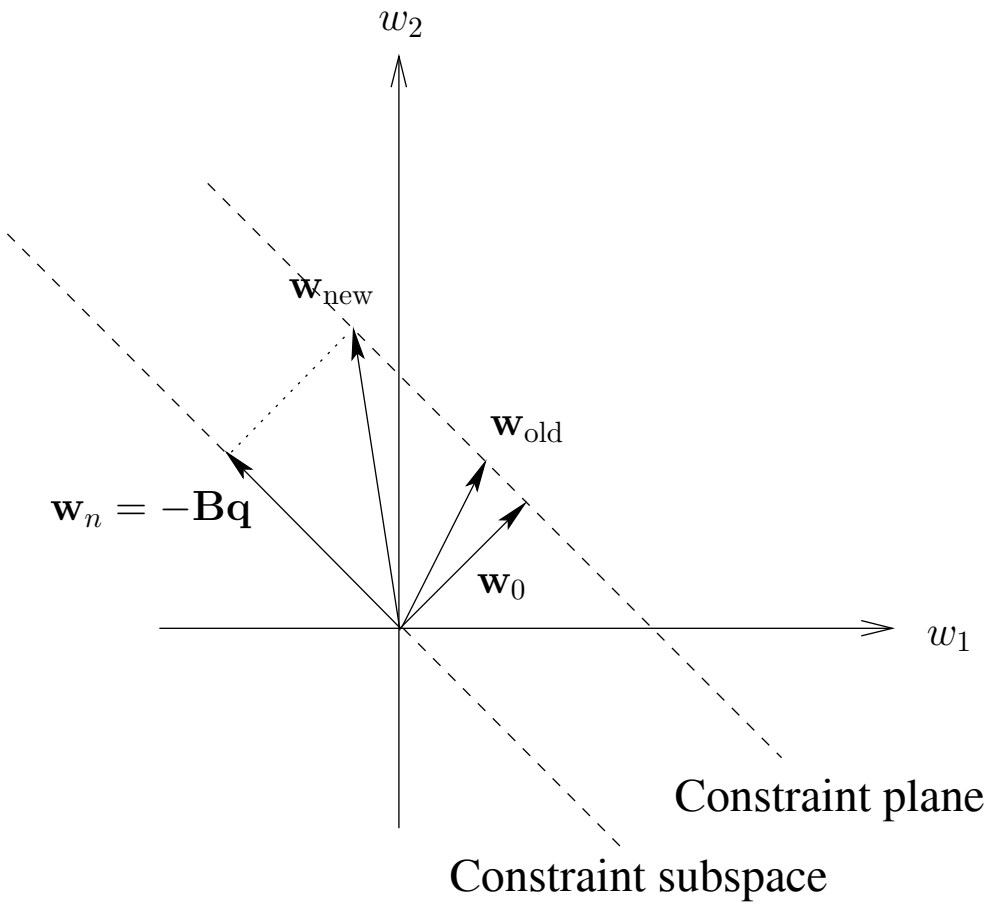
## Graphical Interpretation Frost, 1972



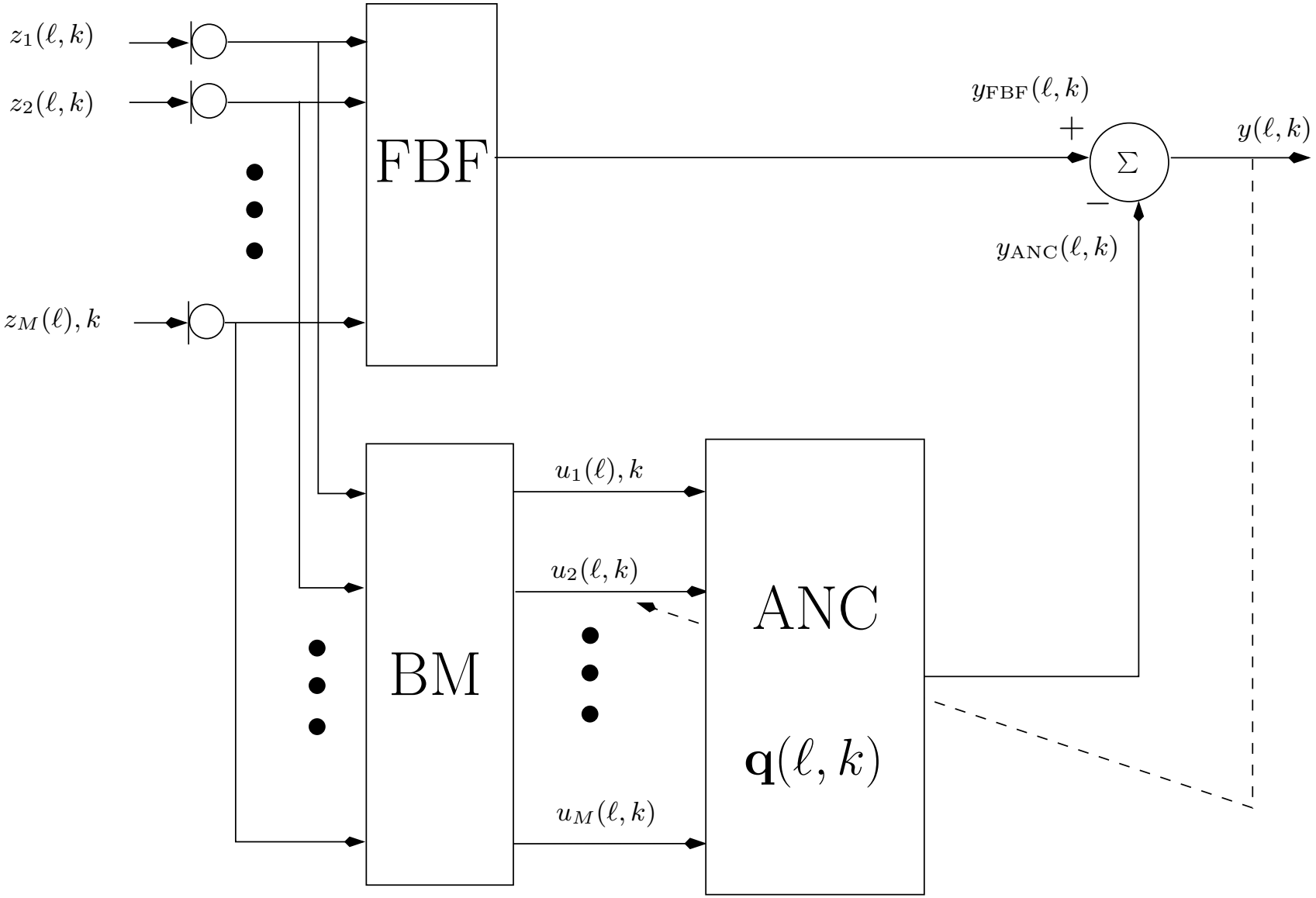
# The Generalized Sidelobe Canceller Implementation

## Split the Beamformer

- $\mathbf{w} = \mathbf{w}_0 - \mathbf{w}_n$
- $\mathbf{w}_n \triangleq \mathbf{B}^H \mathbf{q}$
- Constraints Subspace
  $\mathbf{w}_0 \in \mathrm{Span}\{C\}$
- Null Subspace
  (columns of $\mathbf{B}$ span $\mathcal{N}\{C\}$)



# The GSC Structure Griffiths, Jim, 1982



# The Minimum Variance Distortionless Beamformer

## Single Constraint

### Scenario:

- One desired signal.

### Beamformer Design:

- "Steers beam" towards the desired source (one constraint).
- Minimize all other directions.

$$\mathbf{C} = \mathbf{h}^d$$

$$\mathbf{g} = 1$$

# The MVDR Beamformer

## GSC Implementation Affes, Grenier, 1997; Hoshuyama et al., 1999; Gannot et al., 2000

### GSC Blocks:

- FBF - matched filter to the desired response, $\mathbf{h}^d$.
- BM - projection matrix to the null space of $\mathbf{h}^d$.
- ANC - recursively updated using the LMS algorithm (Shynk, 1992).

### Output signal:

$$y = s^d + \text{residual noise and interference signals}$$

# The Transfer Function GSC

## Relax Dereverberation Requirement Gannot et al., 2001

### Modified Constraint Set:

$$\mathbf{C} = \mathbf{h}^d$$
$$\tilde{\mathbf{g}} = (h_1^d)^*$$

### Equivalent to:

$$\mathbf{C} = \tilde{\mathbf{h}}^d \triangleq \frac{\mathbf{h}^d}{h_1^d} = \left[ 1 \ \frac{h_2^d}{h_1^d} \ \cdots \ \frac{h_M^d}{h_1^d} \right]^T$$

$$\mathbf{g} = 1.$$

# The Transfer Function GSC utilizing RTF

RTF suffices for implementing all blocks of the GSC.

## Output signal:

$$y = \underbrace{h_1^d s^d}_{s_1^d} + \text{residual noise and interference signals}$$

## Tradeoff:

Noise reduction is sacrificed if dereverberation required Habets et al., 2010.

# The Importance of the RTF

## Advantages

- Shorter than the ATF.
- Estimation methods are available:
  - Using ▸ speech non-stationarity (Gannot et al., 2001).
  - Using speech probability and spectral subtraction Cohen, 2004.
- RTF equivalent to Interaural Transfer Function (ITF).

## Drawbacks

- Non-causal

## CTF-GSC

For high $T_{60}$ multiplication in the frequency domain is only valid for very long frames. Hence, the RTFs will be extended to convolution in the STFT domain ▸ (CTF-GSC) Talmon et al., 2009.

# Multi Constraint Beamformer

Based on Multichannel Eigenspace Beamforming Markovich et al., 2009; 2010

## 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):

$$\mathbf{z} = \mathbf{H}^d \mathbf{s}^d + \mathbf{H}^i \mathbf{s}^i + \mathbf{H}^n \mathbf{s}^n + \mathbf{n}$$

# The Constraints Set

## Original

$$\mathbf{C} \triangleq \mathbf{H} = \left[ \mathbf{H}^d \ \mathbf{H}^i \ \mathbf{H}^n \right]$$

$$\mathbf{g} \triangleq \left[ \underbrace{1 \ \ldots \ 1}_{N_d} \ \underbrace{0 \ \ldots \ 0}_{N-N_d} \right]^T$$

## LCMV output

$$y = \sum_{j=1}^{N_d} s_j^d + \text{noise components}$$

# A Modified Constraints Set

## Noise & Interfernce

Replace the ATFs by an equivalent orthonormal basis $\mathbf{Q}$.

## Relax the dereverberation requirements using RTFs:

$$\tilde{\mathbf{g}} \triangleq \left[ \underbrace{(h_{11}^d)^* \ \dots \ (h_{N_d 1}^d)^*}_{N_d} \ \underbrace{0 \ \dots \ 0}_{N-N_d} \right]^T \Rightarrow \tilde{\mathbf{h}}_j^d \triangleq \mathbf{h}_j^d / h_{j1}^d$$

## The modified Constraints Set

$$\tilde{\mathbf{C}} \triangleq \left[ \tilde{\mathbf{H}}^d \ \mathbf{Q} \right]$$

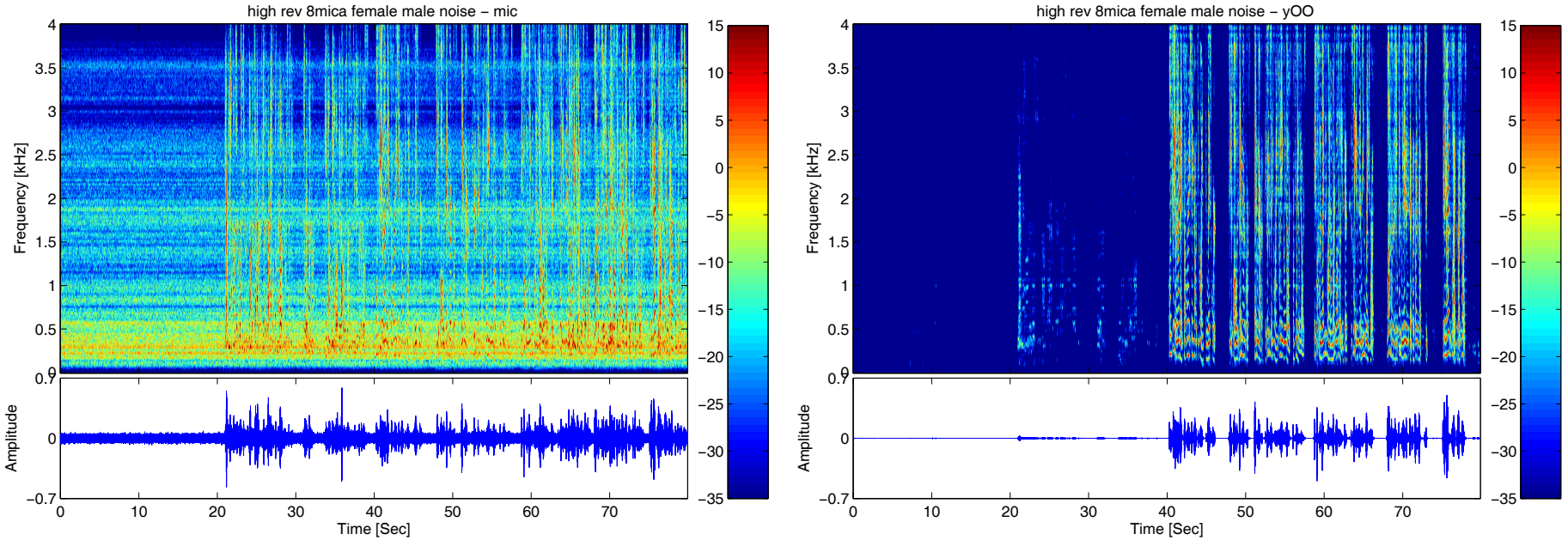
▸ (estimation by applying subspace methods)

## LCMV output

$$y = \sum_{j=1}^{N_d} h_{j1}^d s_j^d + \text{noise components}$$

# Single Desired Speaker

## Directional Noise Field

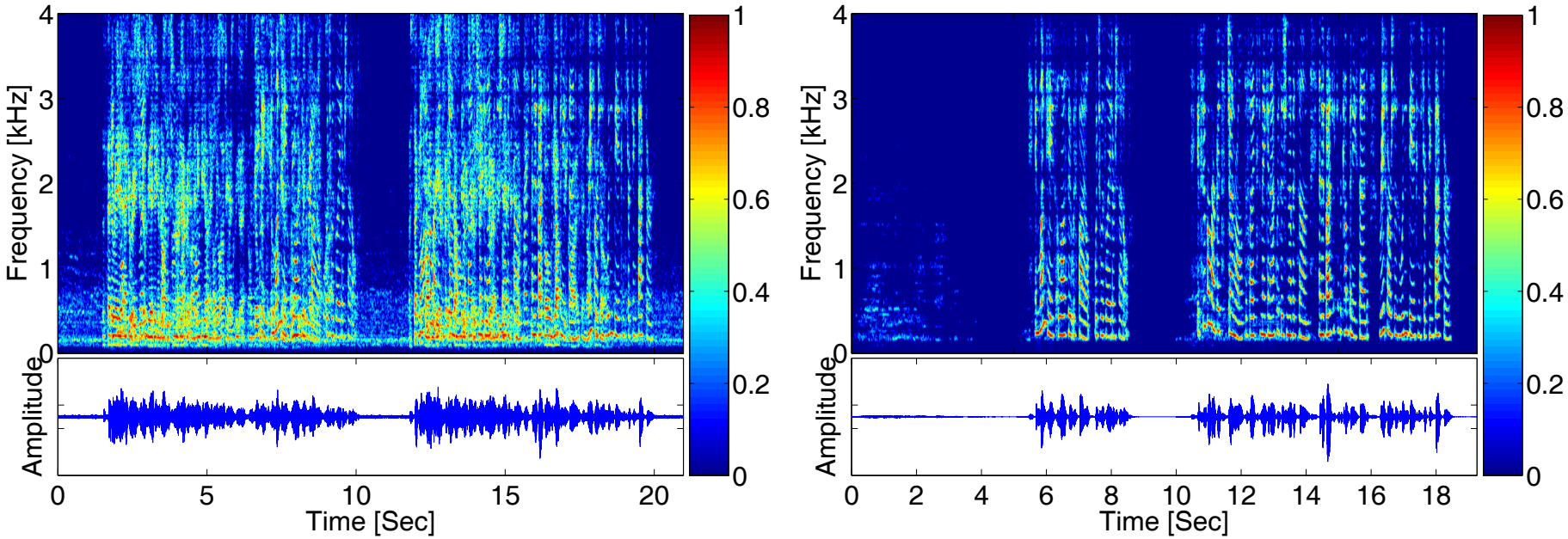



(a) Noisy at mic. #1

(b) Enhanced signal

Figure: Female (desired) and male (interference) with Directional noise. 8 microphones recorded at BIU acoustic lab set to $T_{60} = 300$ms.

# Multi-Speaker




(a) Noisy at mic. #1

(b) Enhanced signal

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

# 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

$$\mathbf{z} = \mathbf{H}^d \mathbf{s}^d + \mathbf{H}^i \mathbf{s}^i + \mathbf{v}$$

## Left & Right Reference Microphones

$$z_\ell = \mathbf{e}_\ell^H \mathbf{z}; \quad z_r = \mathbf{e}_r^H \mathbf{z}$$

where

$$\mathbf{e}_\ell = \begin{cases} 1 & m = m_\ell \\ 0 & \text{otherwise} \end{cases} \qquad \mathbf{e}_r = \begin{cases} 1 & m = m_r \\ 0 & \text{otherwise} \end{cases}$$

## Binaural Spatial Filters

$$y_\ell = \mathbf{w}_\ell^H \mathbf{z}; \quad y_r = \mathbf{w}_r^H \mathbf{z}.$$

# Double LCMV Criterion

## Two BFs Utilizing All Microphones

$$\mathbf{w}_\ell = \mathrm{LCMV}(\mathbf{z}; \mathbf{C}, \mathbf{g}_\ell); \quad \mathbf{w}_r = \mathrm{LCMV}(\mathbf{z}; \mathbf{C}, \mathbf{g}_r)$$

## Orthonormal Basis for the ATFs

$$\{\mathbf{H}_d = \mathbf{Q}_d \mathbf{\Theta}_d; \quad \mathbf{H}_i = \mathbf{Q}_i \mathbf{\Theta}_i\} \Rightarrow \mathbf{C} = \left[\mathbf{Q}_d \ \mathbf{Q}_i\right]$$

## Left & Right Response Vectors

Apply dereverberation relaxation utilizing RTFs.

## Cue Gain Factors:

Desired response $0 < \eta \approx 1$; Interference response $0 < \mu \ll 1$

# Interaural Signal Ratio (ISR)

## Input ISR

$$\mathrm{ISR}^{in} = \frac{z_\ell}{z_r} = \frac{\mathbf{e}_\ell^\dagger \left(\mathbf{H}_d \mathbf{s}_d + \mathbf{H}_i \mathbf{s}_i\right)}{\mathbf{e}_r^\dagger \left(\mathbf{H}_d \mathbf{s}_d + \mathbf{H}_i \mathbf{s}_i\right)}.$$

## Output ISR (in our implementation)

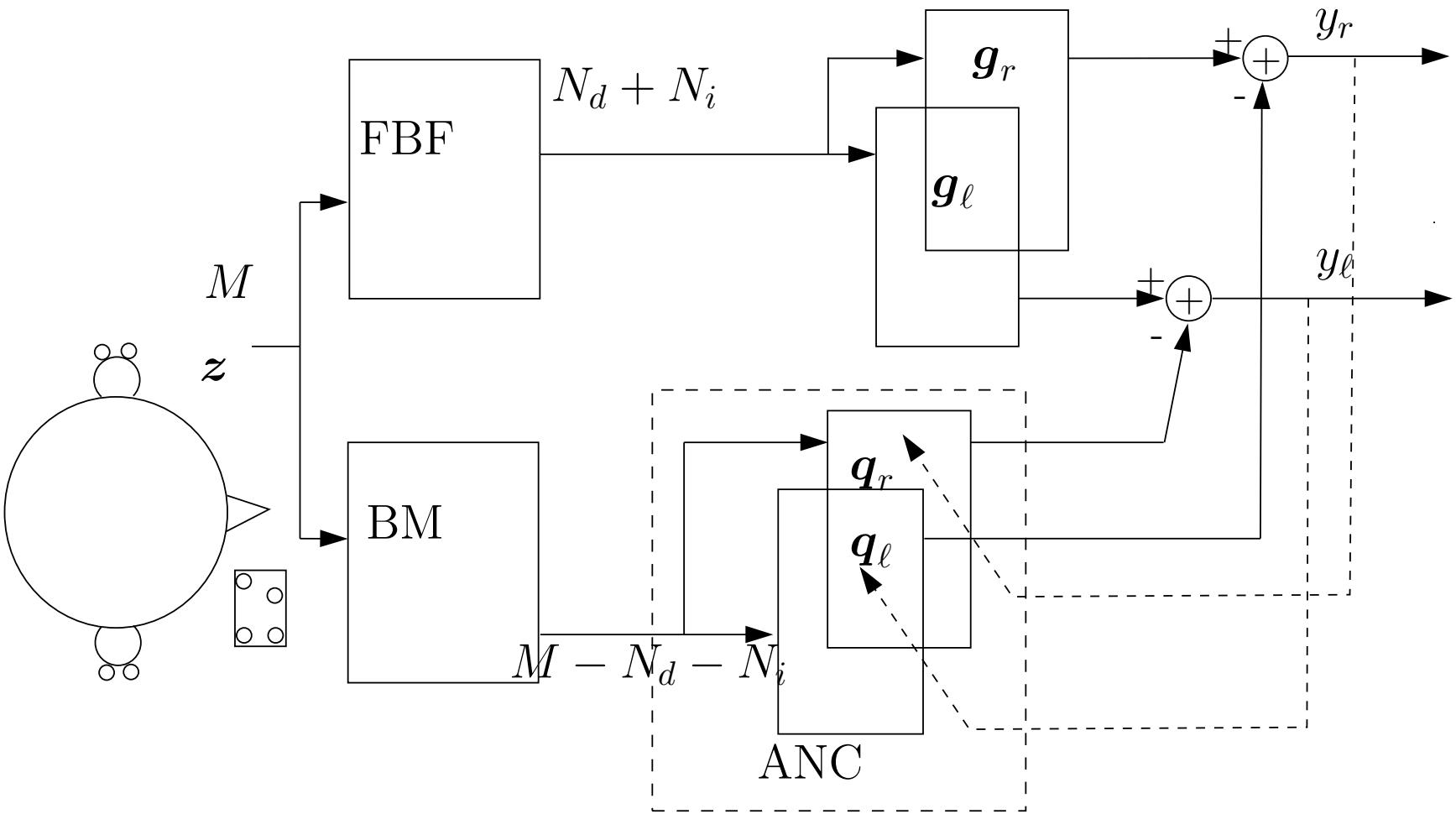
$$\mathrm{ISR}^{out} = \frac{y_\ell}{y_r} = \frac{\mathbf{e}_\ell^\dagger \left(\eta \mathbf{H}_d \mathbf{s}_d + \mu \mathbf{H}_i \mathbf{s}_i\right)}{\mathbf{e}_r^\dagger \left(\eta \mathbf{H}_d \mathbf{s}_d + \mu \mathbf{H}_i \mathbf{s}_i\right)}.$$

# ISR vs. ITF

## Properties

- Single source case: $\mathrm{ISR}^{out} = \mathrm{ISR}^{in}$ 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

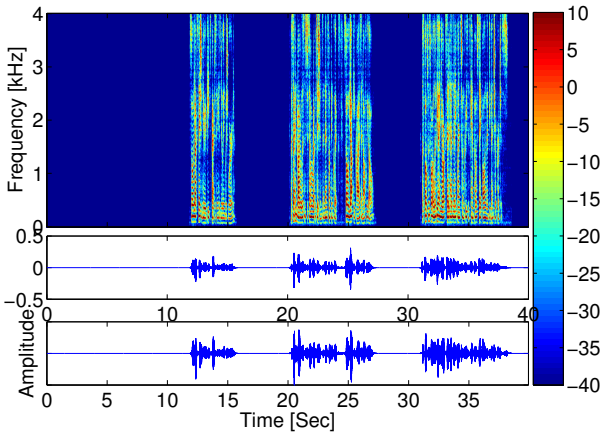


# Sonograms

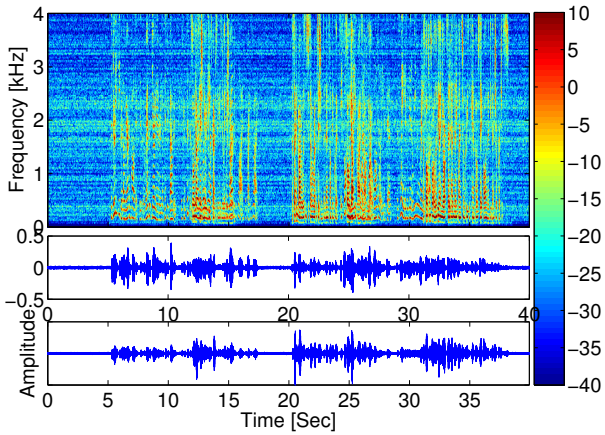
## Right Signal and Stereo Waveforms (Left and Right)

- 2 hearing aid devices, 2 microphones each, utility device with 4 mics.
- Cue gain factors: $\eta = 1$, $\mu = 0.1$ (20dB attenuation).

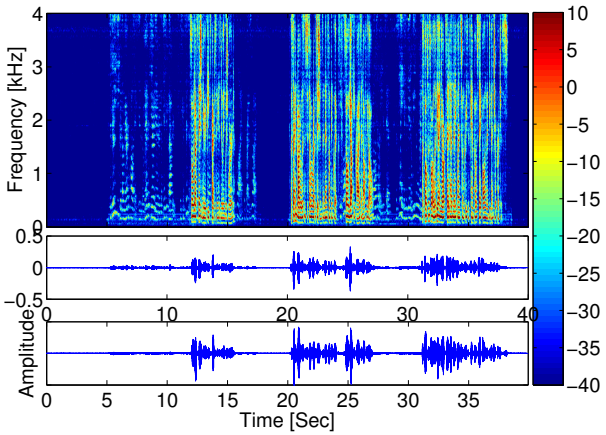
▸ Setup




(a) Desired speaker, reference mic.

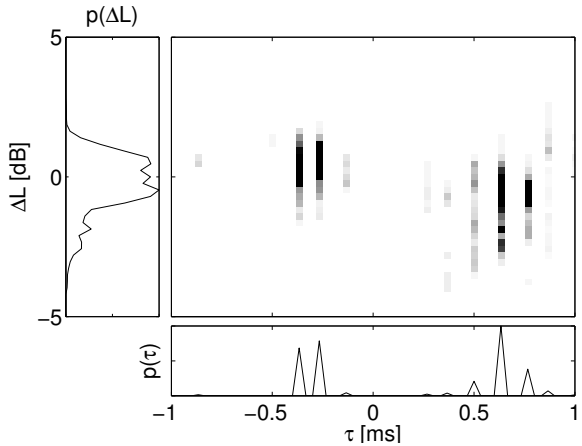



(b) Received reference microphones

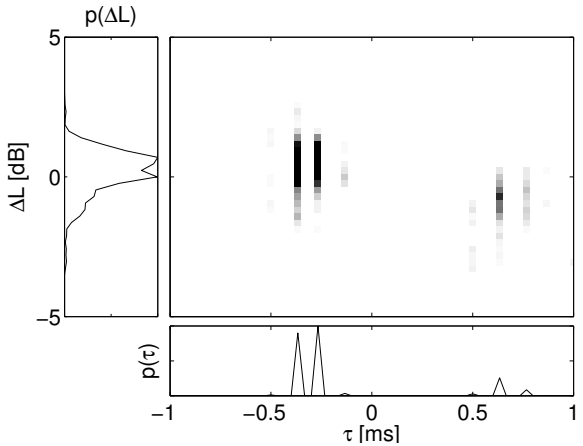



(c) BLCMV outputs

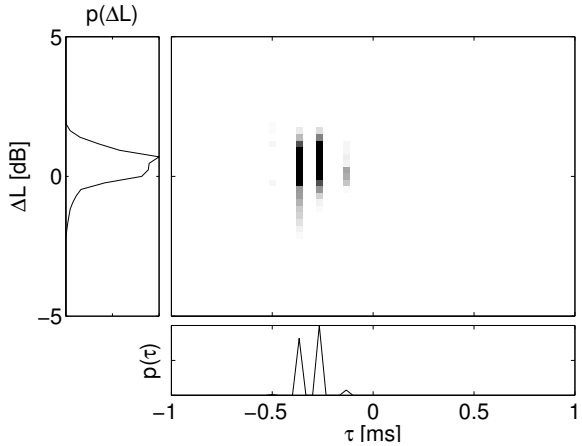
# ILD & ITD Preservation (Faller and Merimaa, 2004)



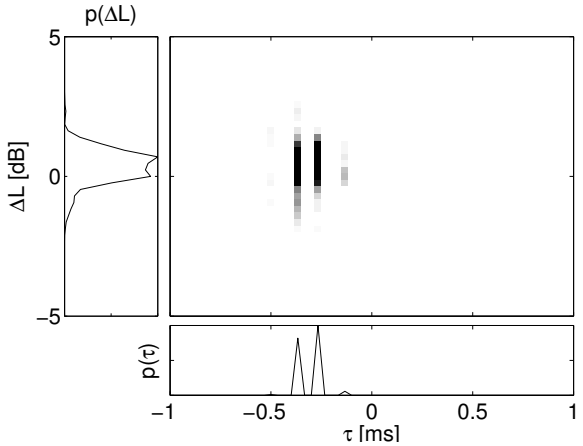

(a) Noisy input




(b) Enhanced output

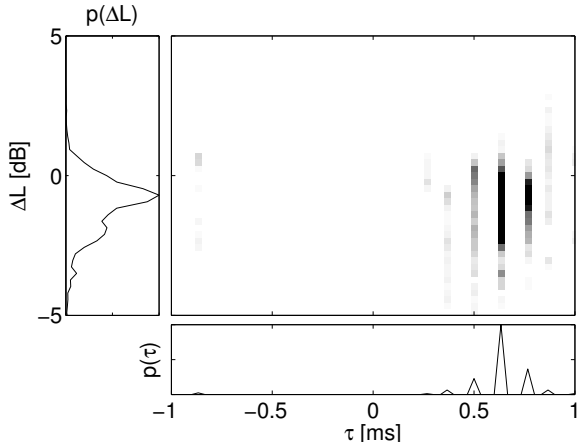



(c) Desired input

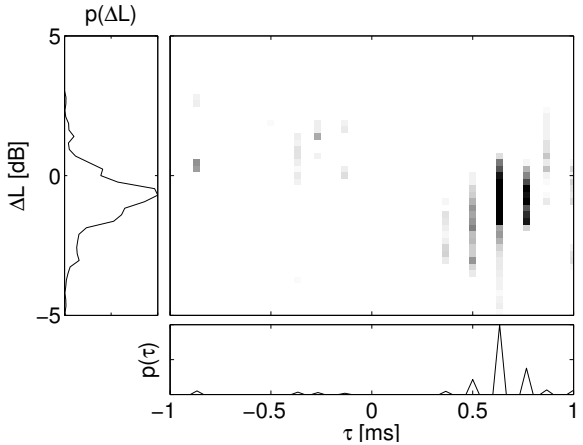



(d) Desired output

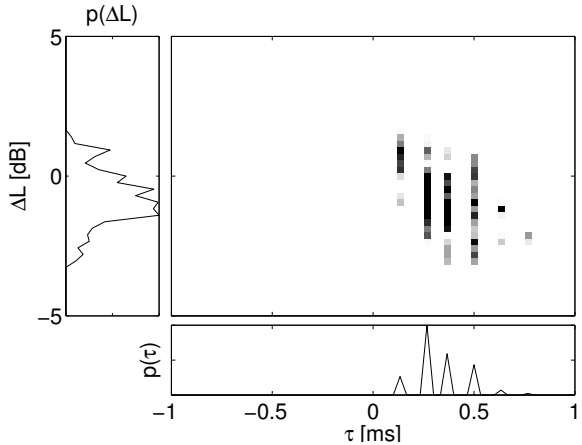
# ILD & ITD Preservation (Faller and Merimaa, 2004) (cont.)



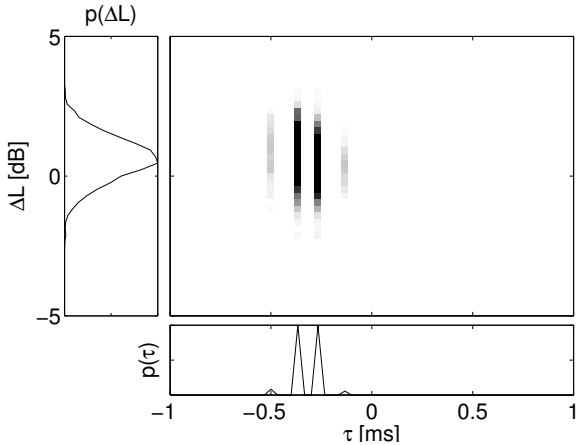

(e) Interference input




(f) Interference output




(g) Stationary input




(h) Stationary output

# Discussion

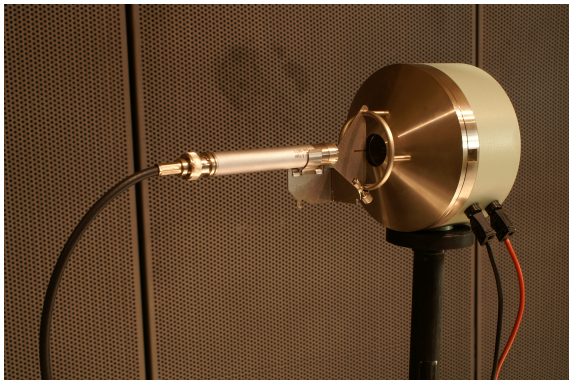
## Advantages of the Proposed family of Algorithms

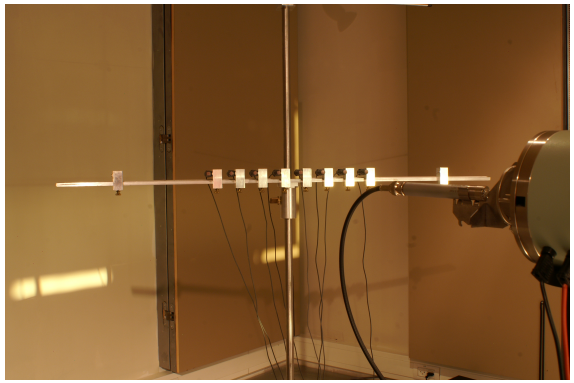
1. LCMV beamformers provide low distortion and high interference cancellation.
2. Practical estimation and tracking procedures.
3. Extendable to binaural algorithms for hearing aid applications.
4. Distributed versions exist (Bertrand, Moonen, 2011; Markovich-Golan, Gannot, Cohen, 2012).
5. The RTF allows for:
   1. Shorter processing frames.
   2. Higher noise reduction (traded-off for dereverberation).
   3. Preservation of binaural cues.

## Required improvements

1. Arbitrary activity patterns.
2. Improved tracking of moving speakers and sensors.
3. Robustness to head movements and to estimation errors.

# Vielen dank !





# Overview of Spatial Noise Reduction Techniques

1. Fixed beamforming Combine the microphone signals using a time-invariant filter-and-sum operation (data-independent).

   Jan, Flanagan, 1996; Doclo, Moonen, 2003; Schäfer, Heese, Wernerus, Vary, 2012

2. 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, Buckley, 1988; Van Trees, 2002

3. 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, 2003

4. Computational Auditory Scene Analysis (CASA) Aims at performing sound segregation by modelling the human auditory perceptual processing. Brown, Wang, 2005

Back

# Relative Transfer Function Estimation

## System Perspective:

$$z_m = \tilde{h}_m^d z_1 + u_m$$

## System Identification:

$$\hat{\Phi}_{z_m z_1} = \tilde{h}_m^d \hat{\Phi}_{z_1 z_1} + \Phi_{u_m z_1} + \varepsilon_m$$

## Estimation is Biased:

$u_m$ and $z_1$ are correlated $\Rightarrow$ Biased estimator for $\tilde{h}_m^d$

# Relative Transfer Function Estimation

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

## Assumptions:

- System is Time-Invariant
- Noise is stationary (true for $\mathbf{\Phi}_{vv}(k)$)
- Speech is non-stationary (use frames $\ell_i$, $i = 1, \ldots, I$)

$$\begin{bmatrix} \hat{\Phi}_{z_m z_1}(\ell_1, k) \\ \hat{\Phi}_{z_m z_1}(\ell_2, k) \\ \vdots \\ \hat{\Phi}_{z_m z_1}(\ell_I, k) \end{bmatrix} = \begin{bmatrix} \hat{\Phi}_{z_1 z_1}(\ell_1, k) & 1 \\ \hat{\Phi}_{z_1 z_1}(\ell_2, k) & 1 \\ \vdots & \\ \hat{\Phi}_{z_1 z_1}(\ell_I, k) & 1 \end{bmatrix} \begin{bmatrix} \tilde{h}_m^d(k) \\ \Phi_{u_m z_1}(k) \end{bmatrix} + \begin{bmatrix} \varepsilon_m(\ell_1, k) \\ \varepsilon_m(\ell_2, k) \\ \vdots \\ \varepsilon_m(\ell_I, k) \end{bmatrix}$$

Back

# The Convolutive TF-GSC Talmon et al., 2009

## Motivation

### The TF-GSC (Gannot *et al.*, 2001)

- The RTFs are incorporated into the MVDR beamformer, implemented in GSC structure.
- The adaptation to reverberant environments, is obtained by time-frequency domain implementation.
- The TF-GSC achieves improved enhancement of a noisy speech signal (compared with the time-domain, delay-only, GSC).

### High $T_{60}$

- The RIRs become very long.
- The 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 deteriorating performance.

# Objectives

## 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., 2009).
- Significant improvements in blocking ability , NR, distortion and SNR in comparison with the TF-GSC.

Back

# Experimental Study

## Setup

### Comparing the CTF-GSC and the TF-GSC

- Image method (Allan and Berkley, 1979, implemented by Habets, 2009).
- Array of 5 microphones.
- Reverberation time $T_{60} = 0.5$s.
- TF-GSC:
  - Frame length - $N = 512$.
  - RTF length - 250.
  - Noise Canceller length - 450.
- CTF-GSC:
  - In FBF and BM - $N = 512, 50\%$ overlap.
  - In adaptive NC - $N = 512, 75\%$ overlap.

# Signal Blocking

The signal blocking factor (SBF) is defined by:

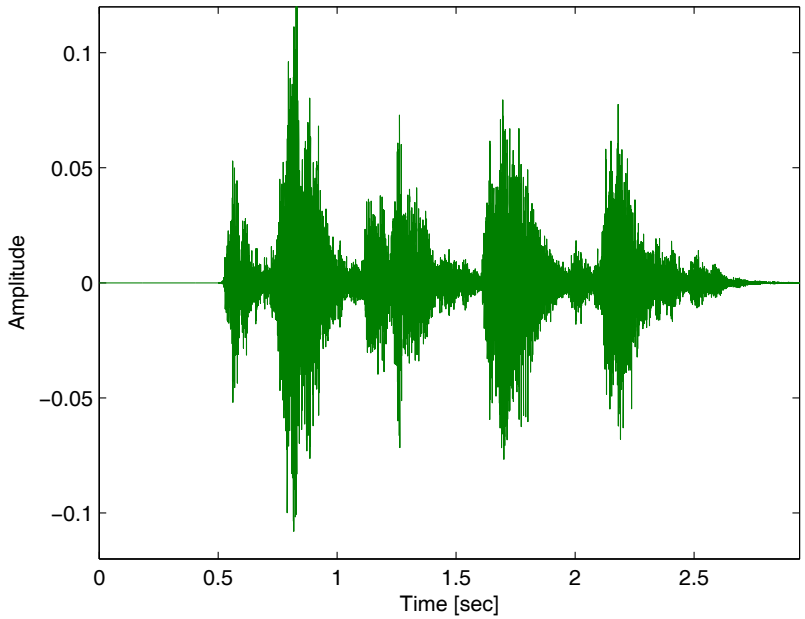
$$\textbf{SBF} = 10 \log_{10} \frac{E\left\{(s^d * h_1^d[n])^2\right\}}{\mathrm{Mean}_m E\left\{u_m^2[n]\right\}}$$

where $u_m[n]$; $m = 2, \ldots, M$ are the blocking matrix outputs.

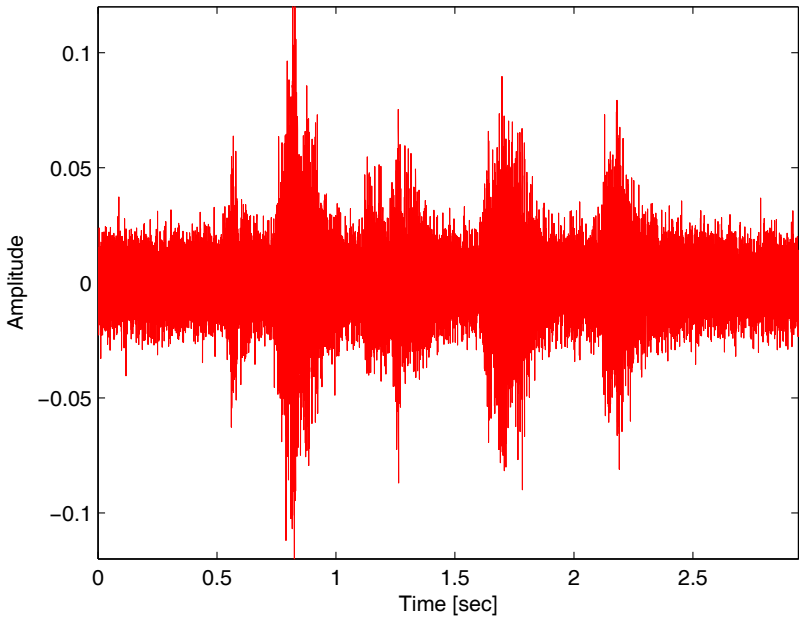
## The blocking ability [dB] (known RTF)

| | mic. 2 | mic. 3 | mic. 4 | mic. 5 |
|---|---|---|---|---|
| TF | 17 | 12 | 14 | 8 |
| CTF | 22 | 17 | 18 | 13 |

# Known RTF, Input SNR=0dB

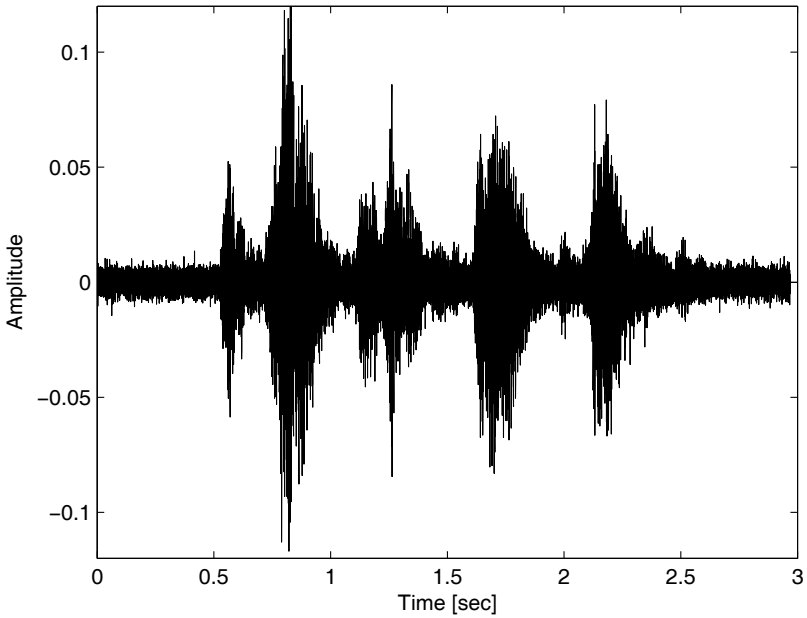



(i) Reverberated speech at microphone #1.

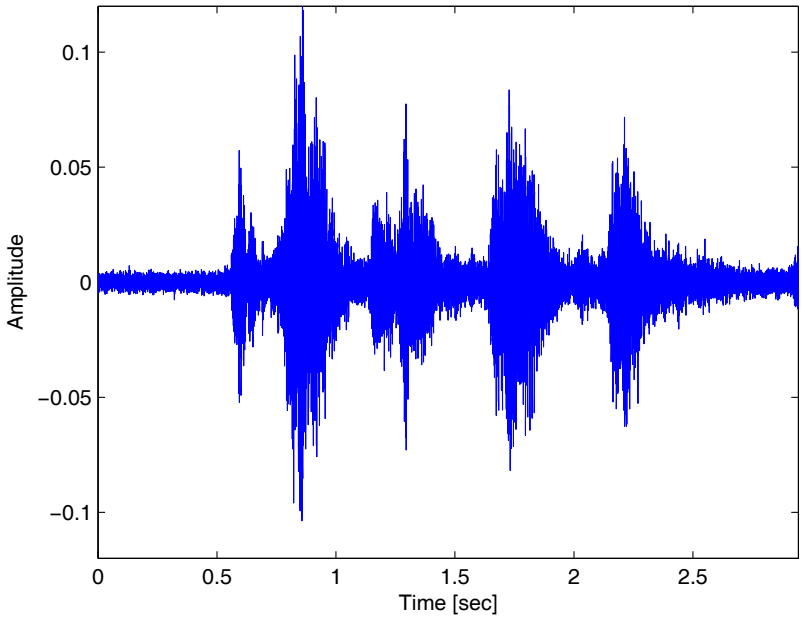



(j) Noisy signal at microphone #1.

# Known RTF, Input SNR=0dB (cont.)




(k) TF-GSC output.



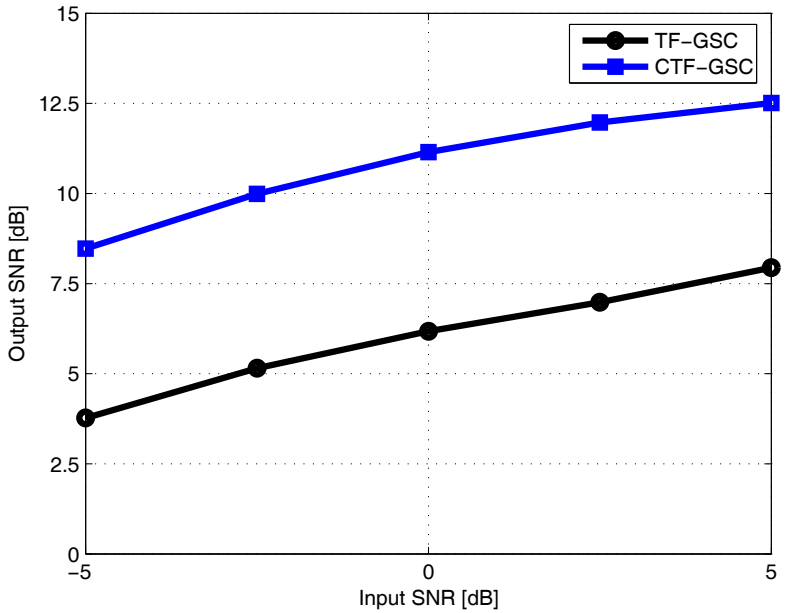

(l) CTF-GSC output.

# Summary

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

| In SNR | SNR | | NR | |
|---|---|---|---|---|
| | TF-GSC | CTF-GSC | TF-GSC | CTF-GSC |
| -5 | 3.8 | 8.5 | -5.9 | -10.9 |
| -2.5 | 5.2 | 10.0 | -6.2 | -10.9 |
| 0 | 6.2 | 11.2 | -6.2 | -10.9 |
| 2.5 | 7.0 | 12.0 | -6.7 | -10.9 |
| 5 | 7.9 | 12.5 | -6.1 | -10.9 |

Back



# Estimation

## Unknowns

- Desired sources RTFs, $\tilde{\mathbf{H}}^d$.
- Interferences subspace basis, $\mathbf{Q}$.

## Limiting Assumptions

- The RIRs are time invariant.
- A segment without interference speakers is available for each desired source (double talk within desired allowed).
- A segment without desired speakers is available for each interfering source (double talk within interfering allowed).

# Interferences Subspace Estimation

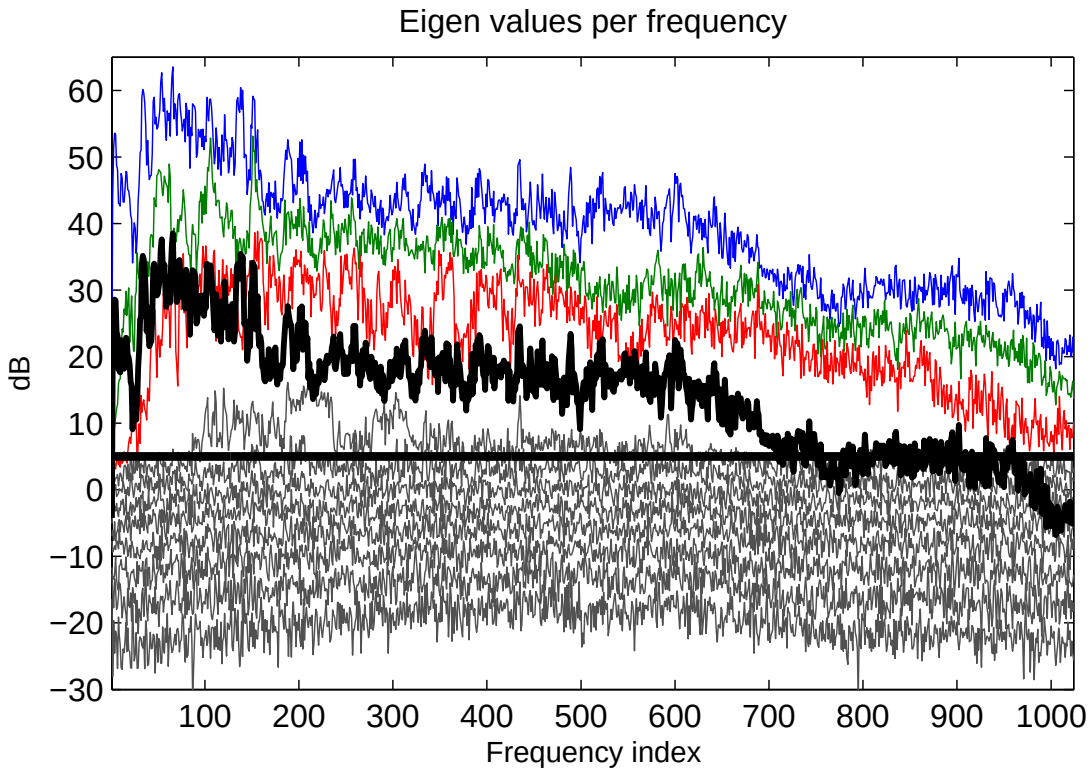
Step 1

## EVD and Pruning

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

$$\hat{\mathbf{\Phi}}^{i}_{zz} = \bar{\mathbf{Q}}_i \mathbf{\Lambda}_i \bar{\mathbf{Q}}_i^H$$

- All eigenvectors corresponding to "weak" eigenvalues are discarded



# Interferences Subspace Estimation

Step 2

## Union of Estimates using Gram-Schmidt

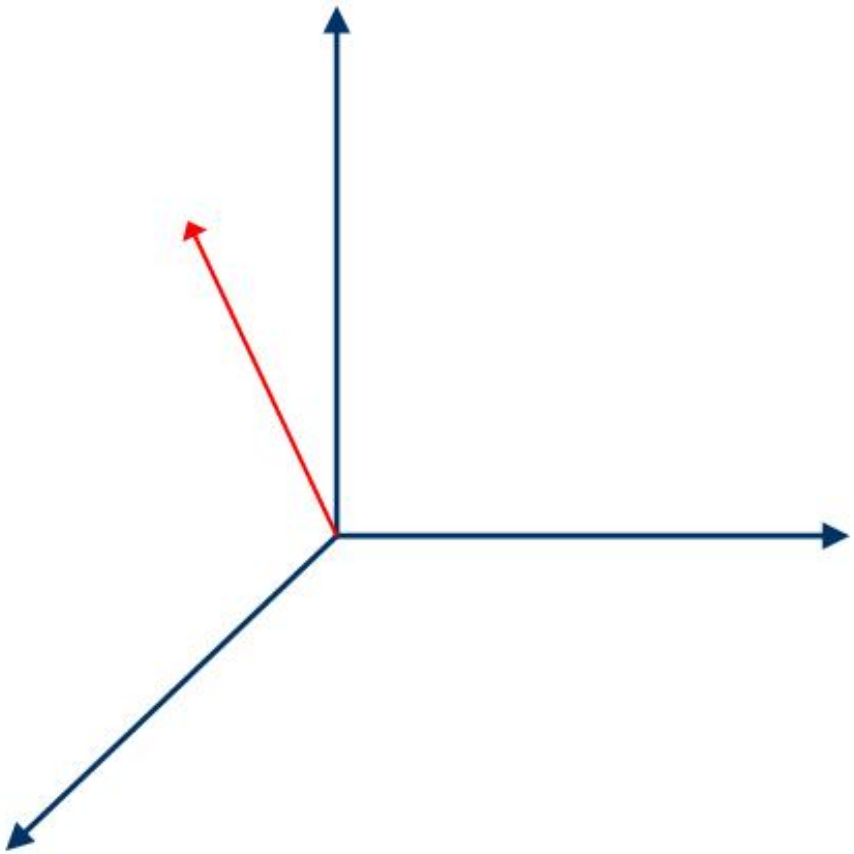
- Apply QRD

$$\left[ \bar{\mathbf{Q}}_1 \bar{\mathbf{\Lambda}}_1^{\frac{1}{2}} \ \ldots \ \bar{\mathbf{Q}}_{N_{seg}} \bar{\mathbf{\Lambda}}_{N_{seg}}^{\frac{1}{2}} \right] \mathbf{P} = \mathbf{Q}\mathbf{R}$$

- Discard vectors from the basis $\mathbf{Q}$ that correspond to "weak" coefficients in $\mathbf{R}$

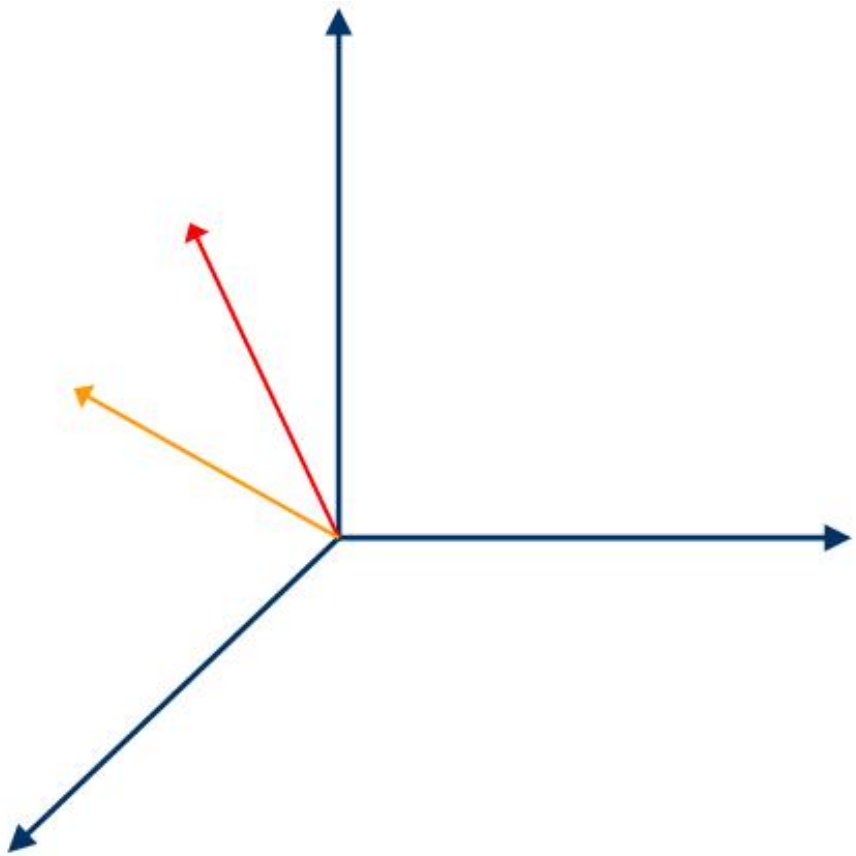
# EVD per Frame - Graphical Interpretation

## Frame 1, strong eigenvectors



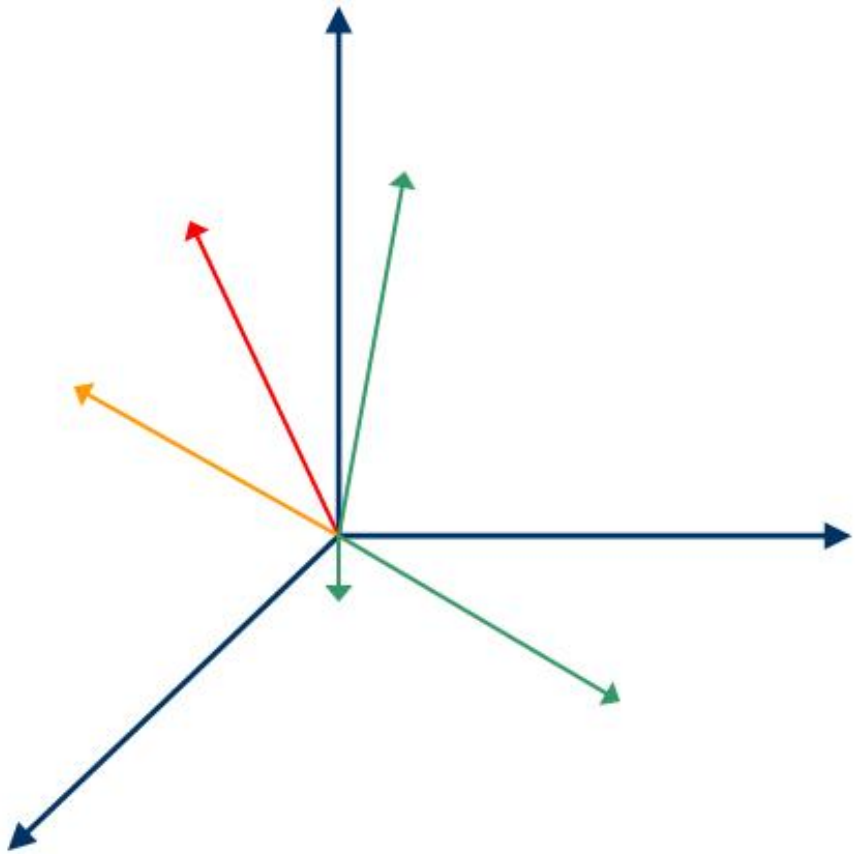
# EVD per Frame - Graphical Interpretation

## Frame 2, strong eigenvectors



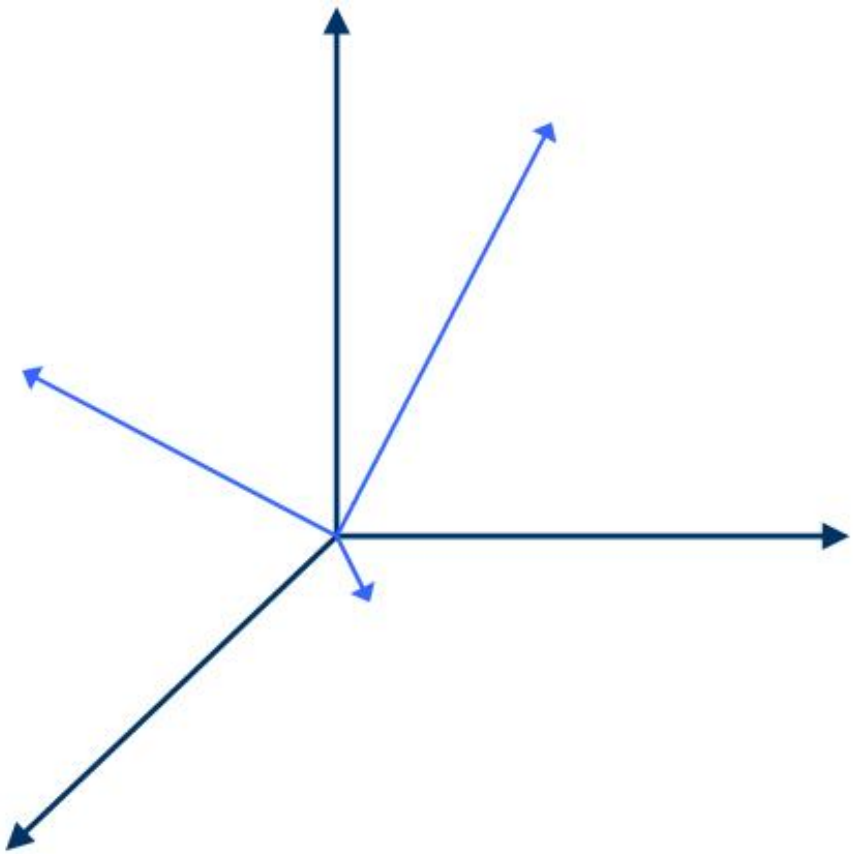
# EVD per Frame - Graphical Interpretation

## Frame 3, strong eigenvectors



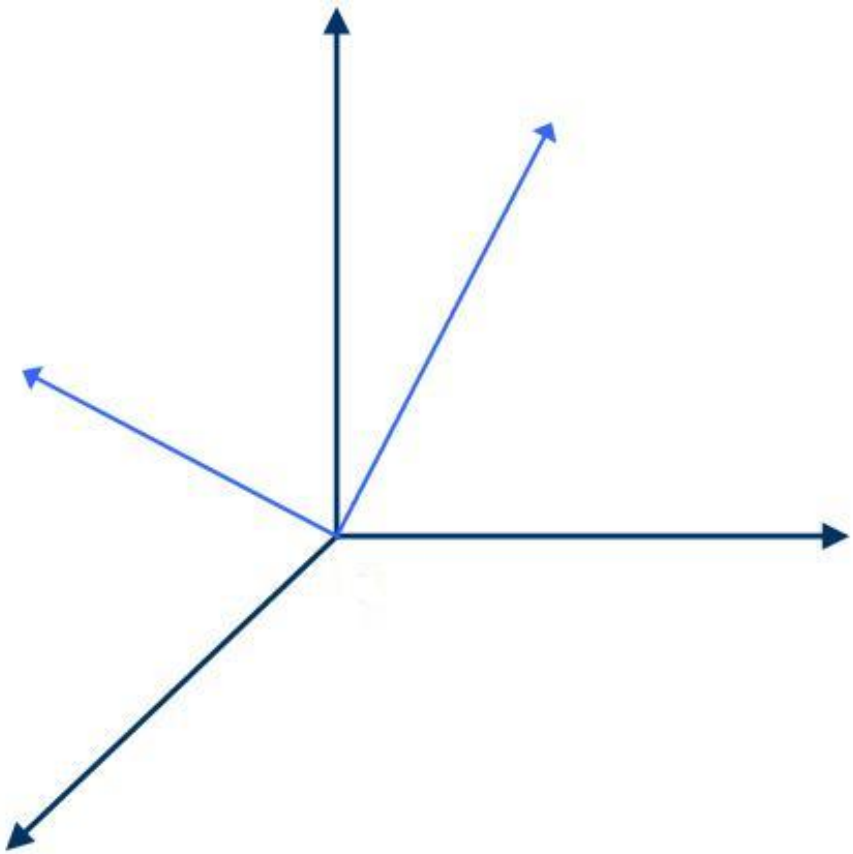
# QRD Calculation

## Graphical Interpretation



# QRD Pruning

## Graphical Interpretation



# Desired Sources RTF Estimation

## For a Single Desired Source

### PSD Estimation

- Estimate spatial PSD of stationary sources (during noise-only segments).
- Estimate spatial PSD of one desired source + stationary sources.

### Largest Generalized Eigenvector

- $\mathbf{f} = \mathrm{GEVD}(\text{“Desired+Noise PSD”}, \text{“Noise PSD”})$.
- Rotate and normalize $\mathbf{f}$ to obtain the RTF $\tilde{\hat{\mathbf{h}}}_j^d$.

Back

# 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, $\theta_d = 30^o$, 1m (constrained).
  - 1 interference speaker at $\theta_i = -70^o$, 1m (constrained).
  - 1 directional stationary noise, $\theta_n = -40^o$, 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). Back