Blind multi-microphone noise reduction and dereverberation algorithms for speech communication applications

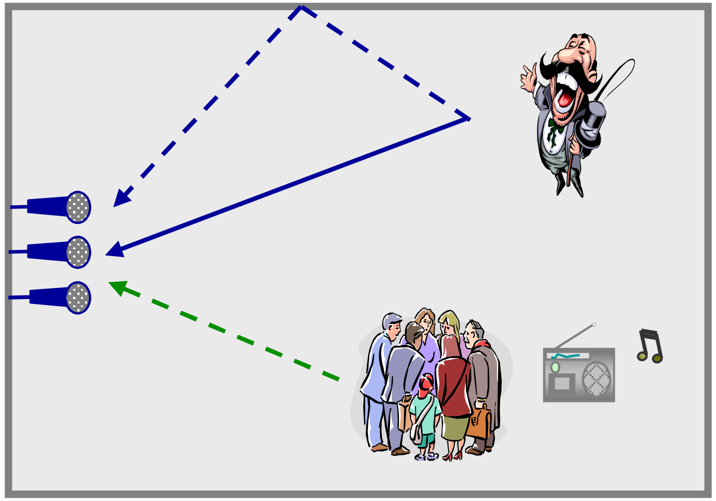
Prof. Dr. Simon Doclo
University of Oldenburg, Germany
Dept. of Medical Physics and Acoustics, Cluster of Excellence Hearing4all

http://www.sigproc.uni-oldenburg.de/

SANE Workshop, 24.10.2019

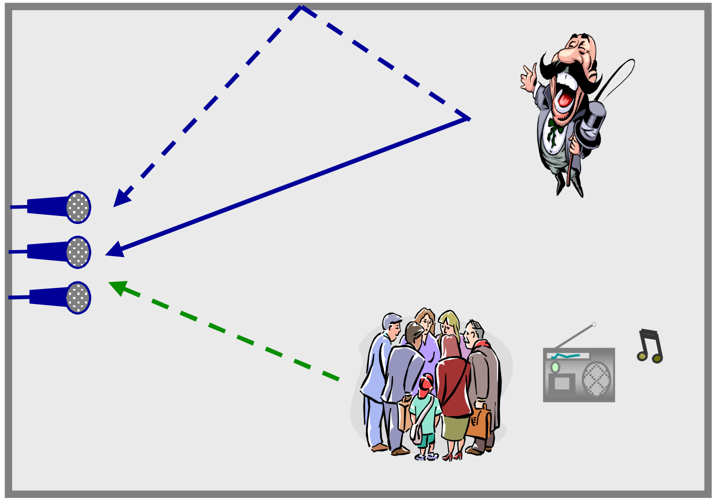
Introduction

- **Problem**
  - **Ambient noise and reverberation** jointly present in typical acoustic environments
  - **Speech quality and intelligibility** degradation for speech communication applications
  - Performance degradation of voice-controlled systems



Introduction

- **Objectives**
  - Single- and **multi-microphone joint noise reduction and dereverberation** algorithms
  - Speech communication applications: **blind and on-line processing** for time-varying dynamic acoustic scenarios
  - Exploit knowledge or (statistical) models of speech signals and room acoustics



- **This presentation**
  1. **Joint estimation of (time-varying) spatial and spectral variables** for multi-microphone speech enhancement
  2. **Binaural hearing devices:** combination of speech enhancement and preservation of auditory scene
  3. **Extension to acoustic sensor networks** with spatially distributed microphones

1. Joint dereverberation and noise reduction

Signal model

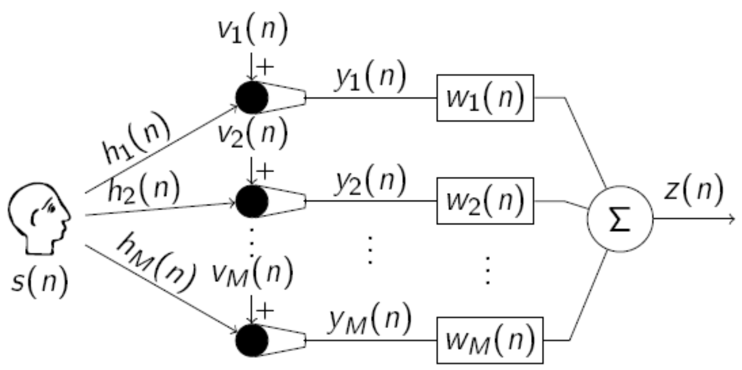
- **Scenario:** speech source in noisy and reverberant environment, *M* microphones
- **Model in Short-Time Fourier Transform (STFT) domain:**

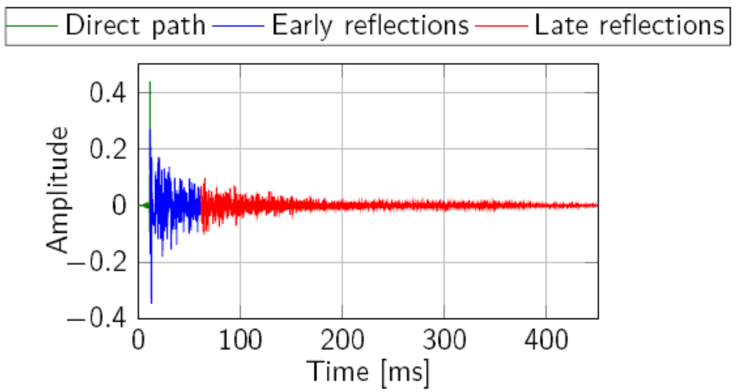
$$\begin{aligned} y_m(k,l) &= h_m(k,l) \star s(k,l) + v_m(k,l) \\ &= a_m(k,l)x_1(k,l) + x_{r,m}(k,l) + v_m(k,l) \end{aligned}$$

direct and early reverberation **late reverberation** **ambient noise**

$$\boxed{\mathbf{y}(k,l) = \mathbf{a}(k,l)x_1(k,l) + \mathbf{x}_r(k,l) + \mathbf{v}(k,l)}$$

a(*k*,*l*) = vector of **relative early transfer functions (RETFs)** of target source

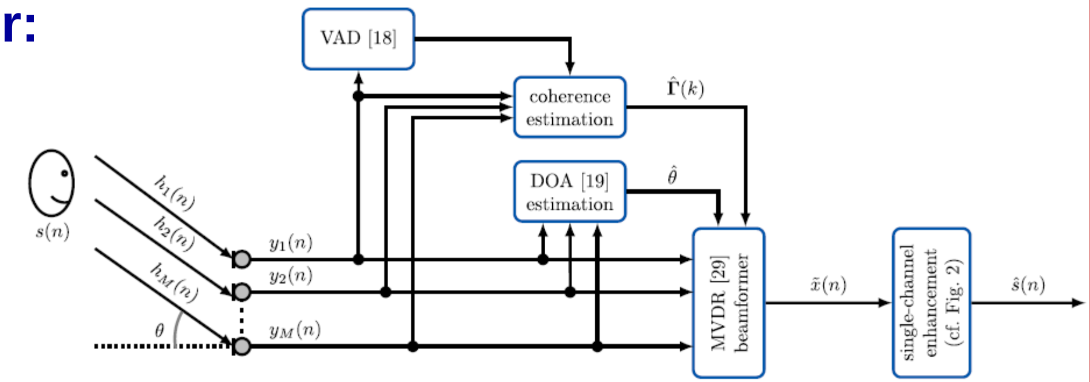




Multi-microphone dereverberation and noise reduction

1. **Beamforming + spectral postfilter:** *multiply* each time-frequency bin with *real-valued* gain

$$\mathbf{y}(l) = \mathbf{a}(l)x_1(l) + \mathbf{x}_r(l) + \mathbf{v}(l)$$

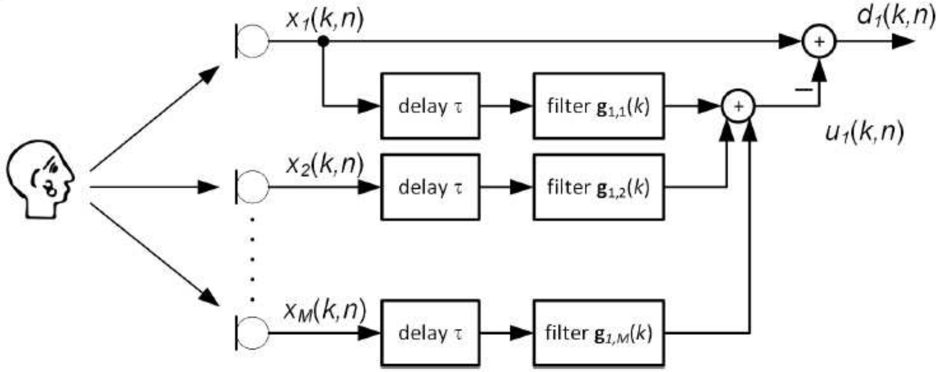


$$\mathbf{w}_{MWF} = \frac{(\phi_d\mathbf{\Gamma}+\mathbf{\Phi_v})^{-1}\mathbf{a}}{\mathbf{a}^H(\phi_d\mathbf{\Gamma}+\mathbf{\Phi_v})^{-1}\mathbf{a}} \cdot \frac{\phi_{x_1}}{\phi_{x_1}+(\mathbf{a}^H(\phi_d\mathbf{\Gamma}+\mathbf{\Phi_v})^{-1}\mathbf{a})^{-1}}$$

2. **Reverberation and noise suppression:** *subtract complex-valued* estimate of late reverberant and noise component

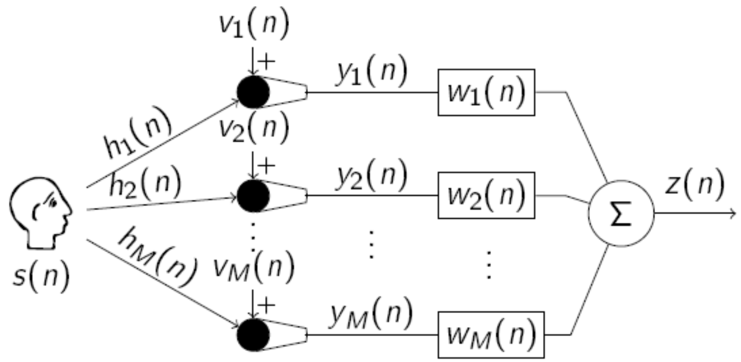
$$y_m(l) = h_m(l) \star s(l) + v_m(l)$$

$$\hat{x}_{e,1}(l) = y_1(l) - \mathbf{Y}_\tau(l)\mathbf{g}(l)$$



Beamforming + spectral postfilter

- **Filter-and-sum structure :** $\boxed{z = \mathbf{w}^H \mathbf{y}}$



Beamforming + spectral postfilter

- **Filter-and-sum structure :** $z = \mathbf{w}^H \mathbf{y}$ $\mathbf{y} = \mathbf{a} x_1 + \mathbf{n}$

- **“Workhorse algorithm”: parametric Multi-channel Wiener filter (MWF)**

  Goal: estimate desired speech component in reference microphone + trade off interference (*noise and/or reverberation*) reduction and speech distortion

$$\min_{\mathbf{w}} \mathcal{E}\{|\mathbf{w}^H \mathbf{x} - x_1|^2\} + \mu \mathcal{E}\{|\mathbf{w}^H \mathbf{n}|^2\} \quad \Rightarrow \quad \mathbf{w}_{MWF} = (\mathbf{\Phi}_x + \mu \mathbf{\Phi}_n)^{-1} \mathbf{\Phi}_x \mathbf{e}$$

  Requires estimate of covariance matrices, e.g., based on speech presence probability (SPP)

  Can be decomposed as **MVDR beamformer and spectral postfilter**

$$\mathbf{w}_{MWF} = \frac{\mathbf{\Phi}_n^{-1} \mathbf{a}}{\mathbf{a}^H \mathbf{\Phi}_n^{-1} \mathbf{a}} \cdot \frac{\phi_{x_1}}{\phi_{x_1} + \mu (\mathbf{a}^H \mathbf{\Phi}_n^{-1} \mathbf{a})^{-1}}$$

  Requires estimate/model of interference **covariance matrix** $\mathbf{\Phi}_n$, estimate/model of **relative (early) transfer function vector a** of desired source, and **PSDs** of speech and interference components at MVDR output

[Doclo, Kellermann, Makino, Nordholm, *IEEE Signal Processing Magazine*, Mar. 2015]
[Gannot, Vincent, Markovich-Golan, Ozerov, *IEEE/ACM Trans. Audio, Speech and Language Processing*, Apr. 2017]

Beamforming + spectral postfilter

- **Signal model**

$$\mathbf{y}(l) = \mathbf{a}(l)x_1(l) + \mathbf{x}_r(l) + \mathbf{v}(l)$$

$$\mathbf{\Phi}_y(l) = \phi_{x_1}(l)\mathbf{a}(l)\mathbf{a}^H(l) + \mathbf{\Phi}_{x_r}(l) + \mathbf{\Phi}_v(l)$$

Late reverberation: model as diffuse sound field $\mathbf{\Phi}_{x_r}(l) = \phi_d(l)\mathbf{\Gamma}$

with $\phi_d(l)$ *time-varying* diffuse PSD and $\mathbf{\Gamma}$ *time-invariant* coherence matrix (also incorporating diffuse noise !)

$$\mathbf{w}_{MWF} = \frac{(\phi_d\mathbf{\Gamma}+\mathbf{\Phi_v})^{-1}\mathbf{a}}{\mathbf{a}^H(\phi_d\mathbf{\Gamma}+\mathbf{\Phi_v})^{-1}\mathbf{a}} \cdot \frac{\phi_{x_1}}{\phi_{x_1}+(\mathbf{a}^H(\phi_d\mathbf{\Gamma}+\mathbf{\Phi_v})^{-1}\mathbf{a})^{-1}}$$

- **Key estimation tasks:**
  - **RETF vector a(*l*)**: *anechoic* (based on DOA estimate) or *reverberant*
  - **Diffuse/late reverberant PSD** $\phi_d(l)$: using single-channel *temporal model* (exponential decay) or based on multi-channel *diffuse sound field model*
  - **Noise covariance matrix** $\mathbf{\Phi}_v(l)$**:** *estimate* (based on SPP) or *model* (e.g., spatially white noise)

Estimation of PSDs

- **Requiring estimate of RETF vector and noise covariance matrix**

$$\hat{\mathbf{\Phi}}_x(l) = \hat{\mathbf{\Phi}}_y(l) - \hat{\mathbf{\Phi}}_v(l) = \phi_{x_1}(l)\mathbf{a}(l)\mathbf{a}^H(l) + \phi_d(l)\mathbf{\Gamma}$$

  – *Maximum-likelihood estimators,* requiring iterative optimisation procedure
  – Closed-form *least-squares estimators*, based on Frobenius norm

$$\min_{\phi_{x_1}(l),\phi_d(l)} ||\hat{\mathbf{\Phi}}_x(l) - \phi_{x_1}(l)\mathbf{a}(l)\mathbf{a}^H(l) - \phi_d(l)\mathbf{\Gamma}||_F^2$$

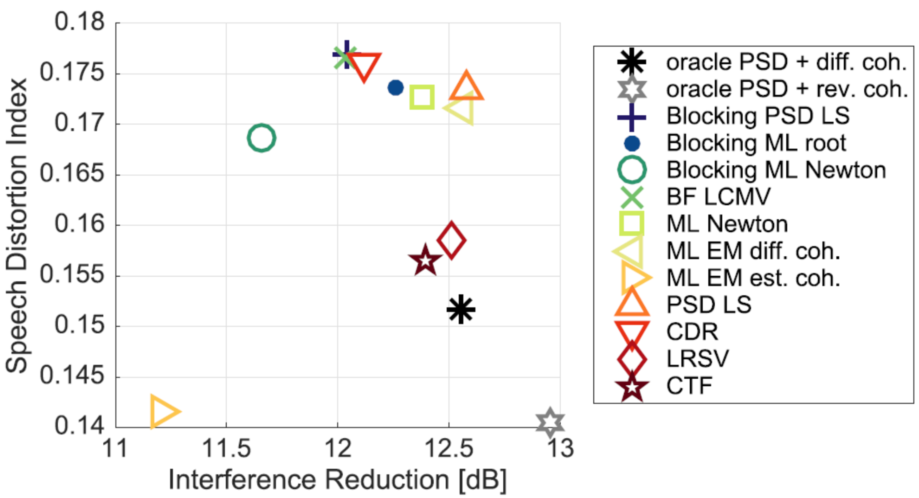


Fig. 9. Speech distortion vs. interference reduction for RSNR = 15 dB.

Similar performance for most methods…

[Braun, Kuklasinski, Schwartz, Thiergart, Habets, Gannot, Doclo, Jensen, *IEEE/ACM Trans. Audio, Speech and Language Processing*, June 2018.]

Joint Estimation of RETF vector and PSDs

1. Covariance whitening (CW) method:

- **Requires estimate of noise covariance matrix**

$$\hat{\mathbf{\Phi}}_x(l) = \hat{\mathbf{\Phi}}_y(l) - \hat{\mathbf{\Phi}}_v(l) = \phi_{x_1}(l)\mathbf{a}(l)\mathbf{a}^H(l) + \phi_d(l)\mathbf{\Gamma}$$

- **Eigenvalue decomposition** of prewhitened signal correlation matrix

$$\hat{\mathbf{\Phi}}_x^w(l) = \mathbf{\Gamma}^{-1/2}\hat{\mathbf{\Phi}}_x(l)\mathbf{\Gamma}^{-H/2} = \phi_{x_1}(l)\mathbf{b}(l)\mathbf{b}^H(l) + \phi_d(l)\mathbf{I}$$

  - **Principal eigenvector $\mathbf{u}(l)$: estimate of RETF vector**

$$\hat{\mathbf{a}}(l) = \frac{\mathbf{\Gamma}^{1/2}\mathbf{u}(l)}{\mathbf{e}^T\mathbf{\Gamma}^{1/2}\mathbf{u}(l)}$$

  - **Eigenvalues: estimate of PSDs**

$$\hat{\phi}_d(l) = \lambda_2\{\hat{\mathbf{\Phi}}_x^w(l)\} \qquad \hat{\phi}_{d,\mu}(l) = \frac{1}{M-1}\left(\mathrm{tr}\{\hat{\mathbf{\Phi}}_x^w(l)\} - \lambda_1\{\hat{\mathbf{\Phi}}_x^w(l)\}\right)$$

$$\hat{\phi}_{x_1}(l) = \lambda_1\{\hat{\mathbf{\Phi}}_x^w(l)\}/||\hat{\mathbf{b}}||_2^2$$

[Markovich-Golan, Gannot, *ICASSP 2015*] [Kodrasi, Doclo, *HSCMA 2017*]
[Kodrasi, Doclo, *IEEE/ACM Trans. Audio, Speech and Language Processing*, June 2018]

Joint Estimation of RETF vector and PSDs

2. **Alternating least squares (ALS) method,** minimizing Frobenius norm
   - **Model noise covariance matrix + estimate noise PSD**

$$\min_{\phi_{x_1}(l),\phi_d(l),\phi_v(l),\mathbf{a}(l)} ||\hat{\mathbf{\Phi}}_y(l) - \phi_{x_1}(l)\mathbf{a}(l)\mathbf{a}^H(l) - \phi_d(l)\mathbf{\Gamma} - \phi_v(l)\mathbf{\Psi}||_F^2$$

   - **No closed-form solution → two-step alternating procedure** (least-squares problem for PSDs, eigenvalue problem for RETF vector)

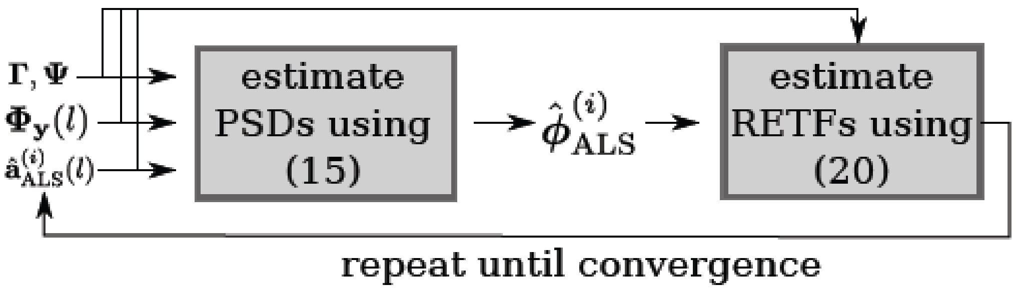


Fig. 1: Block diagram of ALS-based RETF vector and PSD estimation.

Algorithm 1: ALS method to jointly estimate the RETF vector and PSDs.

Input: $\mathbf{\Gamma}(k)$, $\mathbf{\Psi}(k)$, $\hat{\mathbf{\Phi}}_{\mathbf{y}}(k,l)$, iterations N, init. $\hat{\mathbf{a}}^{(0)}(k,1)$
Output: $\hat{\mathbf{a}}_{\text{ALS}}(k,l)$, $\hat{\phi}_{\text{ALS}}(k,l)$
for all (k,l) **do**
  for $i=1:N$ **do**
    compute $\mathbf{A}^{(i-1)}(k,l)$ using (16) and $\mathbf{b}^{(i-1)}(k,l)$ using (17)
    $\hat{\phi}^{(i)}_{\text{ALS}}(k,l) = \left(\mathbf{A}^{(i-1)}(k,l)\right)^{-1}\mathbf{b}^{(i-1)}(k,l)$ (15)
    constrain $\hat{\phi}^{(i)}_{\text{ALS}}(k,l)$ using (25)
    $\hat{\mathbf{\Phi}}^{(i)}_{\mathbf{x}}(k,l) =$
      $\hat{\mathbf{\Phi}}_{\mathbf{y}}(k,l) - (\hat{\phi}^{(i)}_{d,\text{ALS}}(k,l)\mathbf{\Gamma}(k) + \hat{\phi}^{(i)}_{v,\text{ALS}}(k,l)\mathbf{\Psi}(k))$
    $\hat{\mathbf{\Phi}}^{(i)}_{\mathbf{x}}(k,l) = \hat{\mathbf{N}}^{(i)}(k,l)\hat{\mathbf{\Lambda}}^{(i)}(k,l)\hat{\mathbf{N}}^{(i),H}(k,l)$ (EVD)
    $\hat{\mathbf{a}}^{(i)}_{\text{ALS}}(k,l) = \sqrt{\hat{\lambda}^{(i)}_1(k,l)/\hat{\phi}^{(i)}_{s,\text{ALS}}(k,l)}\,\hat{\nu}^{(i)}_1(k,l)$ (20)
  end
  $\hat{\mathbf{a}}^{(1)}_{\text{ALS}}(k,l+1) = \hat{\mathbf{a}}^{(N)}_{\text{ALS}}(k,l)/(\mathbf{e}^T\hat{\mathbf{a}}^{(N)}_{\text{ALS}}(k,l))$ (for next frame)
end

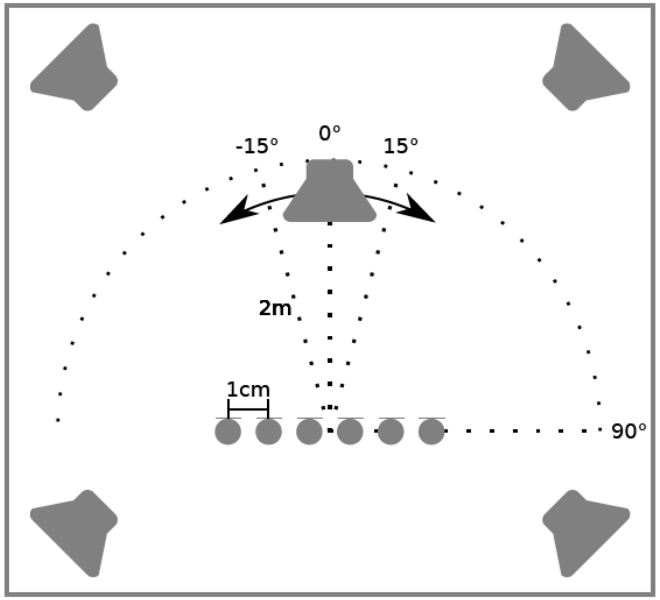
Simulation results

1. Simulated stationary source (ACE)

- Linear microphone array (M=6, d=6cm)
- Target source at 15° (measured room impulse responses, $T_{60} \approx 1.25$ s)
- Simulated diffuse babble noise (SDR=10 dB)

2. Recorded moving source (varechoic lab)

- Linear microphone array (M=6, d=1cm)
- Moving target source ($T_{60} \approx 0.35$ s)
- Recorded pseudo-diffuse babble noise (SDR=10 dB)

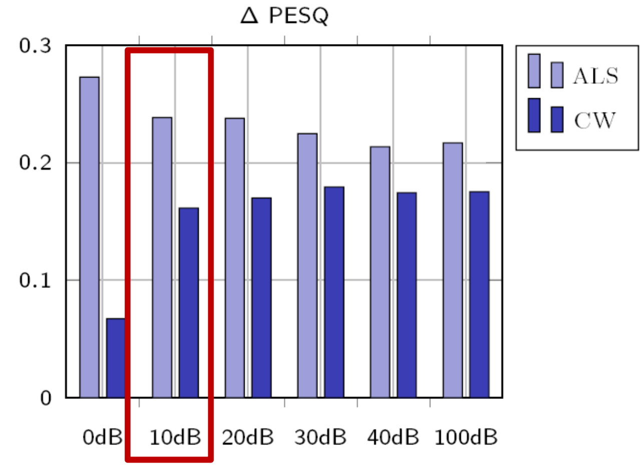


Simulation parameters:

- f_s = 16 kHz, STFT: 64 ms, 75% overlap, Hamming window
- Γ: spherically diffuse; smoothing: 40 ms; speech PSD estimated using decision-directed approach, G_{min} = -10 dB
- CW: noise covariance matrix estimated during first second; ALS: 5 iterations

Simulation results (PESQ improvement)

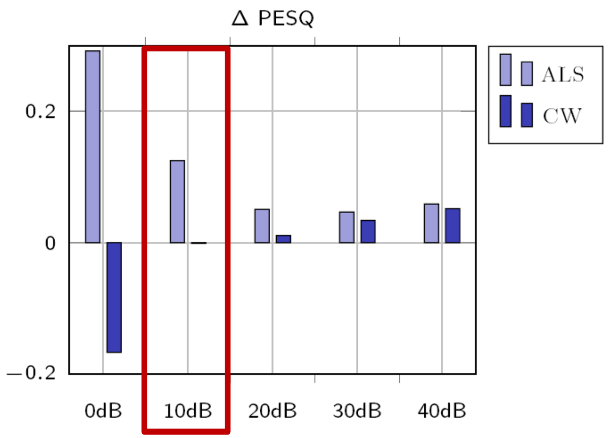
1. Simulated stationary source



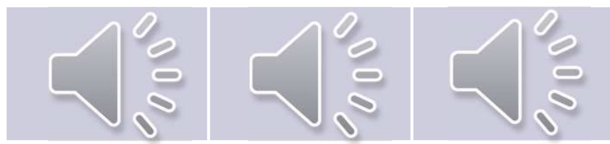
Input	MWF CW	MWF ALS

Linear array (M=6, d=6cm), fs=16kHz, stationary source at θ=15º, perfectly diffuse babble noise (SDR=10dB), sensor noise (DNR=10dB)

2. Recorded moving source



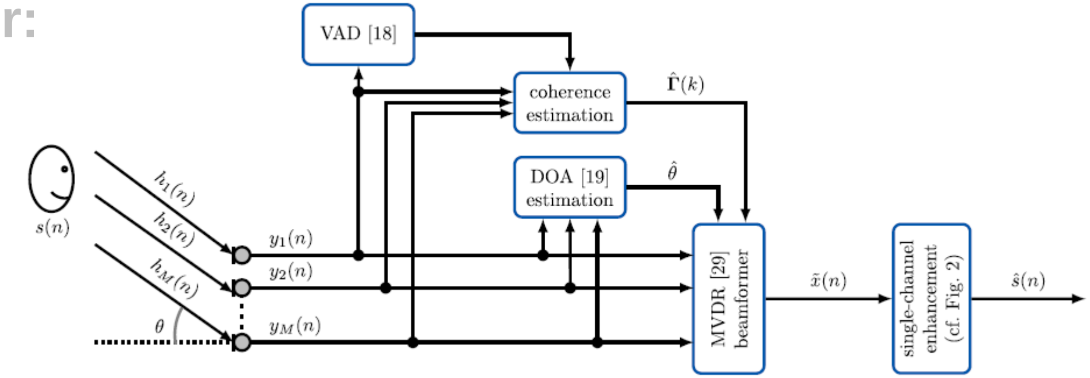
Input	MWF CW	MWF ALS



Linear array (M=6, d=1cm), fs=16kHz, moving source θ=0º to θ=90º pseudo-diffuse babble noise (SDR=10dB), sensor noise (DNR=10dB)

Multi-microphone dereverberation and noise reduction

1. **Beamforming + spectral postfilter:** *multiply* each time-frequency bin with *real-valued* gain

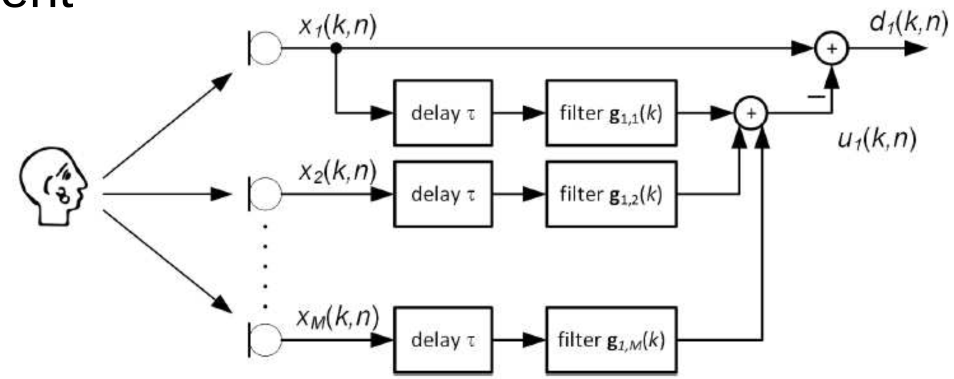


$$\mathbf{y}(l) = \mathbf{a}(l)x_1(l) + \mathbf{x}_r(l) + \mathbf{v}(l)$$

2. **Reverberation and noise suppression:** *subtract complex-valued* estimate of late reverberant and noise component

$$y_m(l) = h_m(l) \star s(l) + v_m(l)$$

$$\hat{x}_{e,1}(l) = y_1(l) - \mathbf{Y}_\tau(l)\mathbf{g}(l)$$

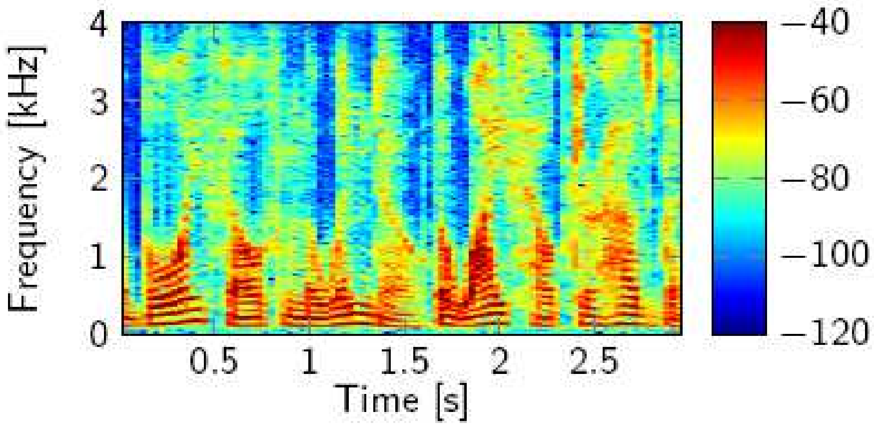


Reverberation suppression

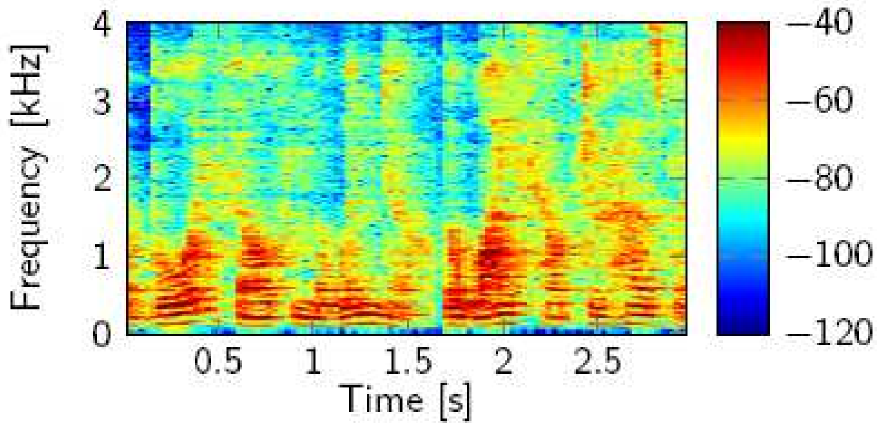
- **Goal**: estimate clean speech STFT coefficients $s(k,l)$ from reverberant (and noisy) STFT coefficients $y_m(k,l)$ by subtracting late reverberant component

$$y_m(k,l) = \underbrace{h_m(k,l) \star s(k,l)}_{x_m(k,l)} + v_m(k,l)$$

  - Probabilistic estimation using (statistical) **models of desired speech signal and reverberation**
  - Exploit **sparsity properties** of speech in STFT-domain



Clean



Reverberant

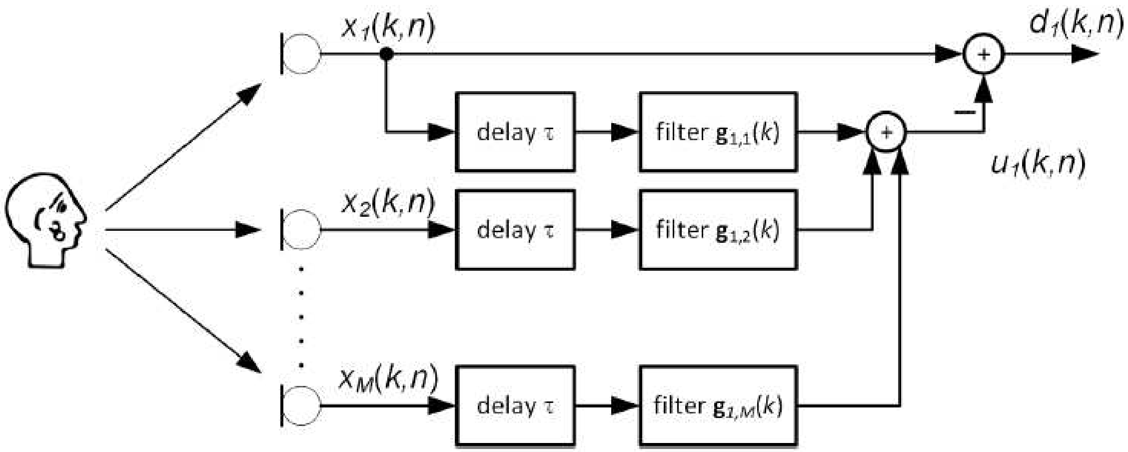
Multi-channel linear prediction

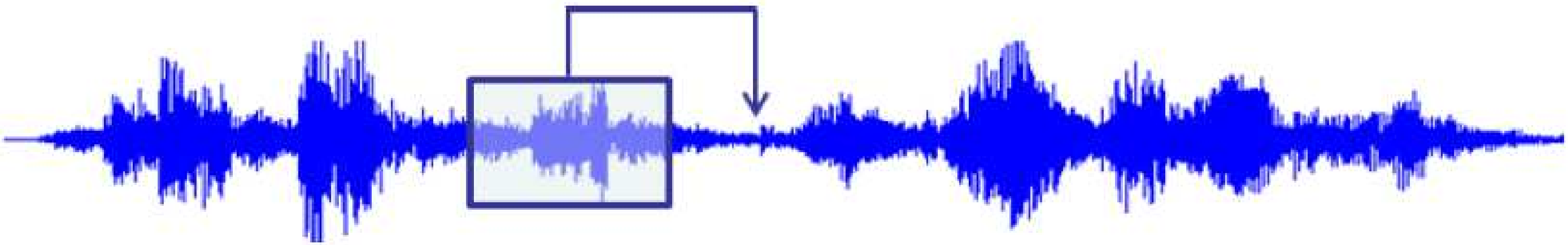
- **AR model of reverberant speech**

$$\mathbf{x}_1(k) = \mathbf{d}(k) + \mathbf{X}_\tau(k)\mathbf{g}(k)$$

$$\hat{\mathbf{d}}(k) = \mathbf{x}_1(k) - \mathbf{X}_\tau(k)\hat{\mathbf{g}}(k)$$

predicted reverberation





How to select suitable cost function for prediction filters ?

Multi-channel linear prediction

- **Approach:**
  - STFT coefficients of desired signal are modelled using **circular sparse/super-Gaussian prior with time-varying variance** $\lambda(n)$

$$\rho(d(n)) = \max_{\lambda(n)>0} \mathcal{N}_{\mathbb{C}}(d(n); 0, \lambda(n)) \boxed{\psi(\lambda(n))}$$

    Scaling function $\psi(.)$ can be interpreted as **hyper-prior on variance**

  - **Maximum-Likelihood Estimation** (batch, per frequency bin)

$$\mathcal{L}(\mathbf{g}) = \prod_{n=1}^{N} \rho(d(n)) \Rightarrow \min_{\boldsymbol{\lambda}>0, \mathbf{g}} \sum_{n=1}^{N} \left(\frac{|d(n)|^2}{\lambda(n)} + \log \pi\lambda(n) \boxed{- \log \psi(\lambda(n))} \right)$$

  - **Alternating optimization procedure**
    1. Estimate **prediction vector** (assuming fixed variances)

$$\hat{\mathbf{g}}^{(i+1)} = \left(\mathbf{X}_\tau^H \mathcal{D}_{\hat{\boldsymbol{\lambda}}^{(i)}}^{-1} \mathbf{X}_\tau \right)^{-1} \mathbf{X}_\tau^H \mathcal{D}_{\hat{\boldsymbol{\lambda}}^{(i)}}^{-1} \mathbf{x}_1$$

    2. Estimate **variances** (assuming fixed prediction vector)

$$\hat{\lambda}^{(i+1)}(n) = \arg\min_{\lambda(n)>0} \frac{|\hat{d}^{(i+1)}(n)|^2}{\lambda(n)} + \log \pi\lambda(n) \boxed{- \log \psi(\lambda(n))}$$

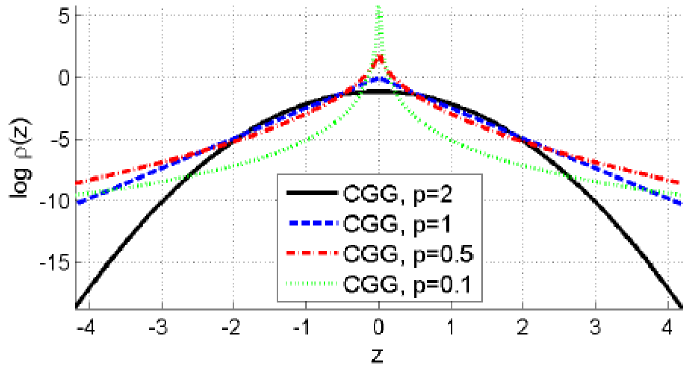
[Nakatani, Yoshioka, Kinoshita, Miyoshi, Juang, *IEEE/ACM Trans. Audio Speech Language Proc.*, Sep. 2010.]
[Jukić, van Waterschoot, Gerkmann, Doclo, *IEEE/ACM Trans. Audio Speech Language Proc.*, Sep. 2015.]

Multi-channel linear prediction

- **Example:** complex generalized Gaussian (CGG) prior with shape parameter *p*

$$\rho(z) = \frac{p}{2\pi\gamma\Gamma(2/p)} e^{-\frac{|z|^p}{\gamma^{p/2}}}$$

$$\boxed{\hat{\lambda}^{(i+1)}(n) = |\hat{d}^{(i+1)}(n)|^{2-p},}$$



- **Remarks:**

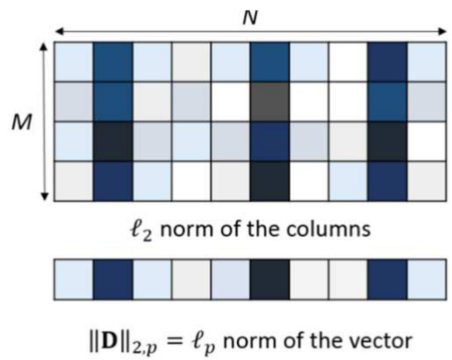
1. ML estimation using CGG prior is equivalent to ***l_p*-norm minimization** → **promotes sparsity of TF-coefficients across time** (for p < 2)

$$\boxed{\min_{\mathbf{g}} \|\mathbf{d}\|_p^p,}$$

2. Incorporate additional knowledge of speech signal, e.g. **low-rank structure** (NMF)

$$|\mathbf{D}|^2 \approx \underbrace{\mathbf{W}}_{\text{spectral dictionary}} \mathbf{H}$$

3. **Group sparsity** for MIMO speech dereverberation → mixed norms

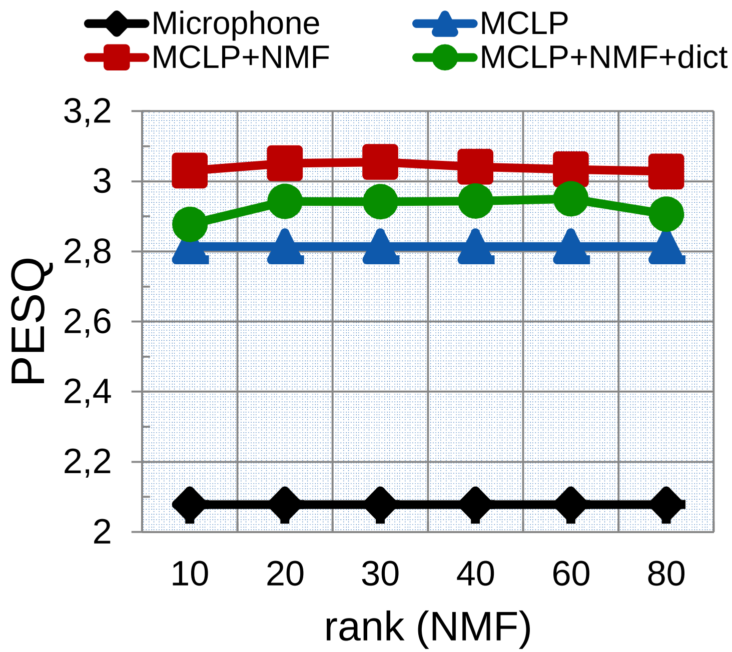


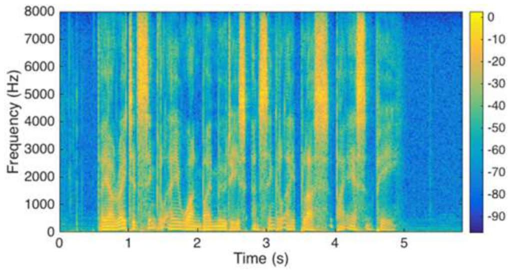
4. **Recursive version** by constraining MCLP-based estimate of undesired component

[Jukić, van Waterschoot, Gerkmann, Doclo, *IEEE/ACM Trans. Audio Speech Language Proc.*, Sep. 2015.]
[Jukić, van Waterschoot, Doclo, *IEEE Signal Processing Letters*, Jan. 2017.]

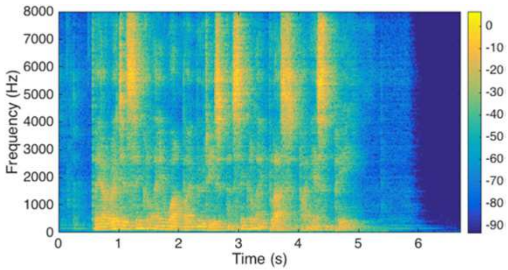
Multi-channel linear prediction

- **Instrumental validation (noiseless, batch)**
  - MCLP exploits sparsity
  - NMF introduces speech structure (unsupervised vs. supervised NMF)

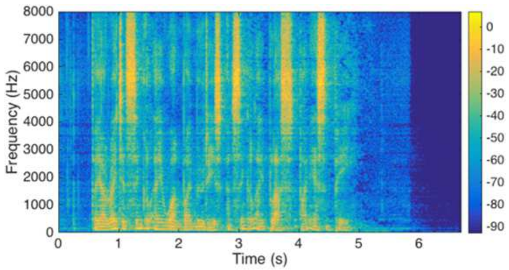




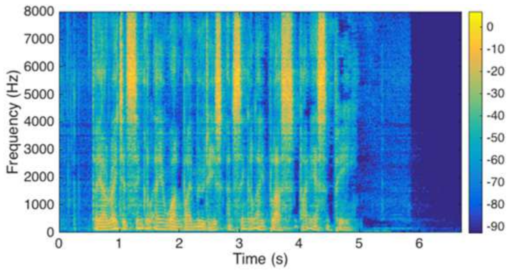
Clean



Microphone



MCLP



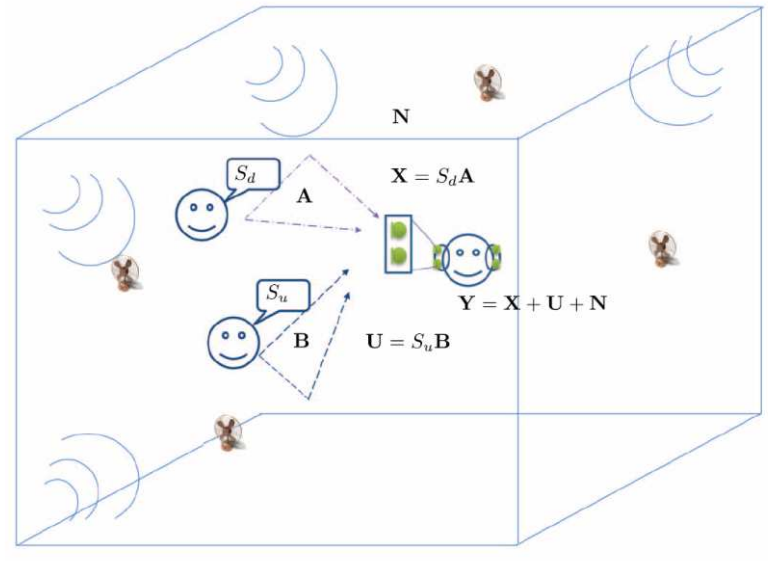
MCLP+NMF

T_{60} ~ 700ms, M=4, fs=16 kHz; STFT: 64ms (overlap 16ms); MCLP: L_g=8, τ=2, p=0

[Jukić, van Waterschoot, Gerkmann, Doclo, *IEEE/ACM Trans. Audio Speech Language Proc.*, Sep. 2015.]

Current/future work

- Estimation of RETF vectors and PSDs for **multi-speaker scenarios** (e.g. based on Procrustes problem)
- **Joint noise reduction and dereverberation:** integration of multi-channel linear prediction and generalized sidelobe canceller



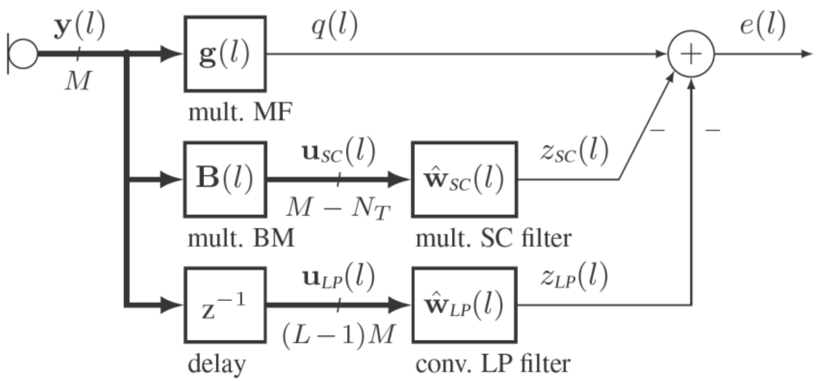


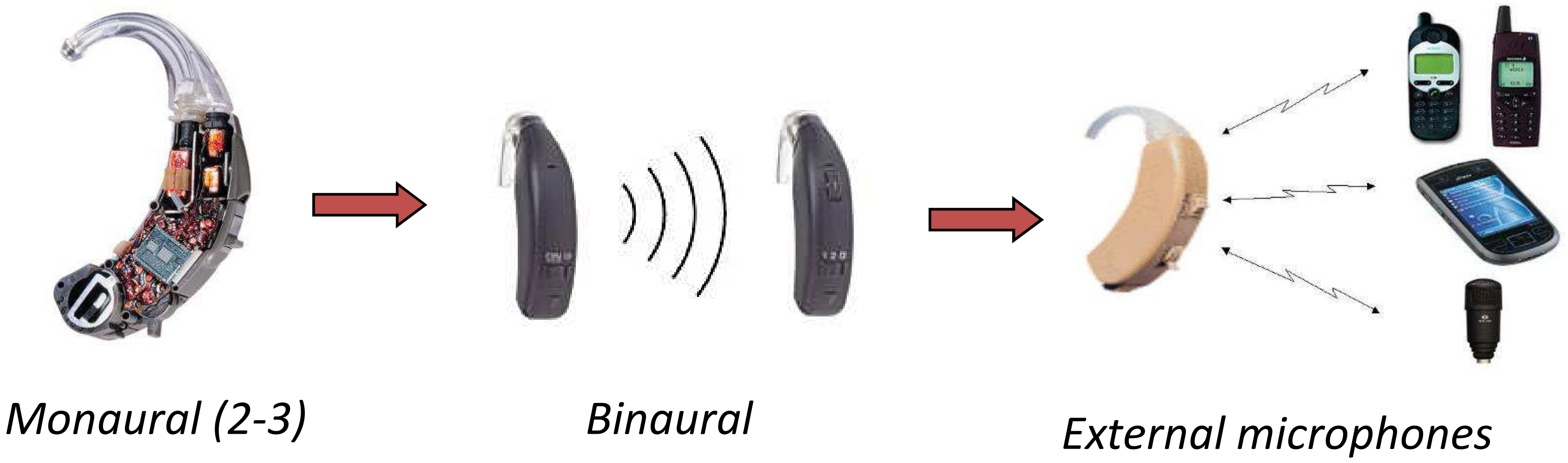
Fig. 1. The integrated sidelobe cancellation and linear prediction (ISCLP) architecture.

[Dietzen, Doclo, Moonen, van Waterschoot, *IEEE/ACM Trans. Audio Speech Language Proc.*, in review.]
[Dietzen, Doclo, Moonen, van Waterschoot, *IWAENC 2018*.]

2. Acoustic signal processing for binaural hearing devices

Hearing devices / hearables

- Hearing devices generally have multiple microphones available and allow for advanced acoustical signal pre-processing



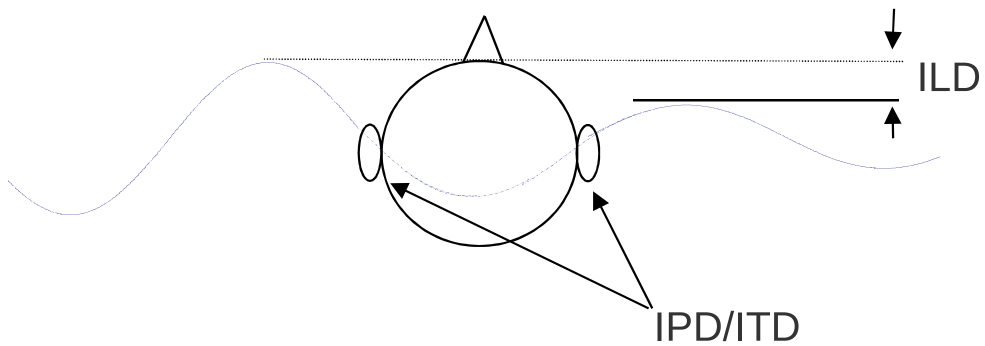
- Main objectives of binaural speech enhancement algorithms: improve speech intelligibility + **preserve spatial awareness (binaural cues)**

[S. Doclo, W. Kellermann, S. Makino, S. Nordholm, *Multichannel signal enhancement algorithms for assisted listening devices*, *IEEE Signal Processing Magazine*, Mar. 2015.]

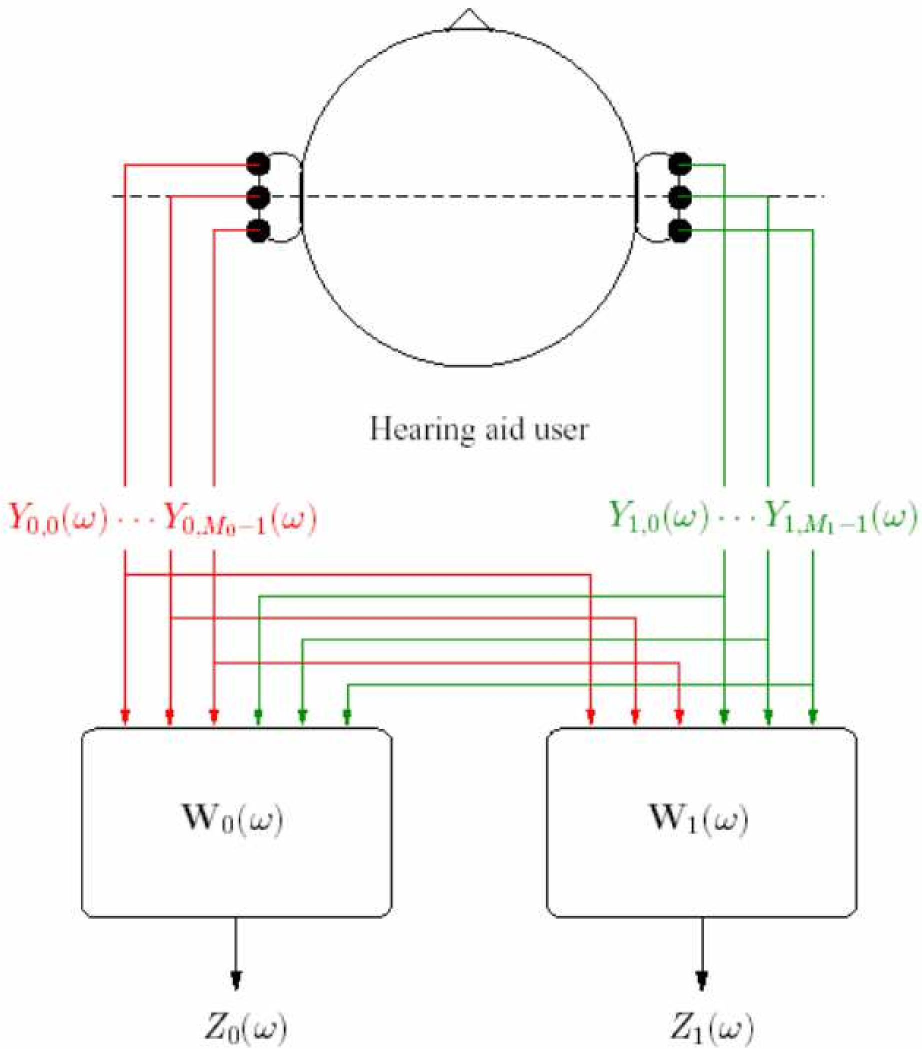
Binaural auditory cues

- **Interaural Time/Phase Difference (ITD/IPD)**
  Interaural Level Difference (ILD)
  Interaural Coherence (IC)
  - ITD: f < 1500 Hz, ILD: f > 2000 Hz
  - IC: describes spatial characteristics, e.g. perceived width, of diffuse noise, and determines when ITD/ILD cues are *reliable*

- Binaural cues, in addition to spectro-temporal cues, play an important role in auditory scene analysis (source segregation) and speech intelligibility



Binaural noise reduction: Configuration



- Binaural hearing aid configuration:
  - Two hearing aids with in total *M* microphones
  - All microphone signals **Y** are assumed to be available at both hearing aids (perfect wireless link)
- Apply a filter $\mathbf{W}_0$ and $\mathbf{W}_1$ at the left and the right hearing aid, generating binaural output signals Z_0 and Z_1

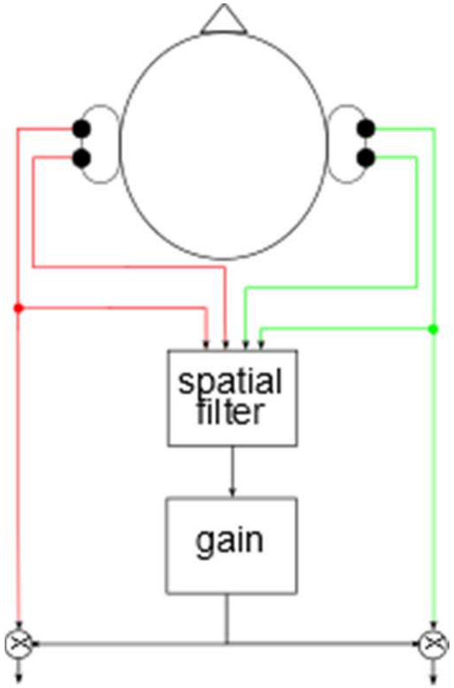
$$\boxed{Z_0(\omega) = \mathbf{W}_0^H(\omega)\mathbf{Y}(\omega), \quad Z_1(\omega) = \mathbf{W}_1^H(\omega)\mathbf{Y}(\omega)}$$

[S. Doclo, W. Kellermann, S. Makino, S. Nordholm, *Multichannel signal enhancement algorithms for assisted listening devices*, *IEEE Signal Processing Magazine*, Mar. 2015.]

Binaural noise reduction: Two main paradigms

Spectral post-filtering (based on multi-microphone noise reduction)

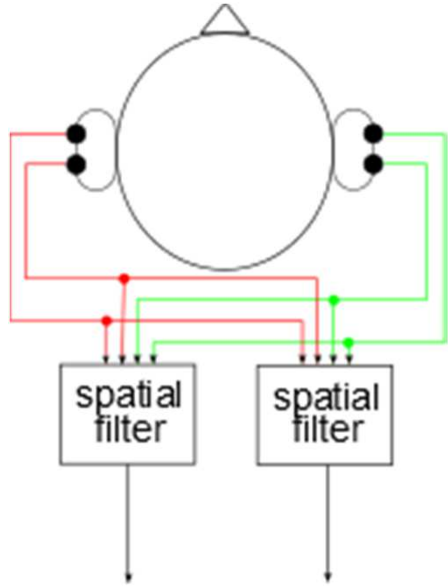
[Wittkop 2003, Lotter 2006, Rohdenburg 2008, Grimm 2009, Kamkar-Parsi 2011, Reindl 2013, Baumgärtel 2015, Enzner 2016]



- ⊕ Binaural cue preservation
- ⊖ Possible single-channel artifacts

Binaural spatial filtering techniques

[Welker 1997, Aichner 2007, Doclo 2010, Cornelis 2012, Hadad 2015-2016, Marquardt 2015-2018, Koutrouvelis 2017-2019]



- ⊕ Larger noise reduction performance
- ⊕ Merge spatial and spectral post-filtering
- ⊖ Binaural cue preservation not guaranteed

Binaural MVDR and MWF

Minimum-Variance-Distortionless-Response (MVDR) beamformer

Goal: minimize output noise power without distorting speech component in reference microphone signals

$$\min_{\mathbf{W}_0} \mathbf{W}_0^H \mathbf{R}_v \mathbf{W}_0 \quad \text{subject to} \quad \mathbf{W}_0^H \mathbf{A} = A_0$$

$$\min_{\mathbf{W}_1} \underbrace{\mathbf{W}_1^H \mathbf{R}_v \mathbf{W}_1}_{\text{noise reduction}} \quad \text{subject to} \quad \underbrace{\mathbf{W}_1^H \mathbf{A} = A_1}_{\text{distortionless constraint}}$$

Requires estimate/model of noise coherence matrix (e.g. diffuse) and estimate/model of relative transfer function (RTF) of target speech source

Multi-channel Wiener Filter (MWF)

Goal: estimate speech component in reference microphone signals + trade off noise reduction and speech distortion

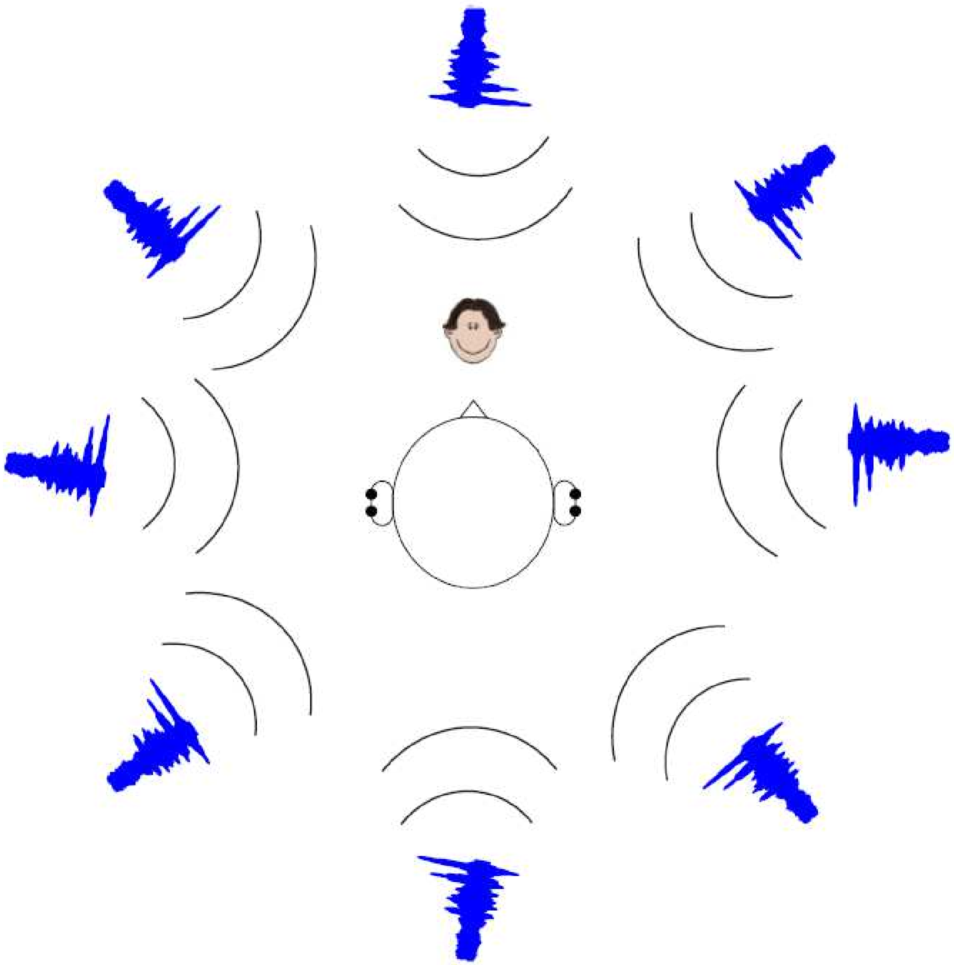
$$J_{\mathrm{MWF}}(\mathbf{W}) = \mathcal{E}\left\{ \underbrace{\left\| \begin{bmatrix} X_0 - \mathbf{W}_0^H \mathbf{X} \\ X_1 - \mathbf{W}_1^H \mathbf{X} \end{bmatrix} \right\|^2}_{\text{speech distortion}} + \mu \underbrace{\left\| \begin{bmatrix} \mathbf{W}_0^H \mathbf{V} \\ \mathbf{W}_1^H \mathbf{V} \end{bmatrix} \right\|^2}_{\text{noise reduction}} \right\}$$

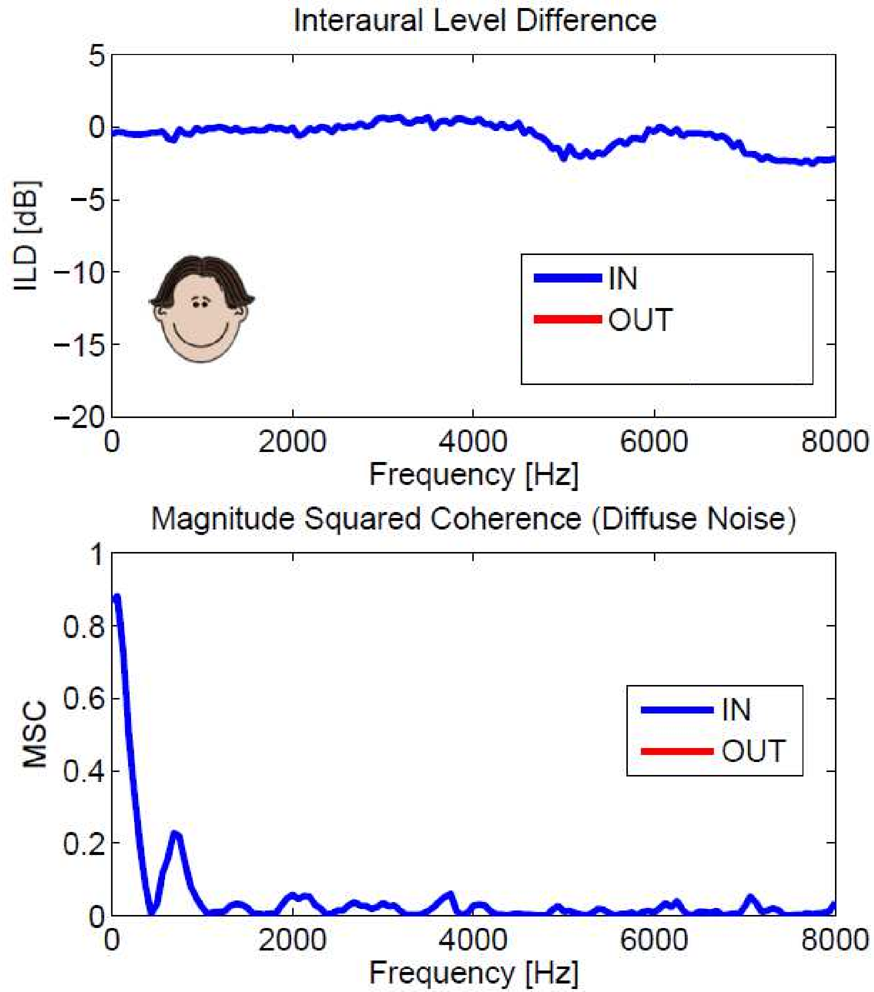
Requires estimate of speech and noise covariance matrices, e.g. based on SPP

Can be decomposed as binaural MVDR beamformer and spectral postfilter

Good noise reduction performance, what about binaural cues ?

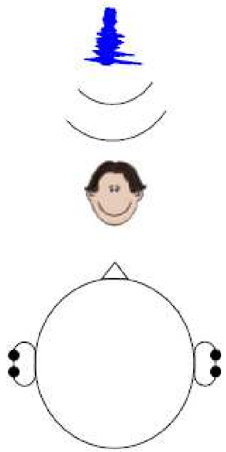
Binaural MVDR/MWF: binaural cues

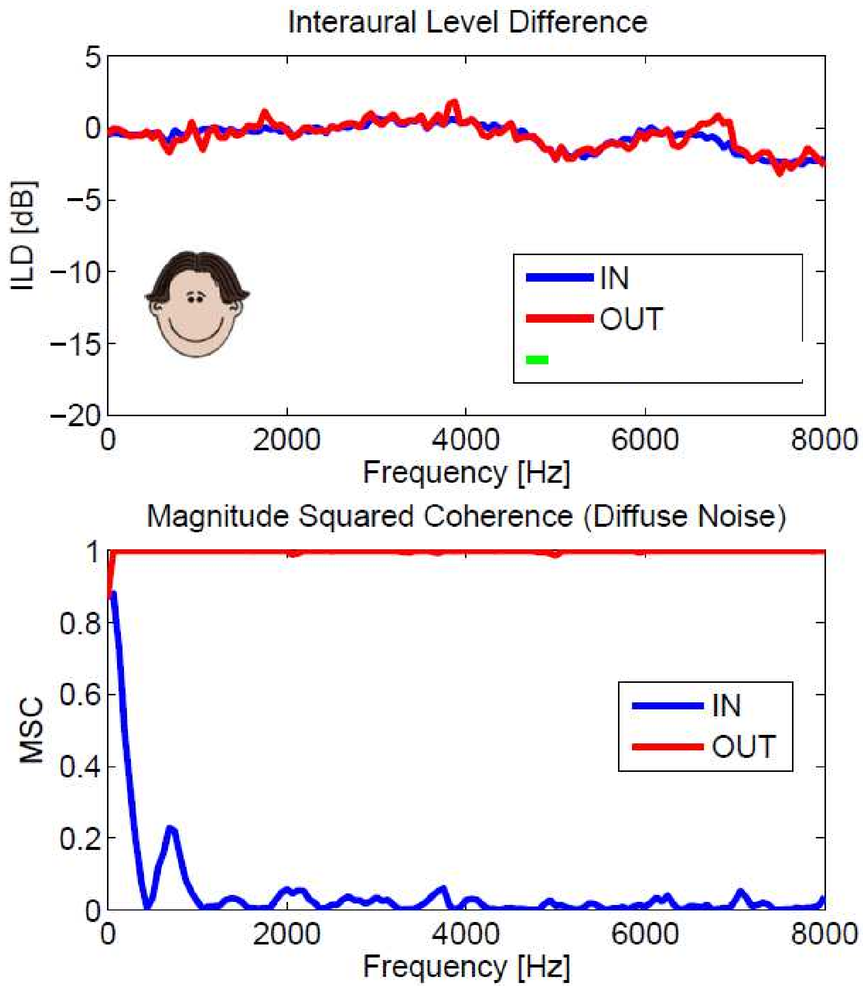




Note: MSC = Magnitude Squared Coherence

Binaural MVDR/MWF: binaural cues





Binaural cues for residual noise/interference in binaural MVDR/MWF not preserved

Binaural MWF: Extensions for diffuse noise

Binaural MWF

- ⊕ SNR improvement
- ⊕ Binaural cues of speech source
- ⊖ Binaural cues of noise

Interaural coherence preservation (MWF-IC)

$$J_{MWF-IC}(\mathbf{W}) = J_{MWF}(\mathbf{W}) + \lambda \left| \frac{\mathbf{W}_0^H \mathbf{R}_v \mathbf{W}_1}{\sqrt{\mathbf{W}_0^H \mathbf{R}_v \mathbf{W}_0 \mathbf{W}_1^H \mathbf{R}_v \mathbf{W}_1}} - IC_v^{des} \right|^2$$

⊖ No closed-form solution, iterative optimization procedures required

Partial noise estimation (MWF-N)

$$J_{\mathrm{MWF-N}}(\mathbf{W}) = \mathcal{E}\left\{ \left\| \begin{bmatrix} X_0 - \mathbf{W}_0^H \mathbf{X} \\ X_1 - \mathbf{W}_1^H \mathbf{X} \end{bmatrix} \right\|^2 + \mu \left\| \begin{bmatrix} \eta V_0 - \mathbf{W}_0^H \mathbf{V} \\ \eta V_1 - \mathbf{W}_1^H \mathbf{V} \end{bmatrix} \right\|^2 \right\}$$

⊕ Closed-form solution (mixing with reference microphone signals)

⊜ **Trade-off** between SNR improvement and binaural cue preservation, depending on **parameters** (η and λ)

[Marquardt, Hohmann, Doclo, *IEEE/ACM Trans. Audio, Speech and Language Processing*, Dec. 2015.]
[Marquardt, Doclo, *IEEE/ACM Trans. Audio, Speech and Language Processing*, Jul. 2018.]

Binaural MWF: Extensions for diffuse noise

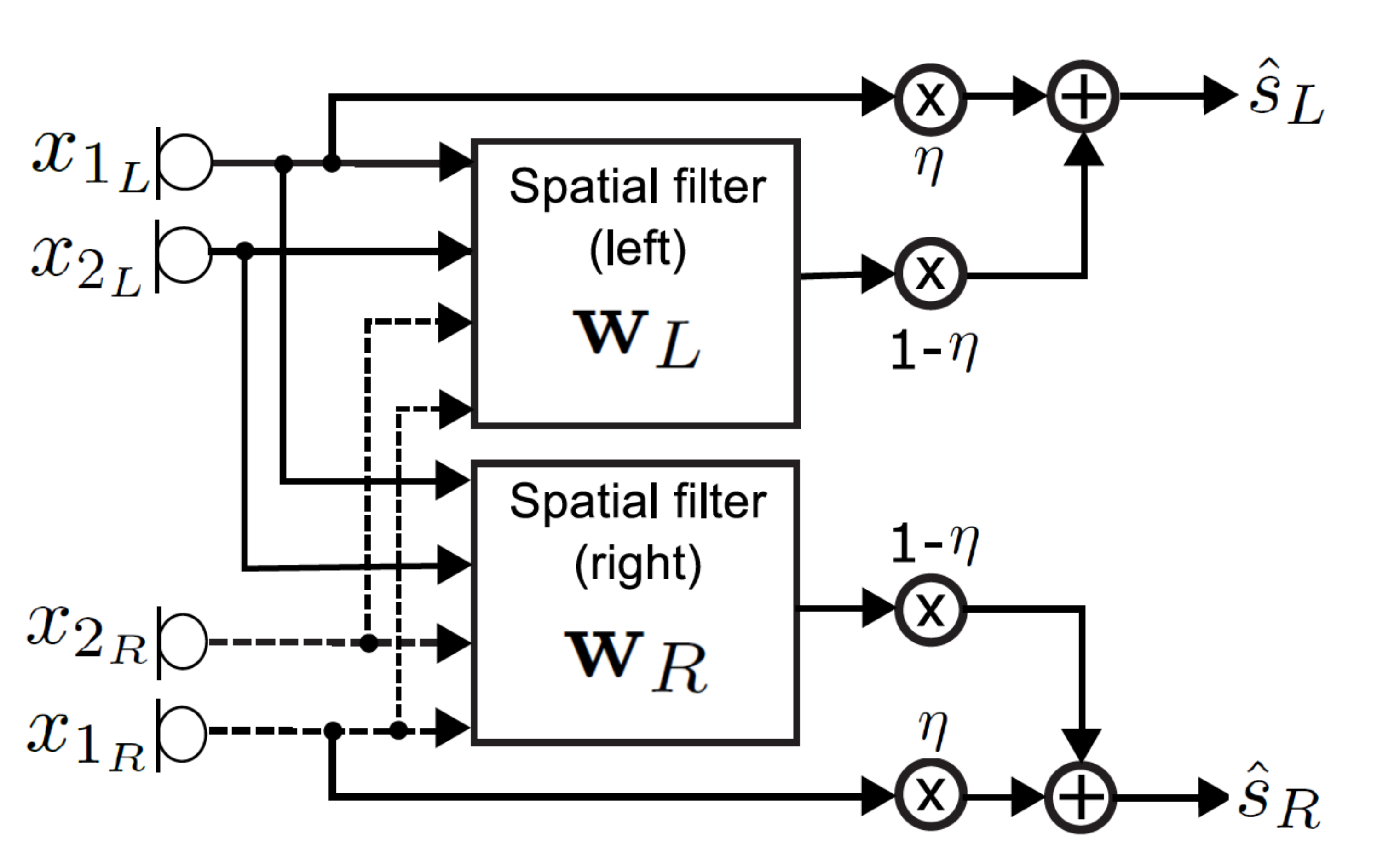
SNR improvement

…eech source

…ise

In… pre…

$J_{MWF-IC}(\mathbf{W}) = J_{M}$ …



… $\mathbf{W}_0^H \mathbf{V}$ … $\mathbf{W}_1^H \mathbf{V}$ …

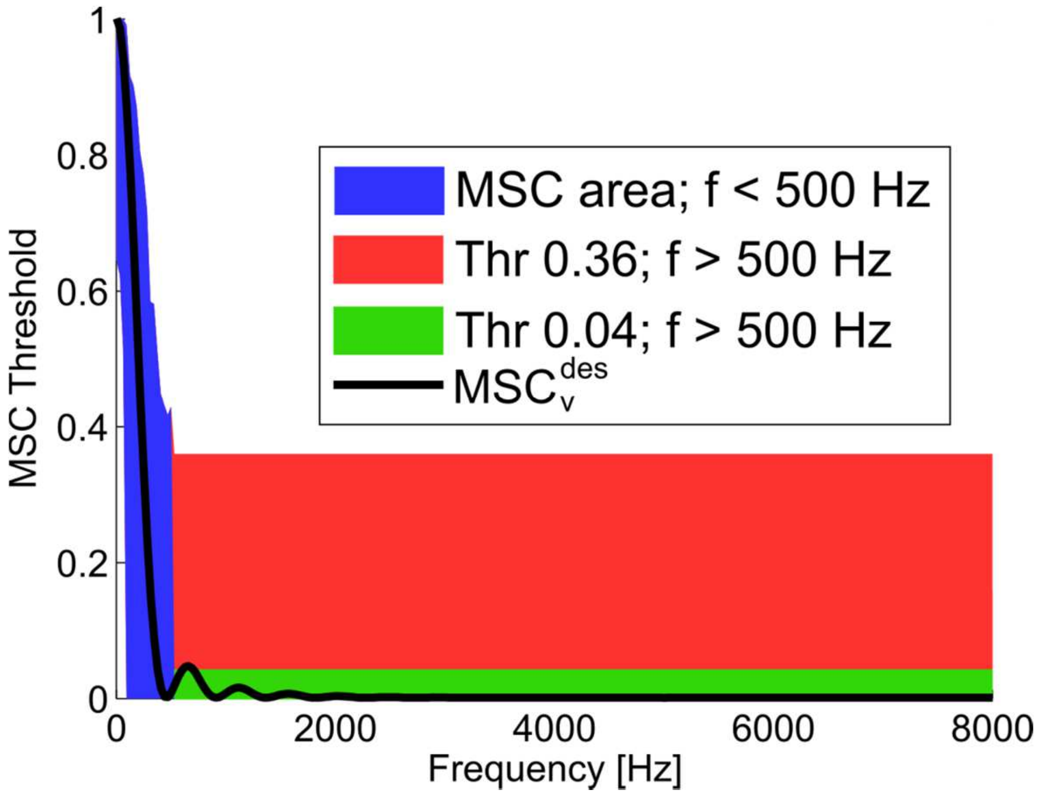
No clo… optim… …ng with …als)

Trade-off between SNR improvement and binaural cue preservation, depending on **parameters** (η and λ)

[Marquardt, Hohmann, Doclo, *IEEE/ACM Trans. Audio, Speech and Language Processing*, Dec. 2015.]
[Marquardt, Doclo, *IEEE/ACM Trans. Audio, Speech and Language Processing*, Jul. 2018.]

Trade-off parameters for binaural MVDR/MWF

- **Fixed broadband values** (η = 0.1 … 0.3)
- **Frequency-dependent values** based on IC discrimination ability of human auditory system

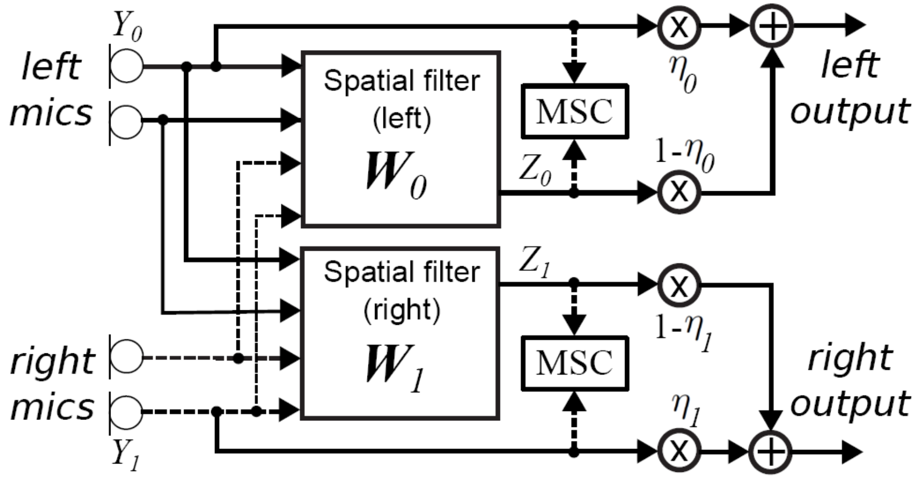


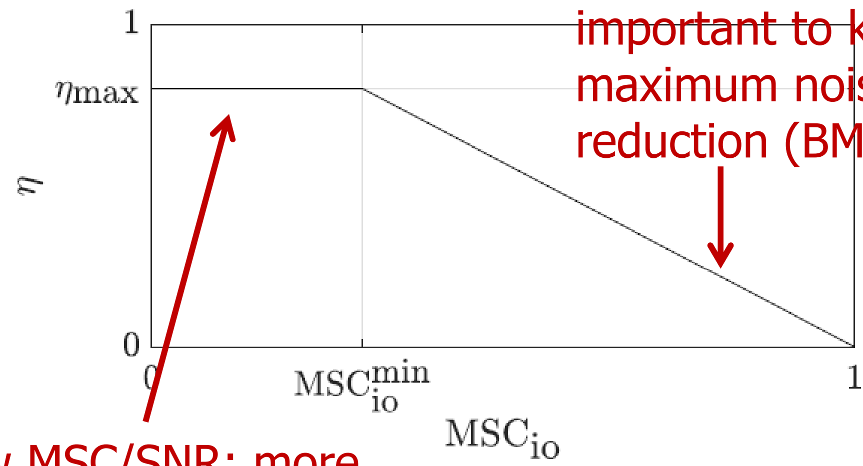
- IC discrimination ability depends on magnitude of reference IC
- **Boundaries on Magnitude Squared Coherence** (MSC=|IC|^2) :
  - For f < 500 Hz ("large" IC): frequency-dependent MSC boundaries (**blue**)
  - For f > 500 Hz ("small" IC): fixed MSC boundary, e.g. 0.36 (**red**) or 0.04 (**green**)

[Marquardt, Doclo, *IEEE/ACM Trans. Audio, Speech and Language Processing*, Jul. 2018.]

Trade-off parameters for binaural MVDR/MWF

- **Fixed broadband values** (η = 0.1 … 0.3)
- **Frequency-dependent values** based on IC discrimination ability of human auditory system
- **Frequency-dependent function** of MSC between noisy reference microphone signals and output signals of BMVDR beamformer



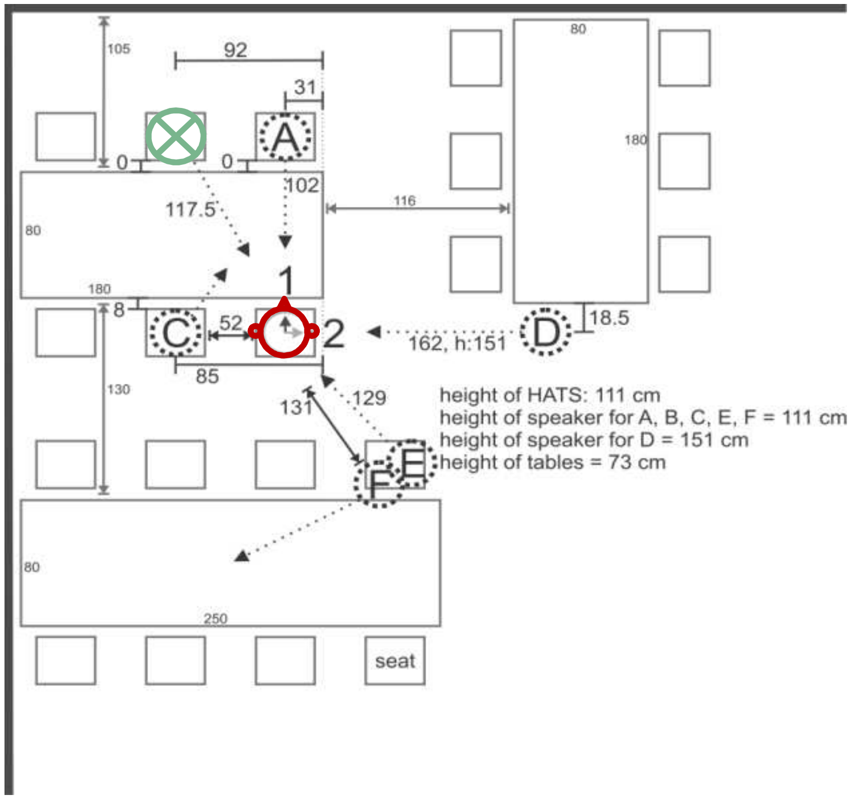


High MSC/SNR: more important to keep maximum noise reduction (BMVDR)

Low MSC/SNR: more important to preserve binaural cues (scaled input signals)

[Marquardt, Doclo, *IEEE/ACM Trans. Audio, Speech and Language Processing*, Jul. 2018.]
[Klug, Marquardt, Gößling, Doclo, *ITG Conference Speech Communication*, Oct. 2018.]

Evaluation: Test setup



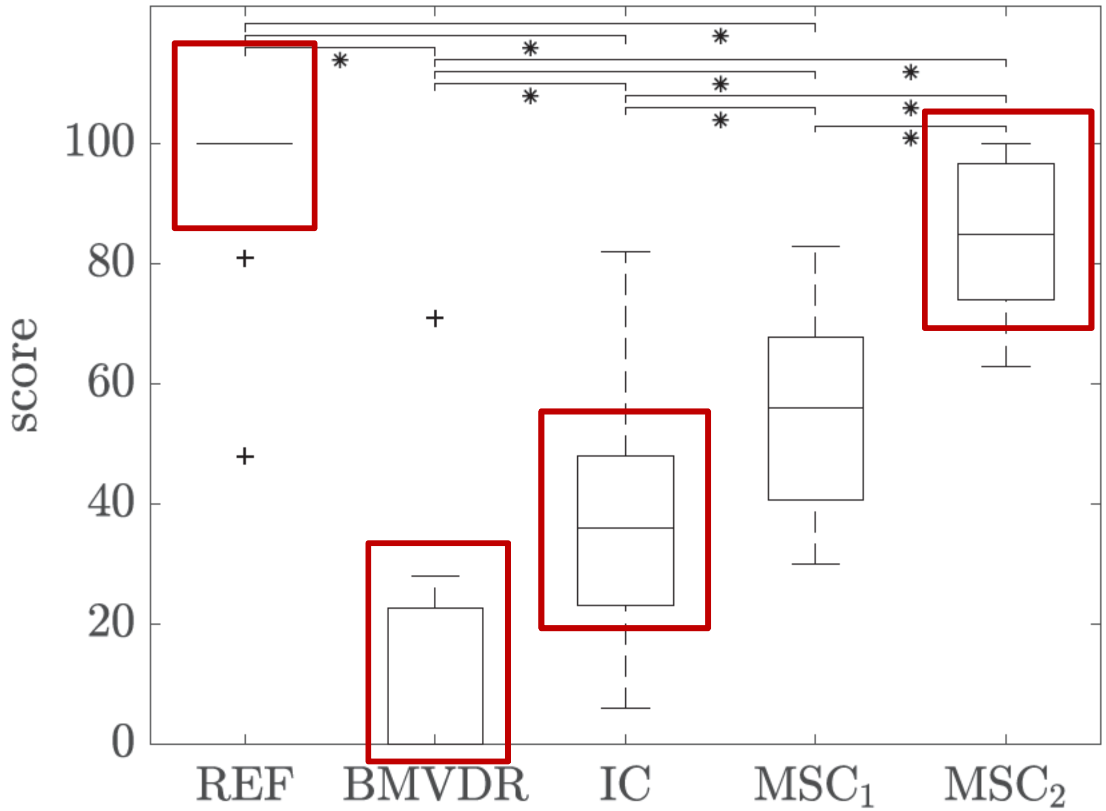
- Binaural hearing aid recordings (M=4 mics) in **cafeteria** (T_{60}≈1250 ms)
  - Target speaker at -35°
  - Realistic cafeteria ambient noise
- **Algorithms**: binaural MVDR and binaural MVDR-N with different trade-off parameters:
  - MVDR-IC
  - MVDR-MSC1: η_{max}=0.7, MSC_{min}=0
  - MVDR-MSC2: η_{max}=1.0, MSC_{min}=0.1
- **Subjective listening experiments**:
  - 11 normal-hearing subjects
  - **SRT** using Oldenburg Sentence Test (OLSA)
  - **Spatial quality (diffuseness)** using MUSHRA

Does binaural unmasking compensate for SNR decrease of cue preservation algorithms (MVDR-N) ?

[Klug, Marquardt, Gößling, Doclo, *ITG Conference Speech Communication*, Oct. 2018.]
[Gößling, Marquardt, Doclo, *Trends in Hearing, in revision*.]

Evaluation: Spatial quality (MUSHRA)

- Evaluate spatial difference between reference microphone signals and binaural output signals
- **MVDR-N outperforms BMVDR**
  - Trade-off parameters: MSC-based better than IC-based
  - Using MSC2 hardly any difference to input !

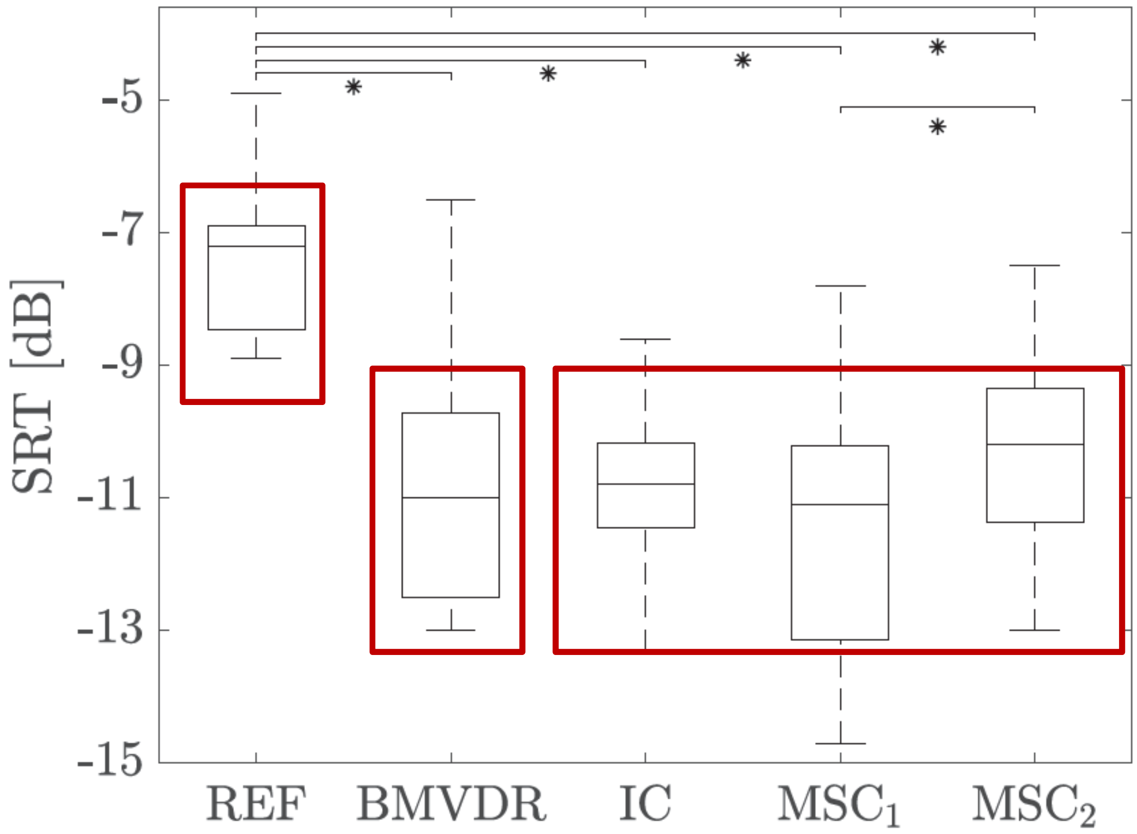


Binaural cue preservation for diffuse noise significantly improves spatial quality

[Klug, Marquardt, Gößling, Doclo, *ITG Conference Speech Communication*, Oct. 2018.]
[Gößling, Marquardt, Doclo, *Trends in Hearing, in revision*.]

Evaluation: Speech intelligibility (SRT)

- All algorithms show a highly significant speech reception threshold (SRT) improvement
- **No significant SRT difference between BMVDR and MVDR-N**

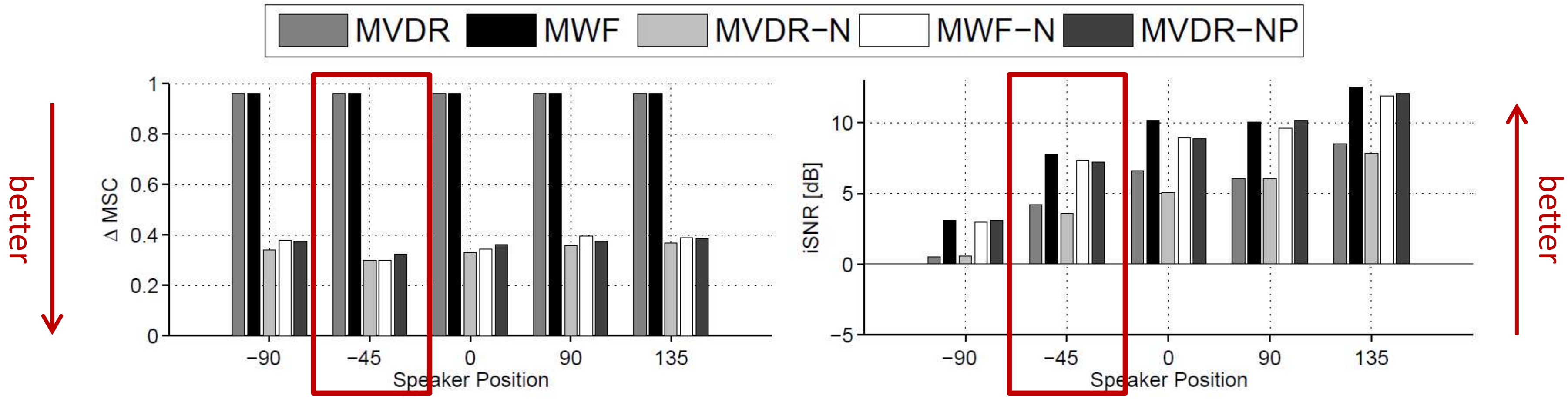


Binaural cue preservation for diffuse noise does not affect speech intelligibility

[Klug, Marquardt, Gößling, Doclo, *ITG Conference Speech Communication*, Oct. 2018.]
[Gößling, Marquardt, Doclo, *Trends in Hearing, in revision*.]

Binaural MVDR/MWF: Sound samples

Input	MVDR	MWF	MVDR-N	MWF-N	MVDR-NP



(a) $\Delta\mathrm{MSC}_{\mathrm{v}}$ (b) $\mathrm{iSNR}_{\mathrm{be}}^{\mathrm{out}}$

Cafeteria with recorded ambient noise, speaker at -45°, 0 dB input iSNR (left hearing aid)
MVDR: anechoic ATF, DOA known, spatial coherence matrix calculated from anechoic ATFs / MWF = MVDR + postfilter (SPP-based)

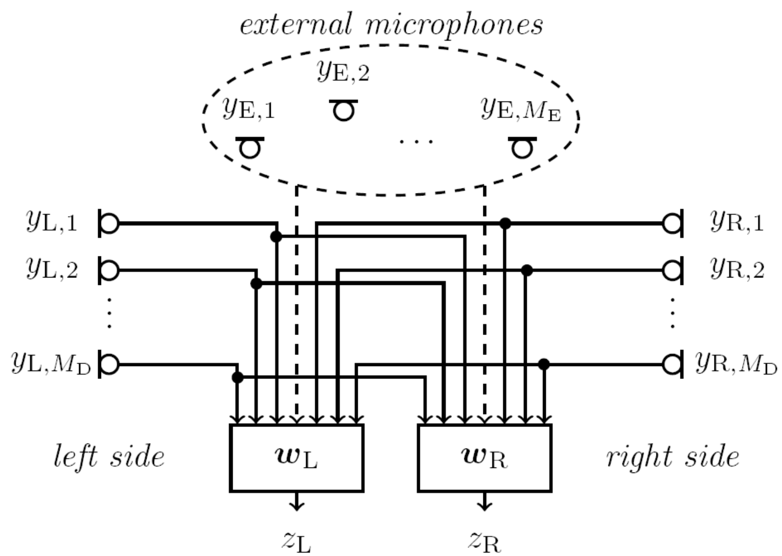
3. Acoustic sensor networks

External microphones

- Exploit the availability of one or more external microphones **(acoustic sensor network)** with hearing aids
  [Bertrand 2009, Szurley 2016, Yee 2018, Farmani 2018, Kates 2018, Ali 2019, Gößling 2019]

- Integrating external microphone(s) with hearing aid microphones may lead to:
  - Low-complexity method to **estimate relative transfer function (RTF)** vector of target speaker
  - Improved **noise reduction** and **binaural cue preservation** performance

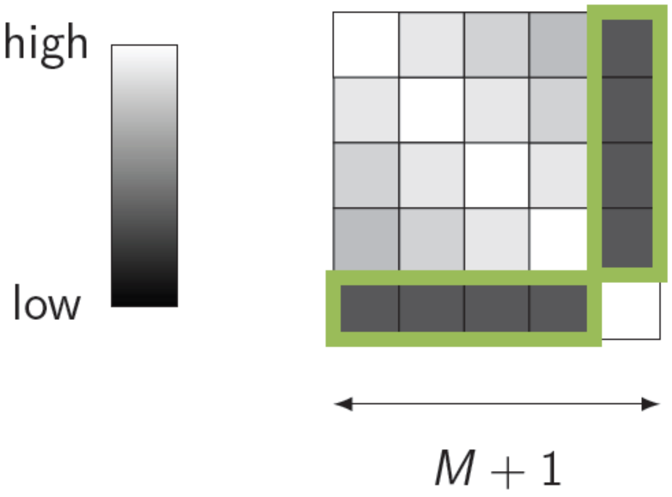
$$\mathbf{w}_L = \frac{\mathbf{R}_v^{-1}\mathbf{a}_L}{\mathbf{a}_L^H\mathbf{R}_v^{-1}\mathbf{a}_L}, \qquad \mathbf{w}_R = \frac{\mathbf{R}_v^{-1}\mathbf{a}_R}{\mathbf{a}_R^H\mathbf{R}_v^{-1}\mathbf{a}_R}$$



One external microphone: RTF estimation

- **Estimate RTF vector of target speaker** to steer binaural MVDR beamformer
- **Spatial coherence (SC) method:** assume that noise components in external microphone and HA microphones are uncorrelated, e.g., when external microphone is spatially separated from HA microphones + diffuse noise field

  → correlate HA microphone signals with external microphone signals and normalize by reference element
- **Low computational complexity** with similar (even better in practice) performance than state-of-the-art covariance whitening (CW) approach



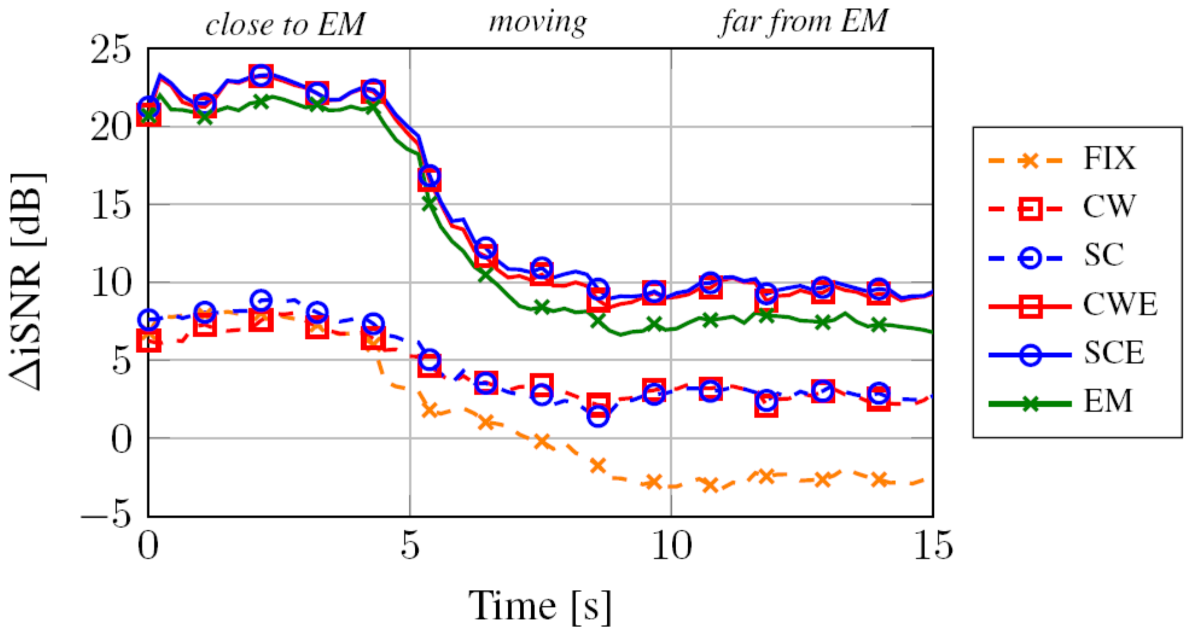
Low spatial coherence between noise components in local array signals and external microphone signal

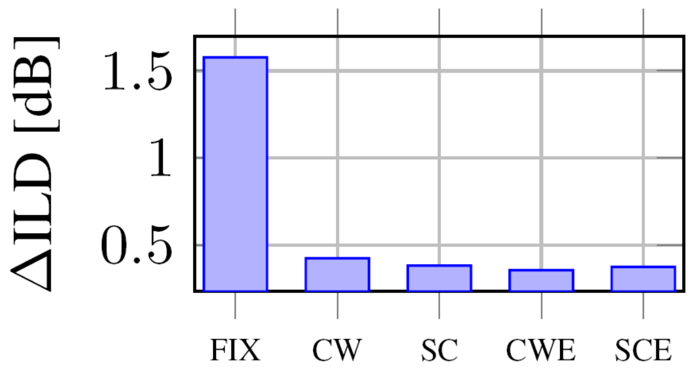
$$\bar{\mathbf{a}}_{\mathrm{L}}^{\mathrm{SCE}} = \frac{\bar{\mathbf{R}}_{\mathrm{y}} \mathbf{e}_{\mathrm{E}}}{\mathbf{e}_{\mathrm{L}}^{T} \bar{\mathbf{R}}_{\mathrm{y}} \mathbf{e}_{\mathrm{E}}}, \quad \bar{\mathbf{a}}_{\mathrm{R}}^{\mathrm{SCE}} = \frac{\bar{\mathbf{R}}_{\mathrm{y}} \mathbf{e}_{\mathrm{E}}}{\mathbf{e}_{\mathrm{R}}^{T} \bar{\mathbf{R}}_{\mathrm{y}} \mathbf{e}_{\mathrm{E}}}$$

$$\bar{\mathbf{w}}_{\mathrm{L}}^{\mathrm{SCE}} = \begin{bmatrix} \alpha \cdot [\mathbf{I}_{2M}, \; \mathbf{0}_{2M\times 1}] \, \bar{\mathbf{w}}_{\mathrm{L}} \\ \alpha(1+\beta) \cdot \mathbf{e}_{\mathrm{E}}^{T} \bar{\mathbf{w}}_{\mathrm{L}} \end{bmatrix}$$

real-valued bias

One external microphone: Simulation results

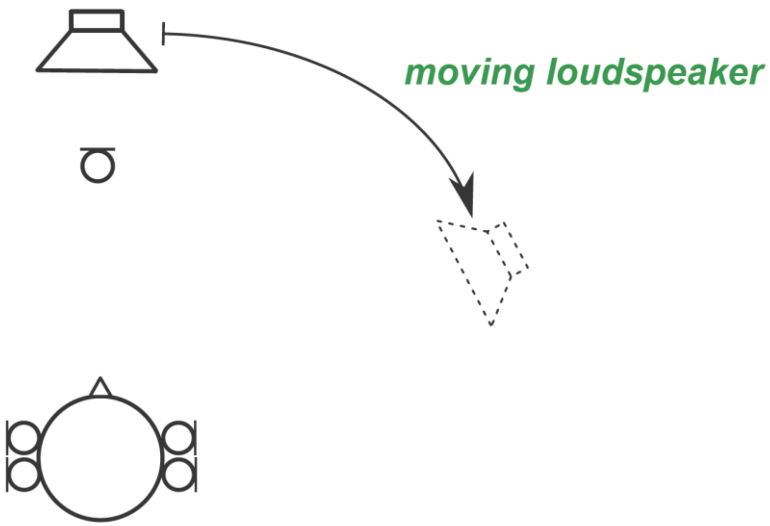


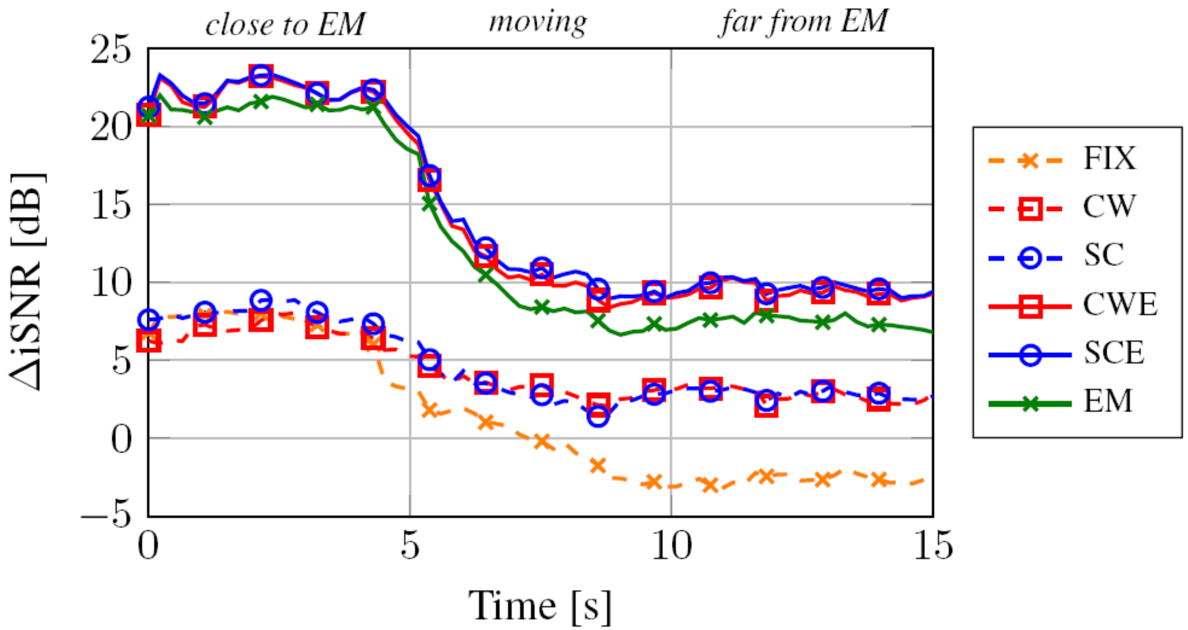


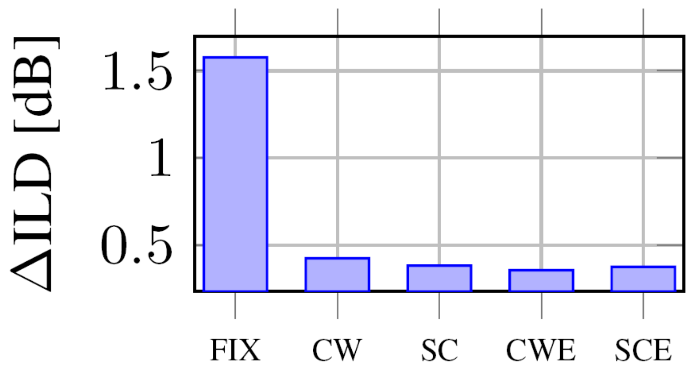
- MVDR with external microphone (SCE) leads to **better SNR** compared to MVDR using only HA microphones (SC,FIX) and external microphone (EM)
- MVDR using estimated RTFs (SCE, SC) **preserves binaural cues of target speaker** compared to fixed MVDR (FIX) and external microphone (EM)

Oldenburg Varechoic Lab ($T_{60} \approx$ 350ms), M=4 + 1 external mic (1.5m/0.5m), moving speaker, pseudo-diffuse babble noise, iSNR=0dB (right HA)
STFT: 32 ms, 50% overlap, sqrt-Hann; SPP in external microphone; smoothing: 100 ms (speech), 1 s (noise)

One external microphone: Simulation results







- MVDR with external microphone (SCE) leads to **better SNR** compared to MVDR using only HA microphones (SC,FIX) and external microphone (EM)
- MVDR using estimated RTFs (SCE, SC) **preserves binaural cues of target speaker** compared to fixed MVDR (FIX) and external microphone (EM)

Oldenburg Varechoic Lab ($T_{60} \approx$ 350ms), M=4 + 1 external mic (1.5m/0.5m), moving speaker, pseudo-diffuse babble noise, iSNR=0dB (right HA)
STFT: 32 ms, 50% overlap, sqrt-Hann; SPP in external microphone; smoothing: 100 ms (speech), 1 s (noise)

[Gößling, Doclo, *Proc. IWAENC* 2018] [Gößling, Doclo, *Proc. ICASSP 2019*]

Multiple external microphones

- Each external microphone yields (different) RTF estimate
- **Linear combination/selection** of RTF estimates (per frequency)

$$\boxed{\mathbf{a}_{\mathrm{L}}^{\mathrm{SC-C}} = \frac{\mathbf{A}_{\mathrm{L}}^{\mathrm{SC}}\mathbf{c}}{\mathbf{e}_{\mathrm{L}}^{T}\mathbf{A}_{\mathrm{L}}^{\mathrm{SC}}\mathbf{c}}}$$

1. Input SNR-based selection

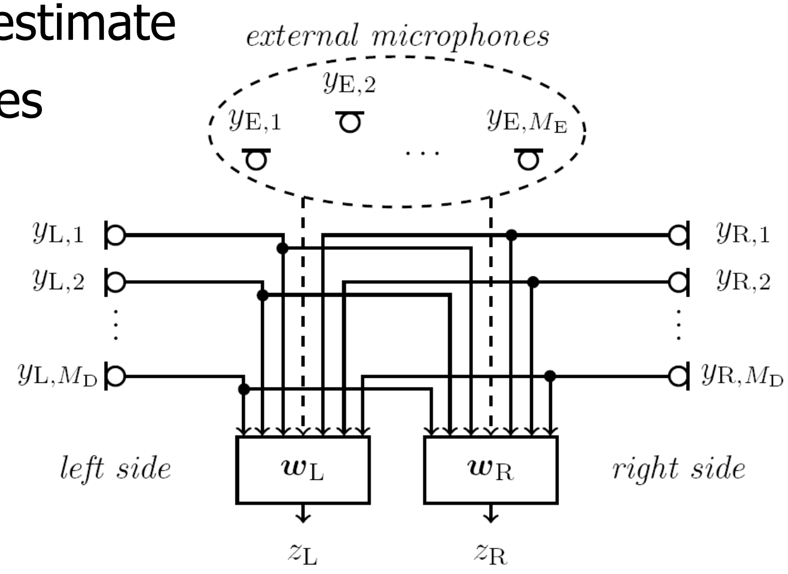
$$\mathbf{c}^{\mathrm{iSNR}} = \mathbf{e}_{\mathrm{E},\hat{i}}\,, \quad \hat{i} = \arg\max_{i} \frac{\mathbf{e}_{\mathrm{E},i}^{T}\mathbf{R}_{\mathrm{y}}\mathbf{e}_{\mathrm{E},i}}{\mathbf{e}_{\mathrm{E},i}^{T}\mathbf{R}_{\mathrm{n}}\mathbf{e}_{\mathrm{E},i}}$$

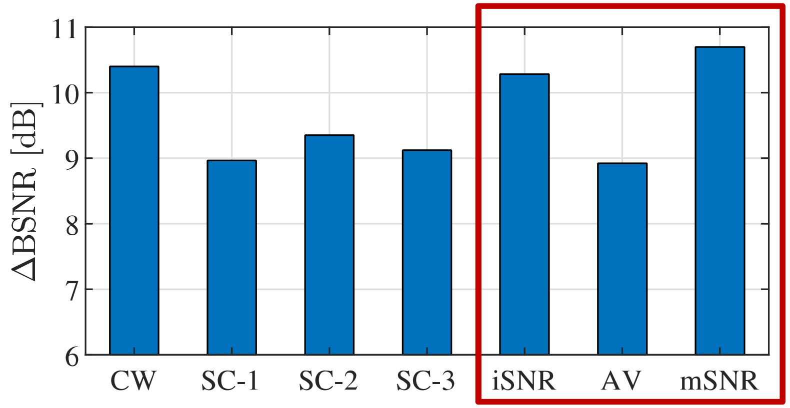
2. Simple averaging

$$\mathbf{c}^{\mathrm{AV}} = \left[\frac{1}{M_{\mathrm{E}}}, \ldots, \frac{1}{M_{\mathrm{E}}}\right]^{T}$$

3. Output SNR-maximizing combination

$$\boxed{\mathbf{c}^{\mathrm{mSNR}} = \arg\max_{\mathbf{c}} \ \mathrm{SNR}_{\mathrm{BMVDR,L}}^{\mathrm{out}} = \mathcal{P}\{\mathbf{\Lambda}_2^{-1}\mathbf{\Lambda}_1\}}$$

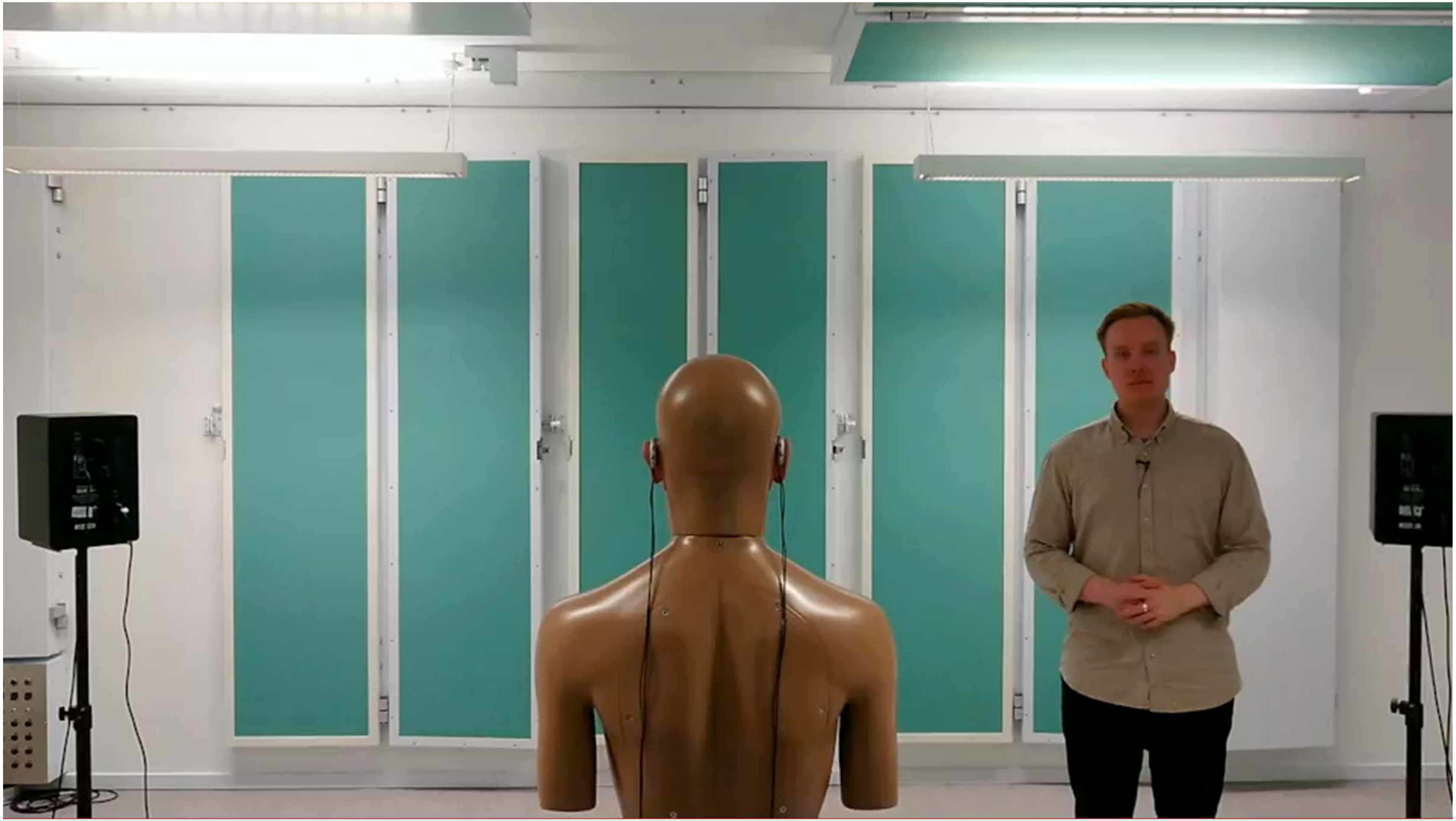




Audio Demo

- Real-world recordings ($T_{60} \approx 300$ ms), **moving speaker**
- KEMAR with **two BTE hearing aids** (2 mics each) and **one external mic**
- Pseudo-diffuse babble noise

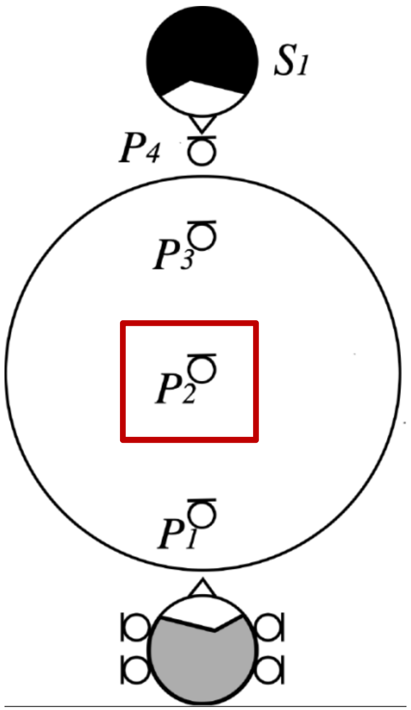
Audio Demo



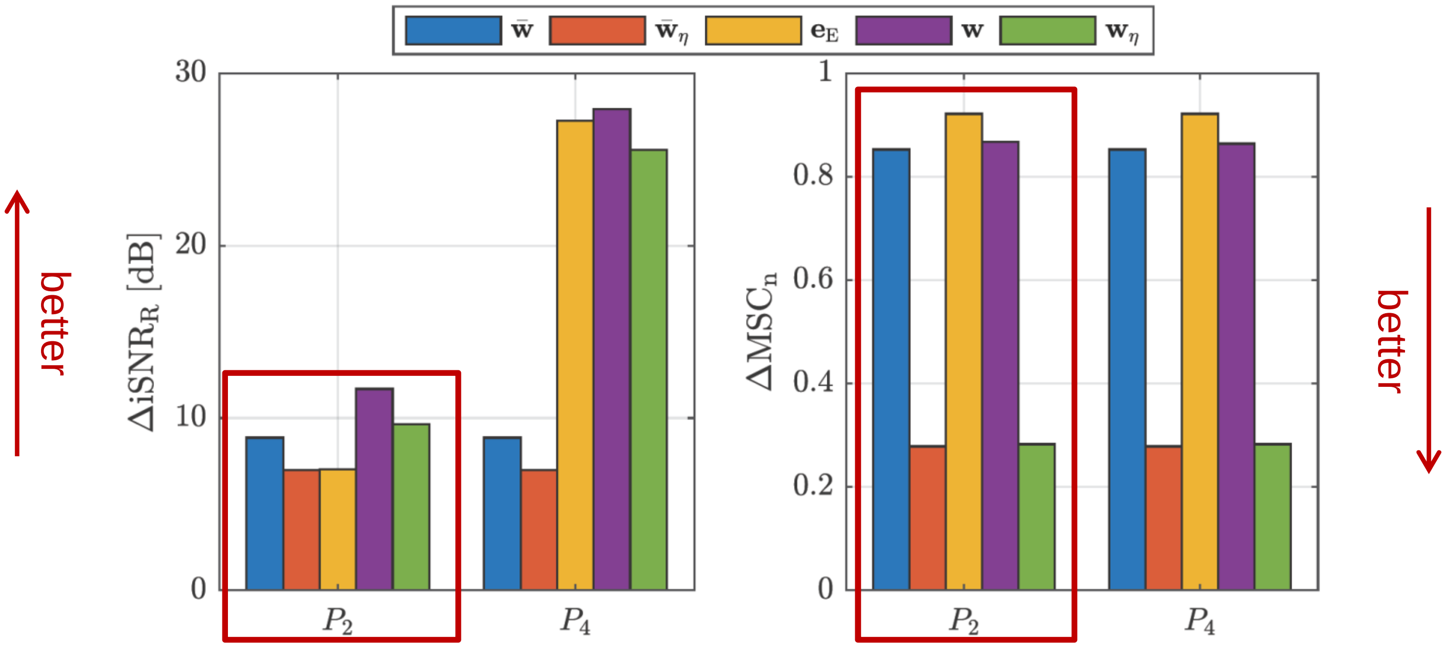
CARL VON OSSIETZKY universität OLDENBURG
DEPARTMENT OF MEDICAL PHYSICS AND ACOUSTICS
SIGNAL PROCESSING GROUP

Binaural MVDR-N beamformer

- Including external microphone in **binaural MVDR-N beamformer** leads to:
  - **Larger output SNR** for same trade-off parameter η
  - Same output SNR with larger trade-off parameter η → **better cue preservation**



MVDR (HA) **MVDR-N (HA)** **External** **eMVDR (HA+ext)** **eMVDR-N (HA+ext)**

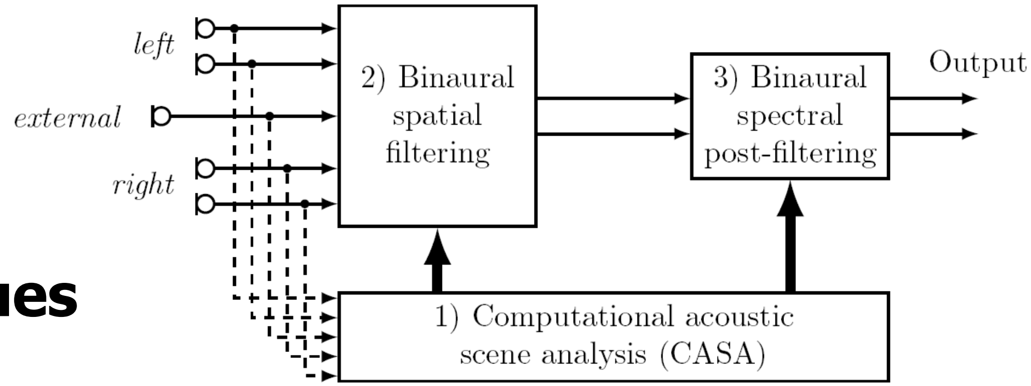


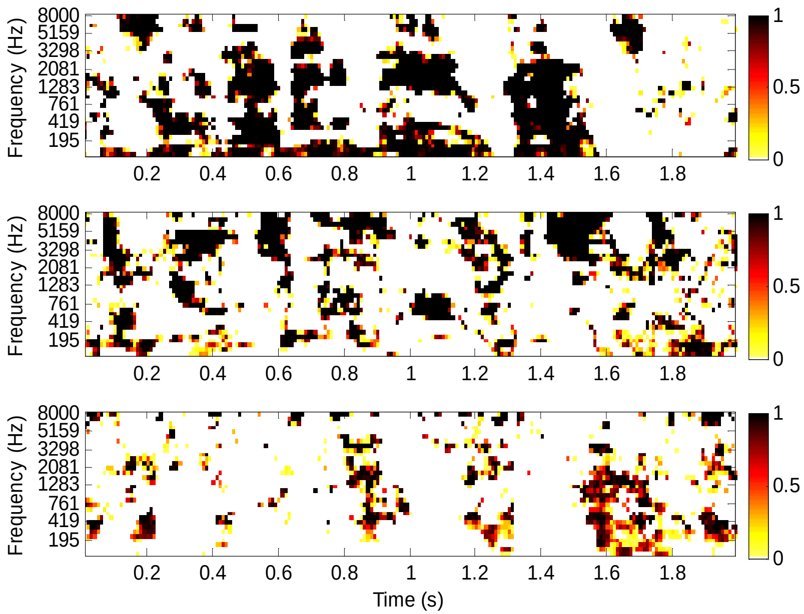
Starkey database with real-world recordings ($T_{60} \approx$ 620ms), M=4, target speaker S_1, multi-talker babble noise, 0 dB input iSNR (right hearing aid)
MVDR: perfectly estimated noise correlation matrix, RTF of target speaker estimated using covariance whitening method

[Gößling, Doclo, *Proc. HSCMA* 2017] [Gößling, Doclo, *submitted to IEEE/ACM TASLP*]

Current/future work

- **Performance analysis** for different acoustic scenarios (interfering speakers)
- **Synchronization/latency issues**
- **Complex and time-varying scenarios**: incorporate computational acoustic scene analysis (CASA) into control path of developed algorithms
- **Subjective evaluation** of binaural speech enhancement algorithms with **HA/CI users** ongoing





Conclusions

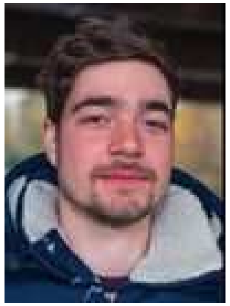
- **Speech communication applications:** on-line speech enhancement algorithms for dynamic acoustic scenarios required
- **Joint noise reduction and dereverberation** using multiple microphones:
  - MVDR beamformer + spectral postfiltering: estimates of time-varying spatial and spectral variables (RETF vector, PSDs)
  - Reverberation suppression: multi-channel linear prediction
- **Binaural hearing devices** with binaural output signals:
  - Extensions of binaural MVDR/MWF enable to improve speech intelligibility while preserving spatial awareness (binaural cues)
  - Improved performance when integrating external microphones (acoustic sensor networks)

Acknowledgments


Dr. Ina Kodrasi


Dr. Ante Jukić

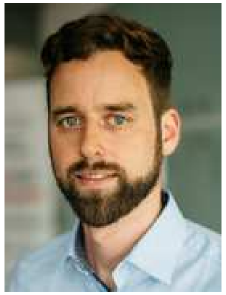

Dr. Daniel Marquardt


Marvin Tammen


Jonas Klug


Nico Gößling


Wiebke Middelberg


Prof. Timo Gerkmann


Prof. Sharon Gannot

- **Funding:**
  - Cluster of Excellence Hearing4all (DFG), Research Unit Individualized Hearing Acoustics (DFG)
  - Marie-Curie Initial Training Network "Dereverberation and Reverberation of Audio, Music, and Speech" (EU)
  - Joint Lower-Saxony Israel Project "Acoustic scene aware speech enhancement for binaural hearing aids" (Partner: Bar-Ilan University, Israel)
  - German-Israeli Foundation Project "Signal Dereverberation Algorithms for Next-Generation Binaural Hearing Aids" (Partners: International Audiolabs Erlangen; Bar-Ilan University, Israel)

DREAMS Dereverberation and Reverberation of Audio, Music, and Speech

DFG Deutsche Forschungsgemeinschaft

VolkswagenStiftung

GIF German-Israeli Foundation for Scientific Research and Development

Questions ?

House of Hearing, Oldenburg