Joint Estimation of RETFs and PSDs for Multi-Channel Speech Enhancement

Marvin Tammen [1], Dr. Ina Kodrasi [2], Prof. Dr. Simon Doclo [1]

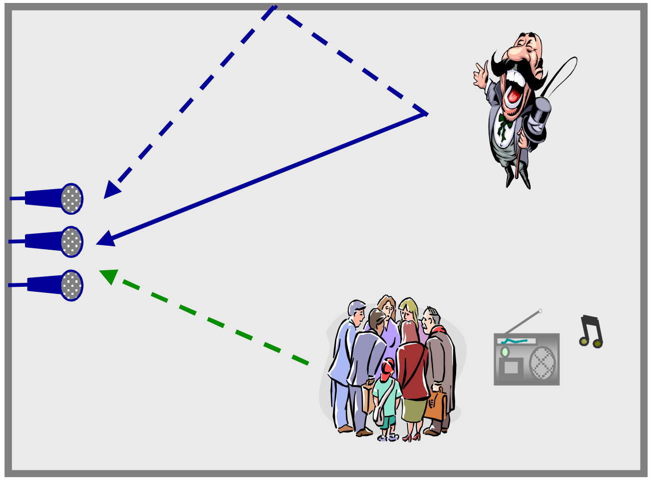
[1] Dept. of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, University of Oldenburg, Germany

[2] Speech and Audio Processing Group, Idiap Research Institute, Switzerland

International Congress on Acoustics, Aachen, 09.09.2019

Introduction

- **Problem**
  - Reverberation and ambient noise jointly present in typical acoustic environments
  - Speech quality and intelligibility degradation
  - Performance degradation of ASR systems



- **Objective**
  - Develop (single- and) **multi-channel joint dereverberation and noise reduction** algorithms
  - Exploit knowledge or (statistical) models of speech signals and room acoustics

- **This presentation:**
  - **Multi-channel Wiener filter (MWF)** = MVDR beamformer + spectral postfilter

  → requires estimate of **relative transfer function vector** of target source and **PSDs** of target source and interference (reverberation and noise)

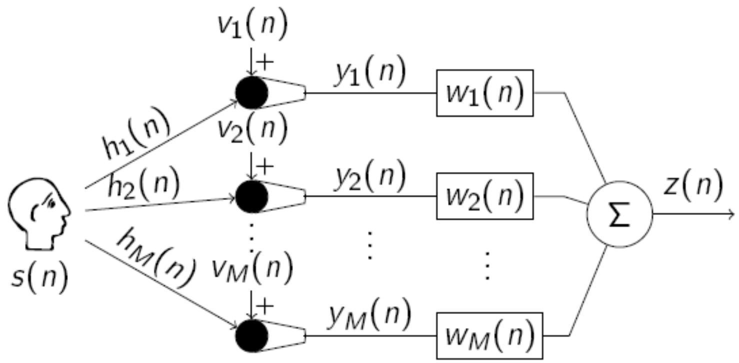
Signal model

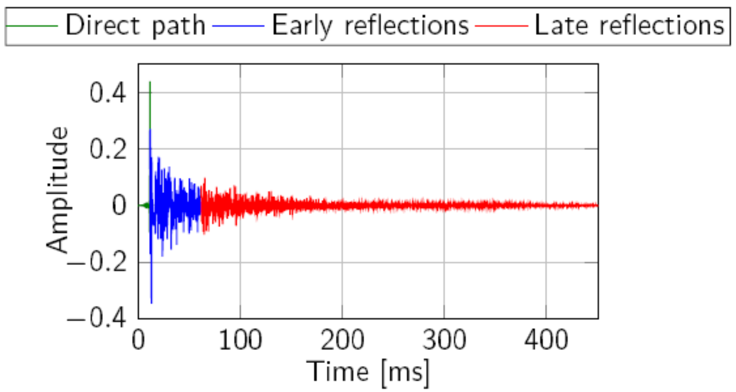
- **Scenario:** speech source in noisy and reverberant environment, *M* microphones
- **Model in Short-Time Fourier Transform (STFT) domain:**

$$\begin{aligned} y_m(k,l) &= h_m(k,l) \star s(k,l) + v_m(k,l) \\ &= \underbrace{a_m(k,l)s(k,l)}_{\text{direct and early reverberation}} + \underbrace{x_{r,m}(k,l)}_{\text{late reverberation}} + \underbrace{v_m(k,l)}_{\text{ambient noise}} \end{aligned}$$

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

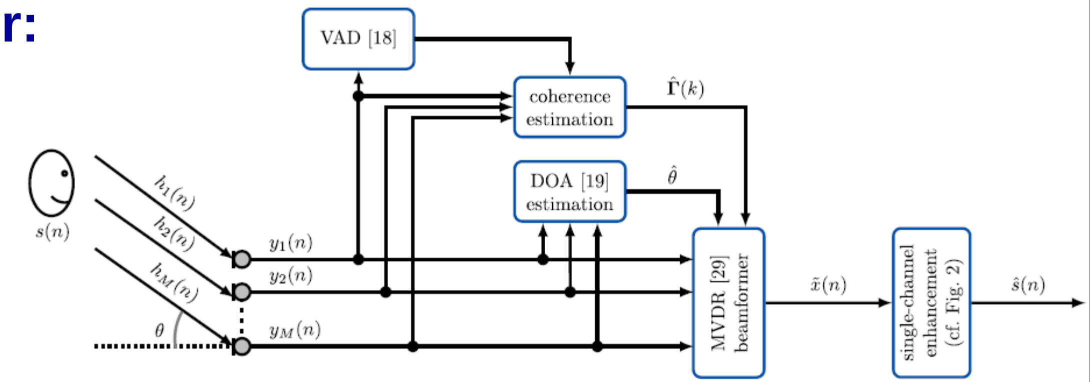
a(*k,l*) = vector of **relative early transfer functions (RETFs)** of target source





Multi-microphone dereverberation and noise reduction

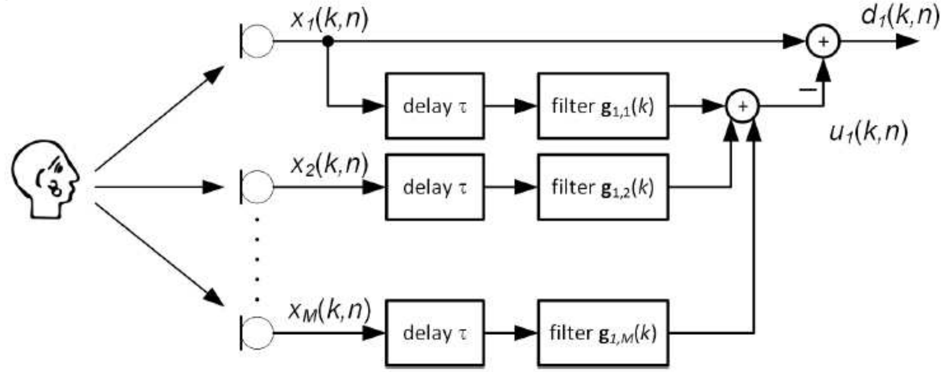
1. **Beamforming + spectral postfilter:** *multiply* each time-frequency bin with *real-valued* gain



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

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

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

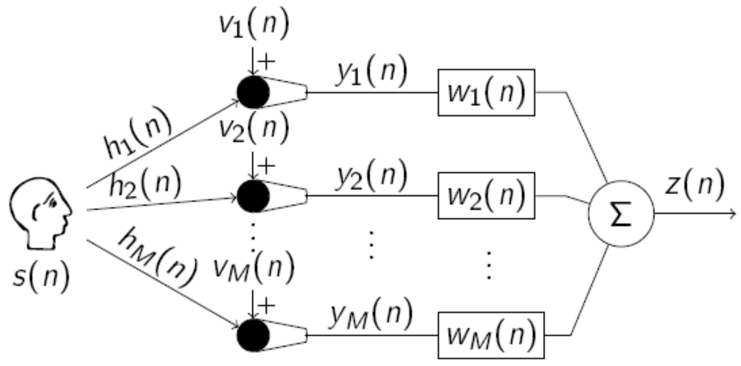


$$\boxed{y_m(l) = h_m(l) \star s(l) + v_m(l)}$$

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

Beamforming + spectral postfilter

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



[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

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

- **"Workhorse algorithm": 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) or voice activity detection (VAD)

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

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

  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 !)

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

- **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:** *estimate* (requiring VAD/SPP) or *model* as spatially white noise $\mathbf{\Phi}_v(l) = \phi_v(l)\mathbf{I}$

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

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

     - **Eigenvalues: estimate of PSDs**

$$\hat{\phi}_x(l) = \lambda_1\{\hat{\mathbf{\Phi}}_x^w(l)\}$$

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

     - **Principal eigenvector: estimate of RETF vector**

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

[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)

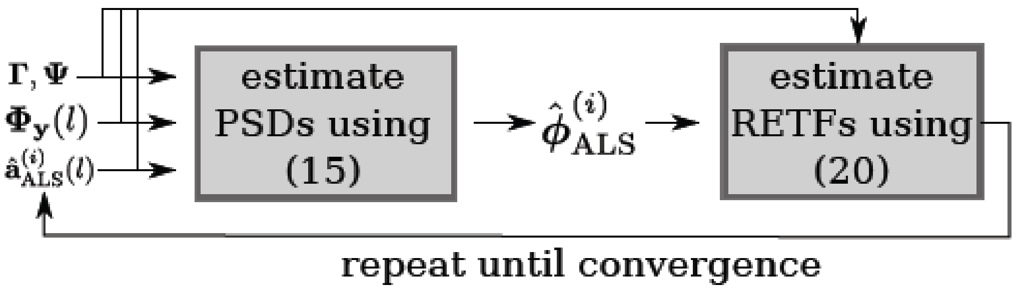


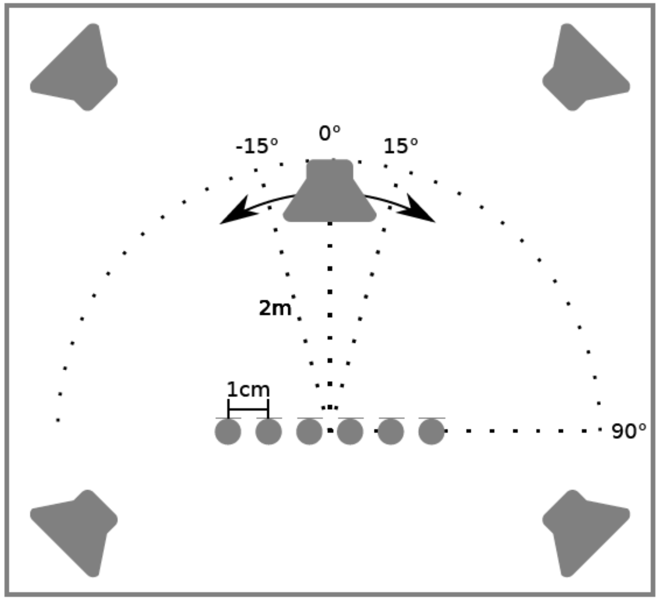
Fig. 1: Block diagram of ALS-based RETF vector and PSD estimation.

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)
    $\hat{\mathbf{a}}^{(i)}_{\text{ALS}}(k,l)=\sqrt{\hat{\lambda}^{(i)}_1(k,l)/\hat{\phi}^{(i)}_{s,\text{ALS}}(k,l)}\hat{\boldsymbol{\nu}}^{(i)}_1(k,l)$ (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)

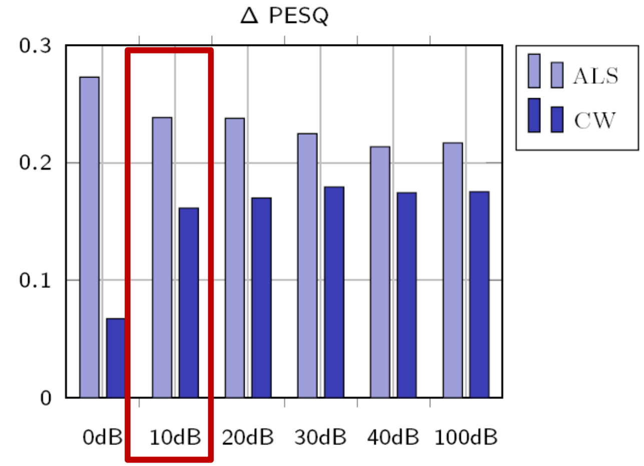


Simulation parameters:

- 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

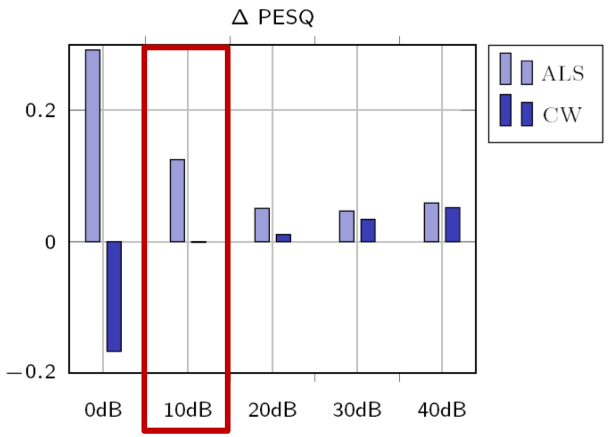
Simulation results (PESQ improvement)

1. Simulated stationary source



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

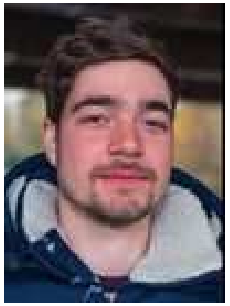


Linear array (M=6, d=1cm), fs=16kHz, moving source θ=0º to θ=90º pseudo-diffuse babble noise (SDR=10dB), sensor noise (DNR=10dB)

Conclusions

- **Multi-microphone noise reduction and dereverberation:**
  - Multi-channel Wiener filter = MVDR beamformer + spectral postfilter
  - Requires **estimates of (time-varying) RETF vector of target source and PSDs of target source and interference** (reverberation, noise)
  - Assumption: model reverberation as diffuse sound field
- **Joint RETF and PSD estimation:**
  - **Covariance whitening**: eigenvalue decomposition of prewhitened signal correlation matrix, requires estimate of noise covariance matrix
  - **Alternating least squares:** minimize Frobenius norm of error covariance matrix; does not require estimate of noise covariance matrix
- **Simulation results show good performance both for stationary as well as for challenging dynamic scenarios**

Acknowledgments



Marvin Tammen

Dr. Ina Kodrasi

- **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)
  - 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