Cognitive-Driven Binaural Speech Enhancement System for Hearing Aid Applications

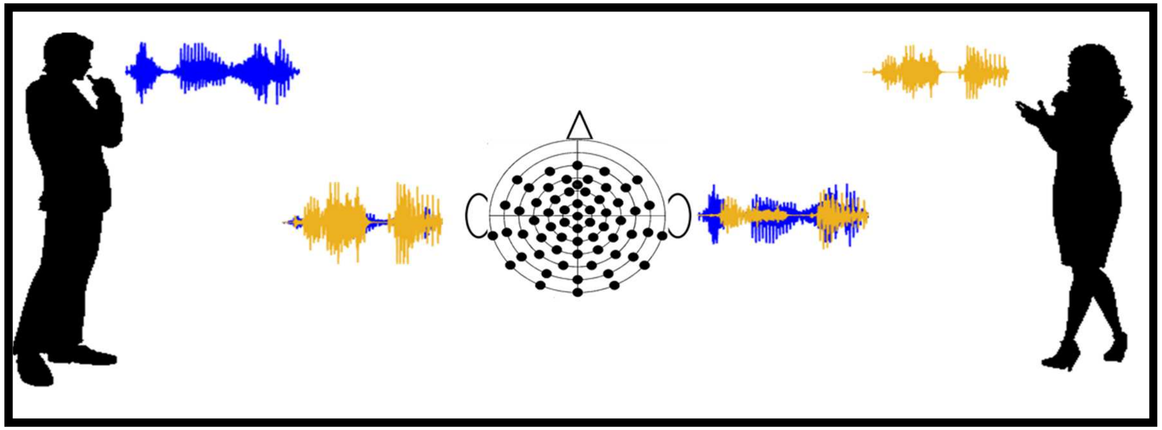
Ali Aroudi[1], Daniel Marquardt[1,2], Simon Doclo[1]

[1] Dept. of Medical Physics and Acoustics and Cluster of Excellence Hearing4All, University of Oldenburg, Germany

[2] Starkey Hearing Technologies, USA

IHCON 2018

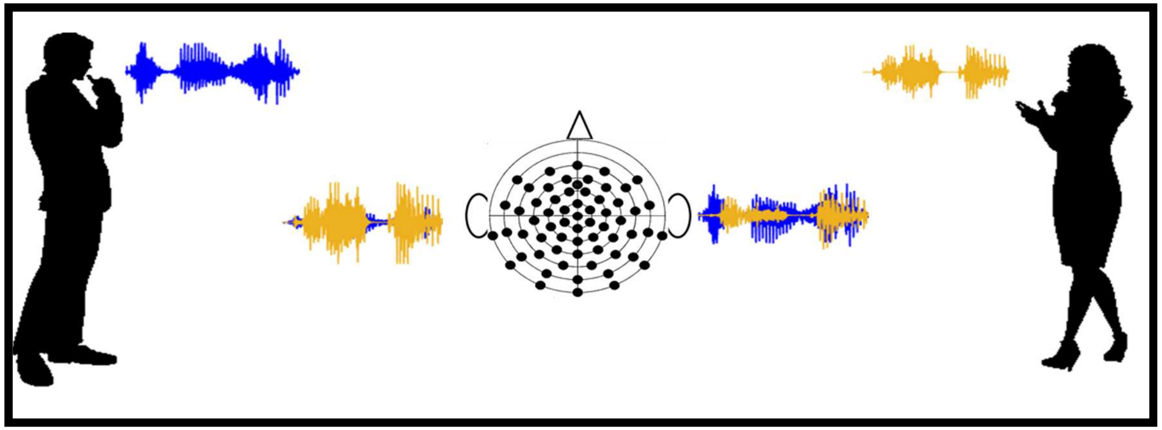
Problem statement



- Identify **target speaker to be enhanced** using hearing aid microphone signals (acoustic scene analysis) + EEG signals
- **Auditory attention decoding (AAD):** least-squares-based method [1]
- Often studied for **anechoic noiseless acoustic conditions** ✗
- Reference signals for decoding: **clean speech signals of speakers** are not available ✗

[1] J. A. O'Sullivan *et al.*, Attentional selection in a cocktail party environment can be decoded from single-trial EEG, Cerebral Cortex, 2014

Problem statement



Goal

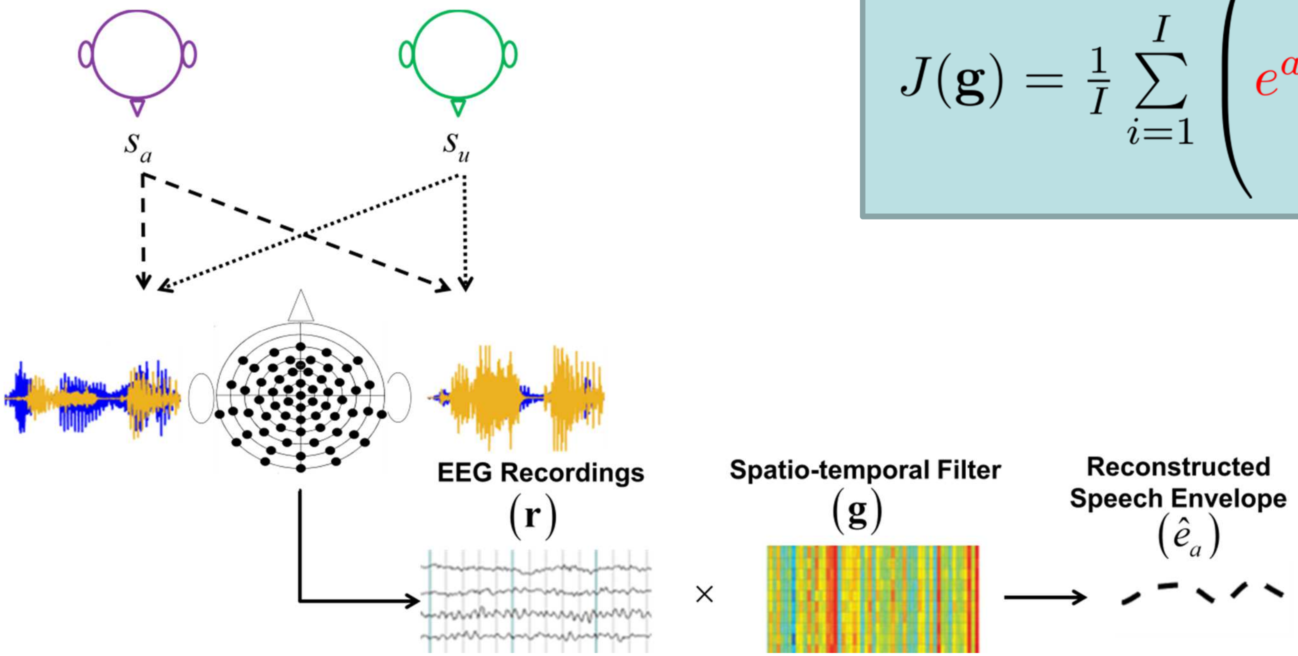
- Investigate **impact of different acoustic conditions** (reverberation + background noise) on AAD filter training and decoding performance
- Generate appropriate **reference signals** from hearing aid microphone signals (here: LCMV beamformer)
- Enhance target speaker while preserving spatial impression of acoustic scene based on AAD → **cognitive-driven binaural speech enhancement**

Impact of different acoustic conditions on AAD

Auditory attention decoding method [1]

- **Training step**: compute spatio-temporal filter **g**

$$J(\mathbf{g}) = \frac{1}{I}\sum_{i=1}^{I}\left(e^{a}[i] - \underbrace{\mathbf{g}^{T}\mathbf{r}[i]}_{\hat{e}_a[i]}\right)^{2} + \beta\mathbf{g}^{T}\mathbf{D}\mathbf{g}$$

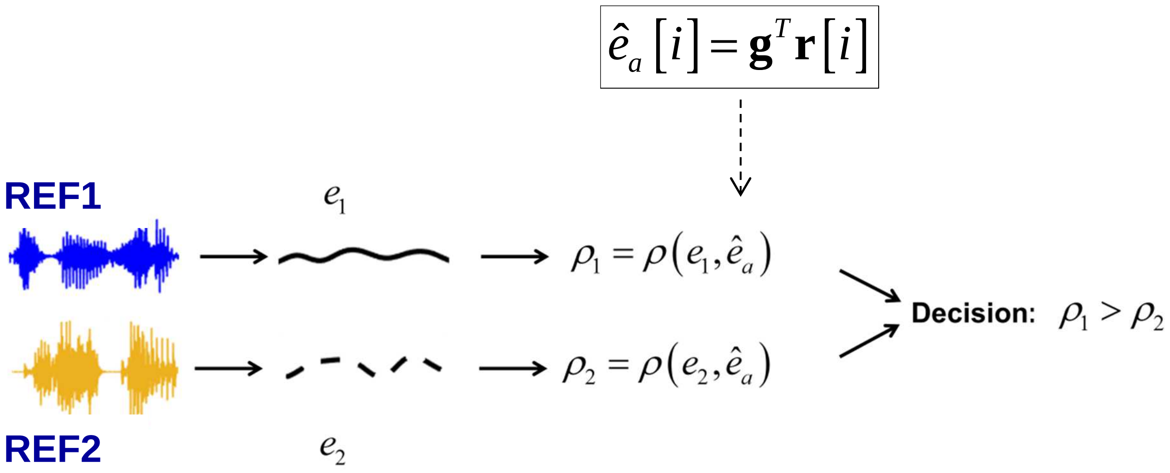


- Reconstruct envelope of attended speech signal (*training signal*) by filtering and combining EEG signals
- Regularization to avoid over-fitting

[1] J. A. O'Sullivan *et al.*, Attentional selection in a cocktail party environment can be decoded from single-trial EEG, Cerebral Cortex, 2014

Auditory attention decoding method [1]

- **Decoding step:** correlate envelope of estimated attended speech signal with envelopes of *reference signals*



[1] J. A. O'Sullivan *et al.*, Attentional selection in a cocktail party environment can be decoded from single-trial EEG, Cerebral Cortex, 2014

EEG condition and signals for filter training and decoding

Training step	Decoding step
$J(\mathbf{g}) = \frac{1}{I}\sum_{i=1}^{I}\left(e^a[i] - \underbrace{\mathbf{g}^T\mathbf{r}[i]}_{\hat{e}_a[i]}\right)^2 + \beta\mathbf{g}^T\mathbf{D}\mathbf{g}$	$\hat{e}_a[i] = \mathbf{g}^T\mathbf{r}[i]$ $\rho_1 = \rho(e_1, \hat{e}_a),\ \rho_2 = \rho(e_2, \hat{e}_a)$

EEG training condition

EEG evaluation condition

EEG training condition	EEG evaluation condition
Anechoic + **Noiseless**	Anechoic + **Noiseless**
Reverberant + Noiseless	**Reverberant** + Noiseless
Anechoic + **Noisy**	Anechoic + **Noisy**
Reverberant + Noisy	**Reverberant + Noisy**
All conditions	**All** conditions

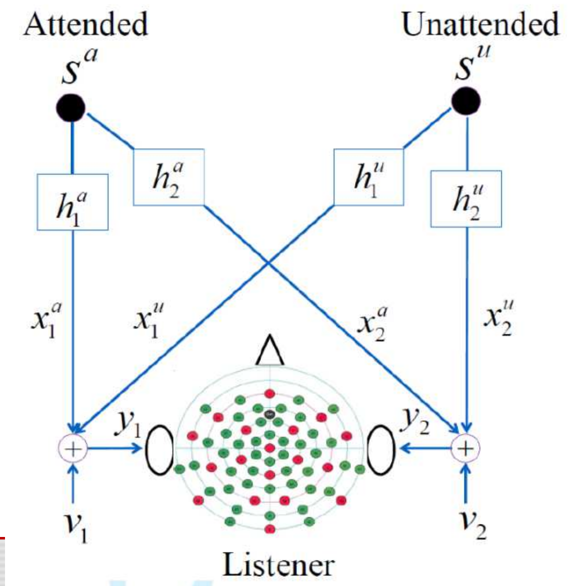
EEG condition and signals for filter training and decoding

Training step	Decoding step
$J(\mathbf{g}) = \frac{1}{I}\sum_{i=1}^{I}\Big(e^a[i] - \underbrace{\mathbf{g}^T\mathbf{r}[i]}_{\hat{e}_a[i]}\Big)^2 + \beta\mathbf{g}^T\mathbf{D}\mathbf{g}$	$\hat{e}_a[i] = \mathbf{g}^T\mathbf{r}[i]$ $\rho_1 = \rho\left(e_1, \hat{e}_a\right),\ \rho_2 = \rho\left(e_2, \hat{e}_a\right)$

training signal

reference signals

- Different **acoustic signals** as reference (and training) signals:
  - *clean* speech signals
  - *anechoic* speech signals (HRTFs)
  - anechoic speech signals affected by
    - noise → *noisy* signals
    - reverberation → *reverberant* signals
    - interfering speaker → *interfered* signals
    - all acoustic components → *binaural* signals



Impact of different acoustic conditions on AAD

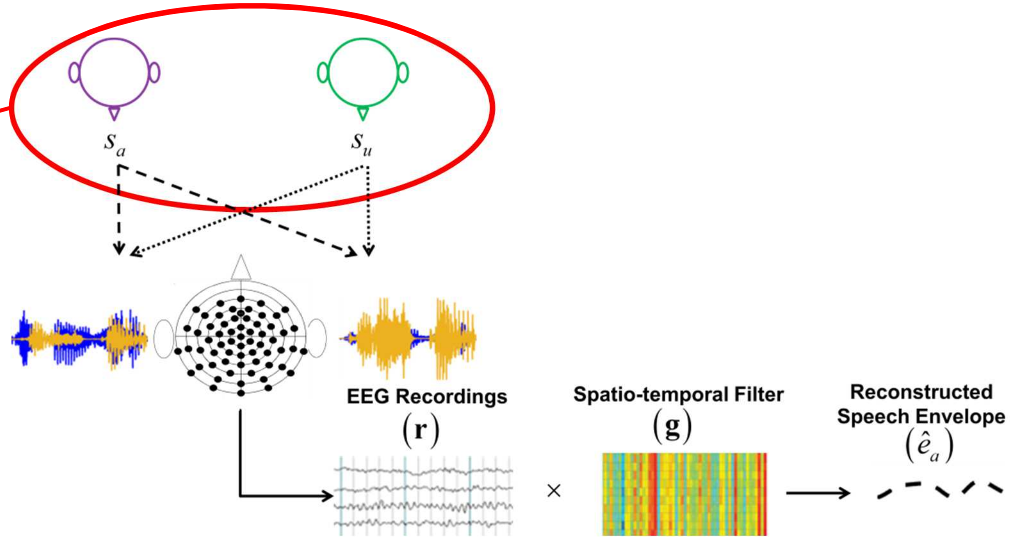
Experimental evaluation

Acoustic setup and simulation

Two audio stories by two different male speakers (German)

Left and right speaker simulated at -45° and 45°

Acoustic stimuli presented to participants using insert earphones



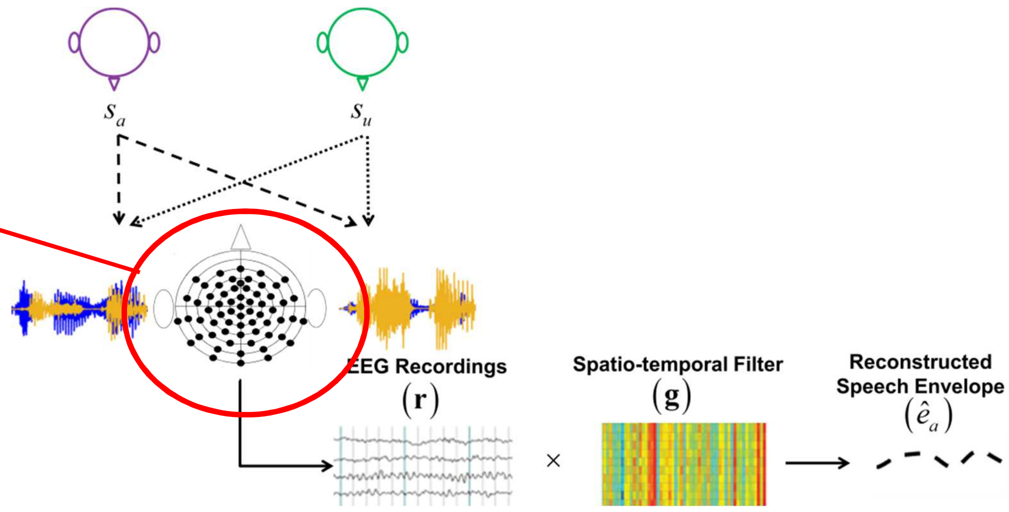
Experimental Analysis Condition	Stimuli Presentation	SNR[dB]	T_{60}[s]
Noiseless	Noiseless	∞	<0.05
Reverberant	Reverberant I	∞	0.50
	Reverberant II	∞	1.00
Noisy	Noisy I	9.0	<0.05
	Noisy II	4.0	<0.05
Reverberant-noisy	Reverberant-noisy I	9.0	0.50
	Reverberant-noisy II	4.0	0.50
	Reverberant-noisy III	9.0	1.00

diffuse babble noise

[2] A. Aroudi, B. Mirkovic, M. De Vos, S. Doclo, *IEEE Trans. Neural Systems and Rehabilitation Engineering, under revision.*

EEG setup, training and decoding

- **Subjects:**
  - *N*=18 German-speaking participants
  - 8 instructed to attend to left speaker, 10 instructed to attend to right speaker
- **EEG signals:**
  - 64 channels (Easycap GmbH)
  - band-pass filtered (2-8 Hz), f_s = 64 Hz
- **Training and decoding:**
  - trial length: **60 seconds**
  - each participant's own data
- **Decoding performance:**
  - percentage of correctly decoded trials over all considered trials and participants
  - leave-one-out cross validation approach

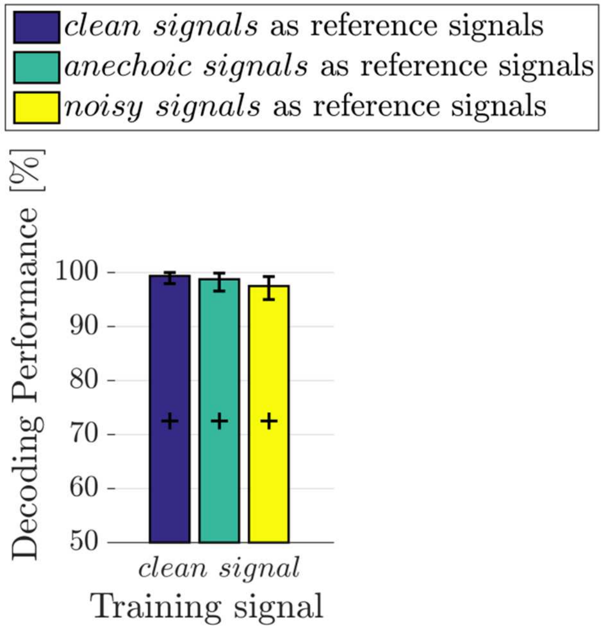


[2] A. Aroudi, B. Mirkovic, M. De Vos, S. Doclo, *IEEE Trans. Neural Systems and Rehabilitation Engineering, under revision.*

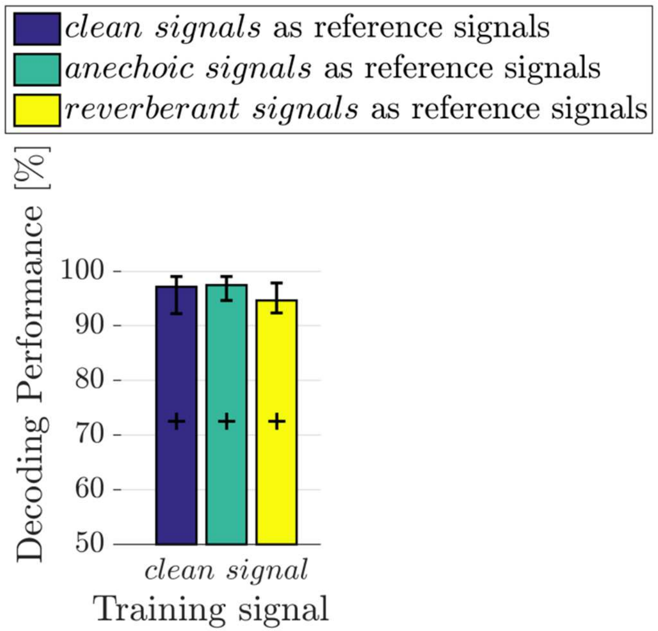
Experimental results

Reference signals: influence of noise, reverberation and interfering speaker

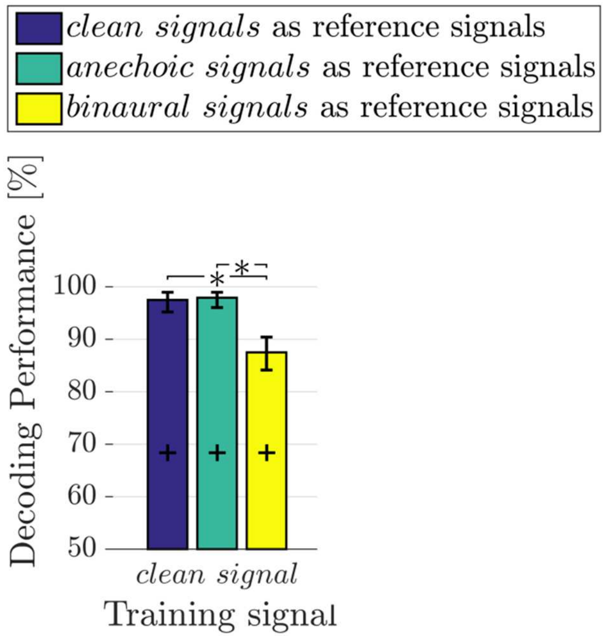
Anechoic - Noisy



Reverberant - Noiseless



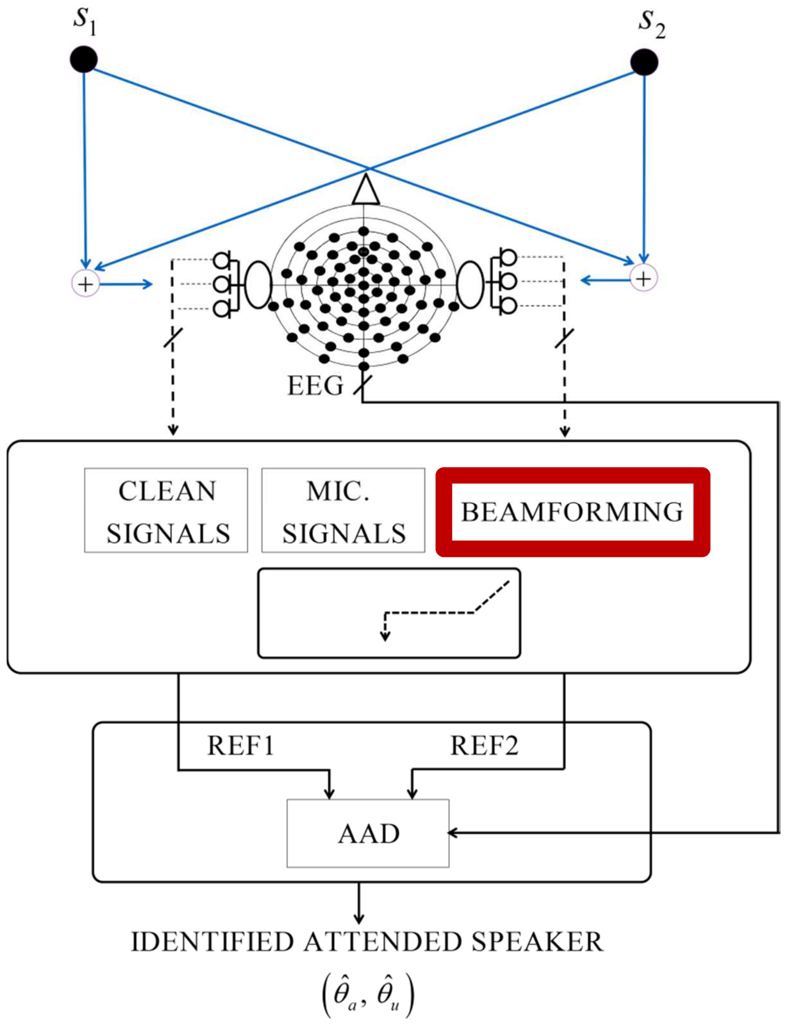
Reverberant - Noisy



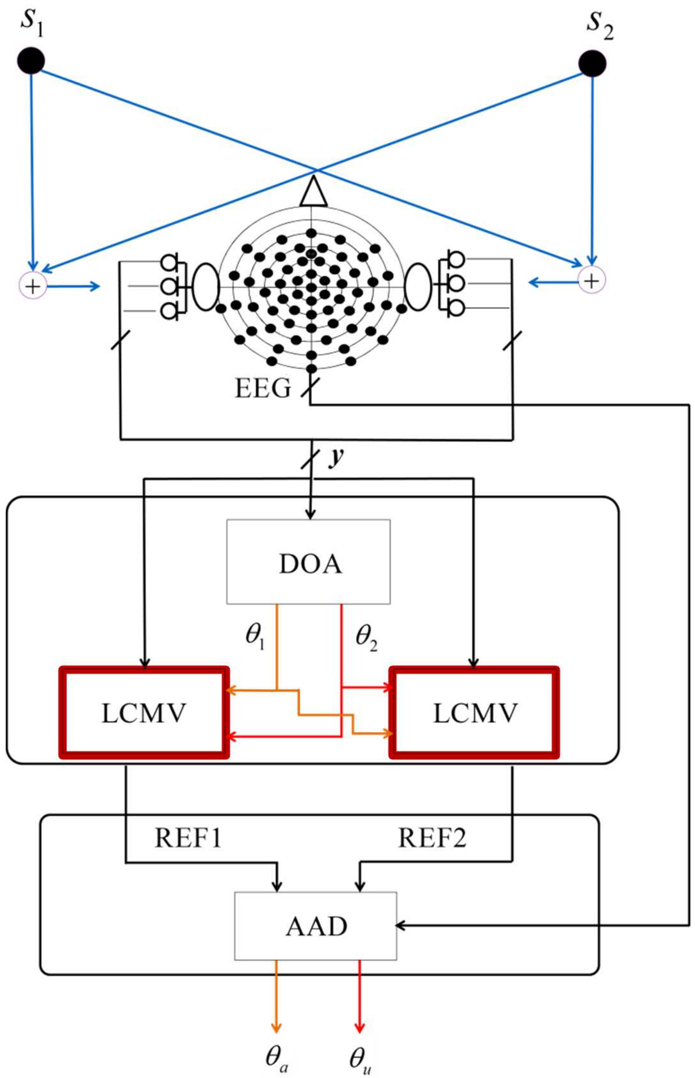
- ❑ Reference signals affected by **reverberation or noise** → comparable decoding performance as when using clean reference signals
- ❑ Reference signals affected by **interfering speaker** → decoding performance significantly decreases

Cognitive-driven binaural speech enhancement system

Beamformer output signals as reference signals



Beamformer output signals as reference signals

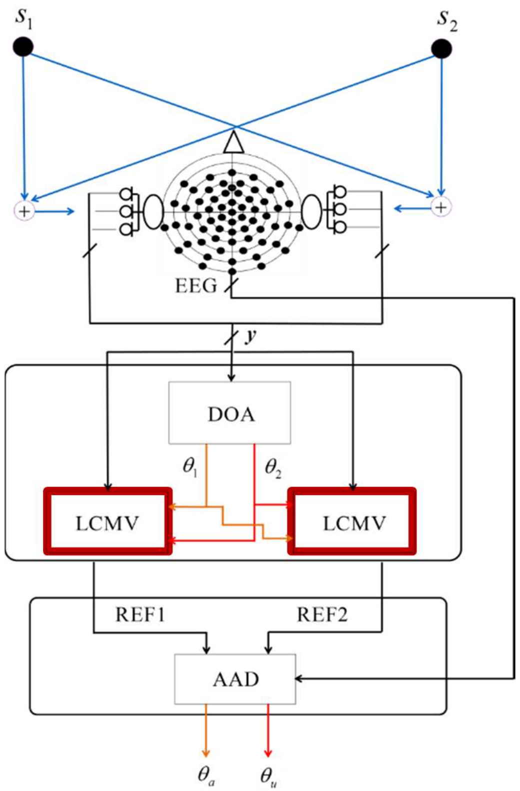


Beamformer output signals as reference signals

- **Linearly Constrained Minimum Variance (LCMV) beamformer [3]** aims at
  1. minimizing (diffuse) noise output PSD
  2. passing signals from *target direction* θ_t without distortion
  3. suppressing signals from *interference direction* θ_i with suppression factor δ

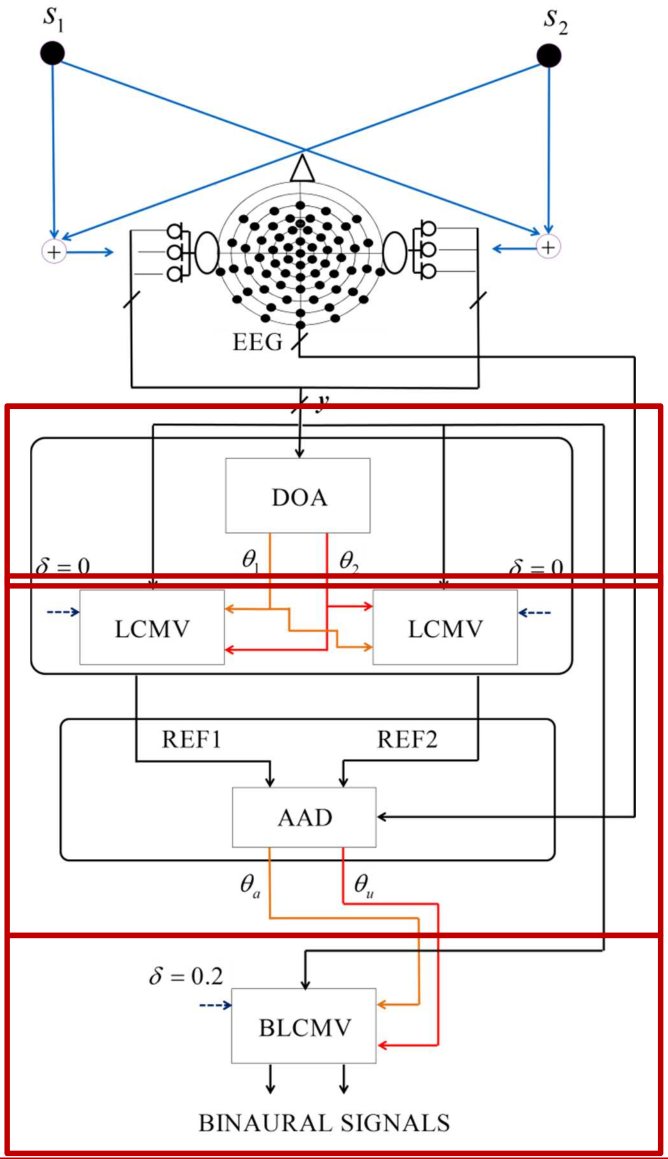
$$\min_{\mathbf{w}} \underbrace{\mathbf{w}^H \mathbf{\Phi}_n \mathbf{w}}_{\text{noise PSD}} \quad \text{subject to} \quad \underbrace{\mathbf{w}^H \mathbf{a}(\theta_t) = 1}_{\text{target}}, \underbrace{\mathbf{w}^H \mathbf{a}(\theta_i) = \delta}_{\text{interference}}$$

- **Requires**
  - *Noise covariance matrix*, e.g., diffuse noise assumption
  - Relative transfer functions (RTFs) of sources:
    - *Reverberant RTFs* (oracle/estimate)
    - *Anechoic RTFs* based on HRTF measurements and direction-of-arrival (DOA) of target and interfering speaker (oracle/estimate)



[3] E. Hadad, S. Doclo, S. Gannot, *The Binaural LCMV Beamformer and its Performance Analysis*, *IEEE/ACM TASLP, 2016.*

Cognitive-driven binaural speech enhancement system



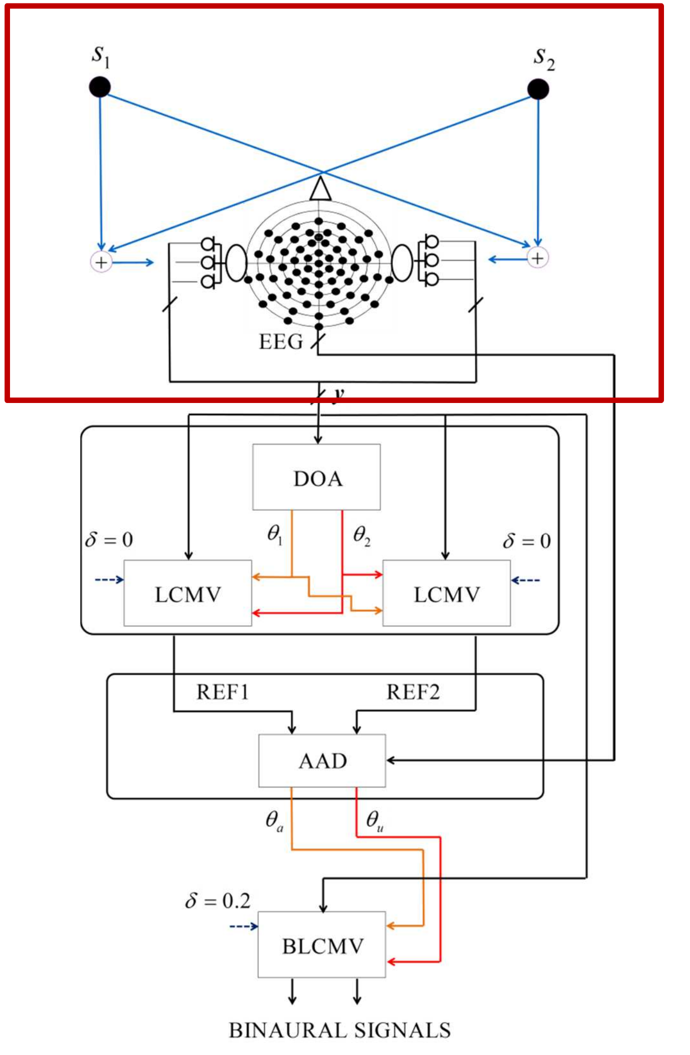
1. **Acoustic scene analysis:** DOA of speakers
2. **AAD using LCMV beamformer output signals** (steered to both speakers) decides which speaker is attended/unattended
3. AAD information is used in **binaural LCMV beamformer** to:
   - Pass (estimated) attended speaker
   - Suppress (estimated) unattended speaker with factor δ=0.2
   - Preserve binaural cues of both speakers

Cognitive-driven binaural speech enhancement system

Experimental evaluation

Acoustic setup and simulation

Experimental Analysis Condition	Stimuli Presentation	SINR [dB]	T_{60} [s]
Anechoic + Noisy	Noisy I	-1.00	<0.05
	Noisy II	-2.50	<0.05
Reverberant + Noisy	Reverberant-noisy I	-1.00	0.50
	Reverberant-noisy II	-2.50	0.50

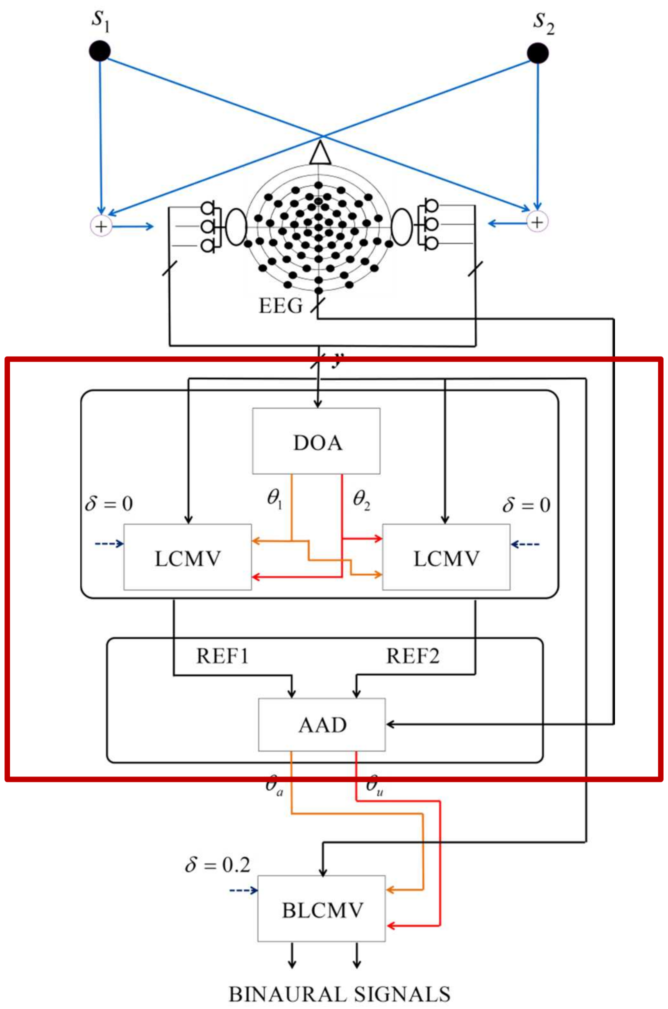


AAD using LCMV output signals as reference signals

- **DOA estimation of speakers**
  - oracle DOA (ODOA)
  - estimated DOA (EDOA) from binaural microphone signals with SVM-based method using GCC-PHAT features [4]

- **LCMV beamformer**
  - Noise covariance matrix: diffuse noise assumption
  - Relative transfer functions:
    - oracle reverberant RTFs (ORTF)
    - anechoic RTFs using oracle DOA (ODOA-RTF)
    - anechoic RTFs using estimated DOA (EDOA-RTF)

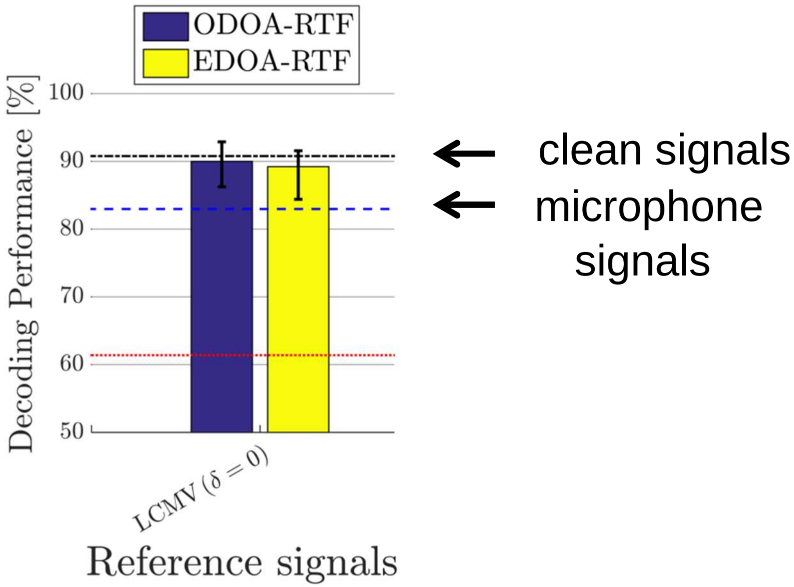
- **Auditory attention decoding**
  - trial length: **30 seconds**
  - oracle AAD (OAAD) or estimated AAD (AAD)



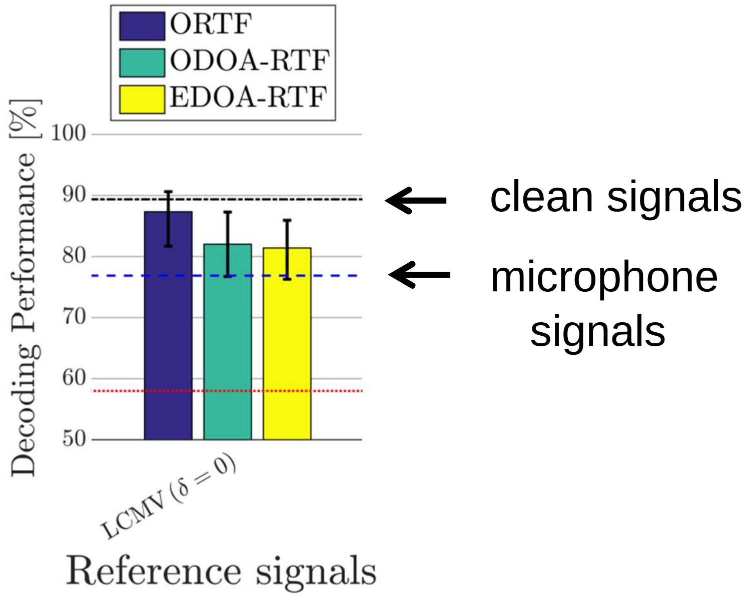
[4] H. Kayser *et al.*, “A discriminative learning approach to probabilistic acoustic source localization,” in Proc. International Workshop on Acoustic Signal Enhancement (IWAENC), pp. 99–103, Sep. 2014.

Experimental results: AAD performance

Anechoic-noisy condition



Reverberant-noisy condition

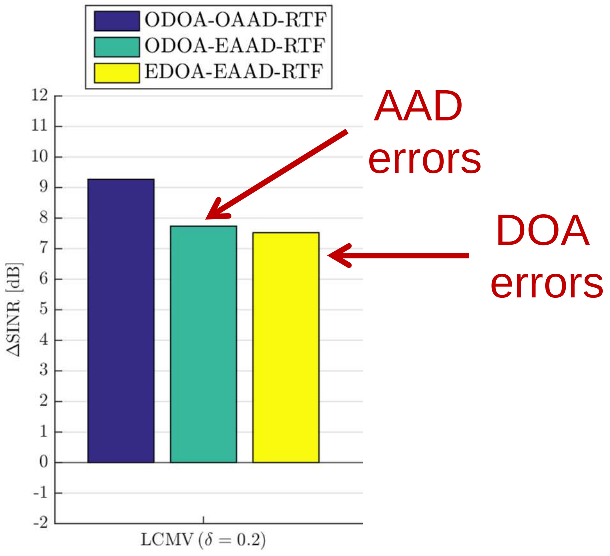


- **Improved AAD performance using LCMV output signals compared to using microphone signals**
- Reverberant condition: anechoic RTFs decrease AAD performance compared to reverberant RTFs
- Anechoic + reverberant condition: **robust to DOA estimation errors**

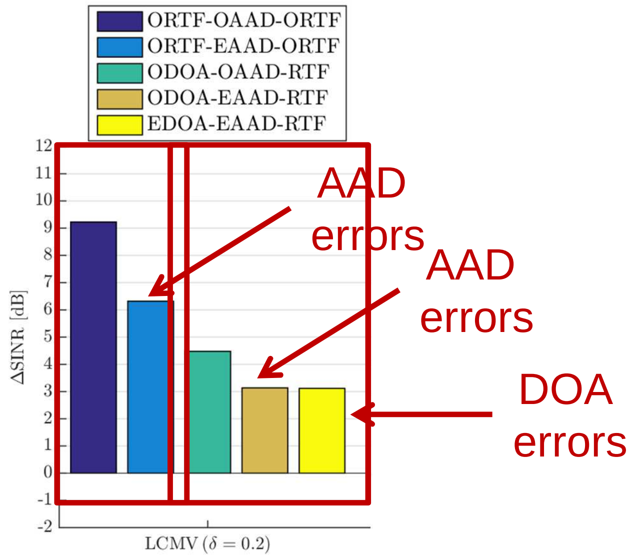
Experimental results: speech enhancement

Binaural SINR improvement averaged over all trials

Anechoic-noisy condition



Reverberant-noisy condition



- **Large binaural SINR improvement on average**
- **AAD errors degrade binaural SINR improvement** (attended speaker wrongly suppressed)
- **Robust to DOA estimation errors**
- Reverberant condition: lower SINR improvement than anechoic condition when using anechoic RTFs

Summary

- **Least-squares-based AAD method**
  - **clean speech signals** are not available as reference signals in practice
  - reference signals affected by **reverberation or noise** → comparable decoding performance as when using clean signals
  - reference signals affected by **interfering speaker** → decoding performance significantly decreases
  - **Improved decoding performance using LCMV output signals compared to using microphone signals**

- **Cognitive-driven binaural speech enhancement system**
  - **Large binaural SINR improvement on average although AAD errors degrade performance**
  - In reverberant conditions: better SINR improvement using reverberant than anechoic RTFs

Thanks for your attention!

Questions?