Linear and Parametric Microphone Array Processing

Part 5 - Joint Linear and Parametric Spatial Processing

Emanuël A. P. Habets[1] and Sharon Gannot[2]

[1] International Audio Laboratories Erlangen, Germany
A joint institution of the University of Erlangen-Nuremberg and Fraunhofer IIS

[2] Faculty of Engineering, Bar-Ilan University, Israel

ICASSP 2013, Vancouver, Canada

Overview

1. Motivation

"Classical" Linear Spatial Filtering:

+ High amount of noise plus interference reduction
+ Controllable tradeoff between speech distortion and noise reduction
+ Controllable tradeoff between different noise types
– Not very robust w.r.t. estimation errors, position changes, etc.
– Relatively slow response time

Parametric Spatial Filtering:

+ Fast response time
+ Relatively robust w.r.t. estimation errors, position changes, etc.
+ Possibility to manipulate parameters (e.g., virtual source displacement)
– Inherent tradeoff between speech distortion and noise reduction
– Model violations can introduce audible artifacts [Thiergart and Habets, 2012]
– Relatively poor interference reduction due to the tradeoff and model violations

Overview

The main idea behind **informed spatial filtering** is to incorporate relevant information about the specific problem into the design of spatial filters and the estimation of required statistics.

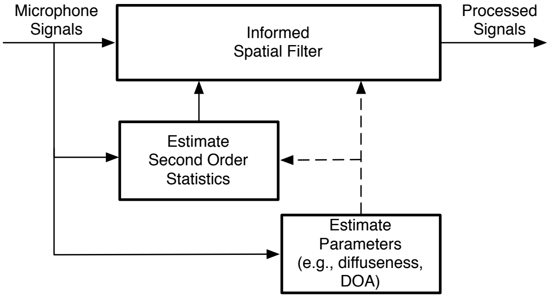


Figure: Informed spatial filtering approach.

2. Informed Spatial Filtering

A selection of parameters that can be used (see Part 4):

- Signal-to-diffuse ratio (SDR):

$$\Gamma(k, m, \mathbf{p}_i) = \frac{P_{\text{dir}}(k, m, \mathbf{p}_i)}{P_{\text{diff}}(k, m)},$$

  where P_{dir} is the power of the direct component at position $\mathbf{p}_i$ and P_{diff} is the power of the diffuse component (assuming a spatially homogenous sound field).

- Time and frequency dependent direction-of-arrival estimates.

- Time and frequency dependent interaural level differences.

- Time and frequency dependent interaural phase differences.

- ...

Overview

3.1 Example A: Extracting Coherent Sound Sources

- **Signal model:** $\mathbf{y}(k,m) = \mathbf{x}(k,m) + \mathbf{v}(k,m)$.

- **Assumption:** Desired signals are strongly coherent across the array.

- **Aim:** Estimate $X_1(k,m)$ using a parametric multichannel Wiener filter [Benesty et al., 2011]:

$$\mathbf{h}_{\mathrm{PMWF}}(k,m) = \frac{\mathbf{\Phi}_{\mathbf{v}}^{-1}(k,m)\mathbf{\Phi}_{\mathbf{x}}(k,m)}{\lambda(k,m) + \mathrm{tr}\left\{\mathbf{\Phi}_{\mathbf{v}}^{-1}(k,m)\mathbf{\Phi}_{\mathbf{x}}(k,m)\right\}}\mathbf{u}_1$$

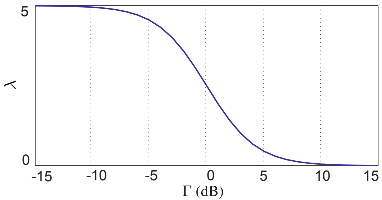


Figure: Mapping from the input signal-to-diffuse ratio to the tradeoff parameter λ [Taseska and Habets, 2012].

Proposed Solution [Taseska and Habets, 2012]

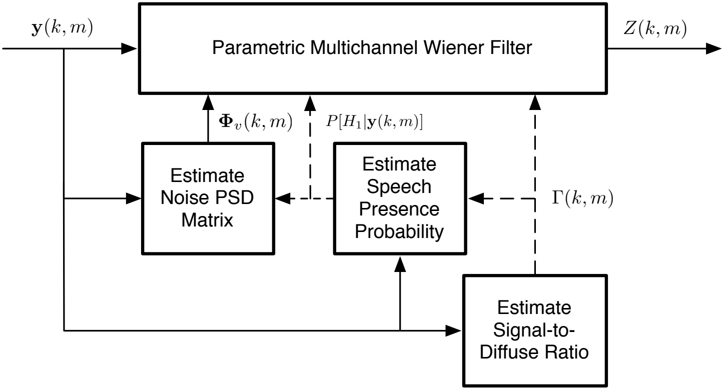


Figure: Block diagram of the proposed system.

Algorithm Summary

High-level description of the proposed algorithm [Taseska and Habets, 2012]:

1. Compute signal-to-diffuse ratio (SDR) using [Thiergart et al., 2012].
2. Compute *a priori* speech presence probability (SPP) based on the SDR.
3. Compute multichannel *a posteriori* SPP [Souden et al., 2010].
4. Update noise PSD matrix using the *a posteriori* SPP.
5. Compute the tradeoff parameter for the parametric multichannel Wiener filter (PMWF) based on the SDR:
   - When the SDR is high, we decrease the amount of speech distortion.
   - When the SDR is low, we increase the amount of noise reduction.
6. Compute and apply the parametric multichannel Wiener filter.

Results (1)

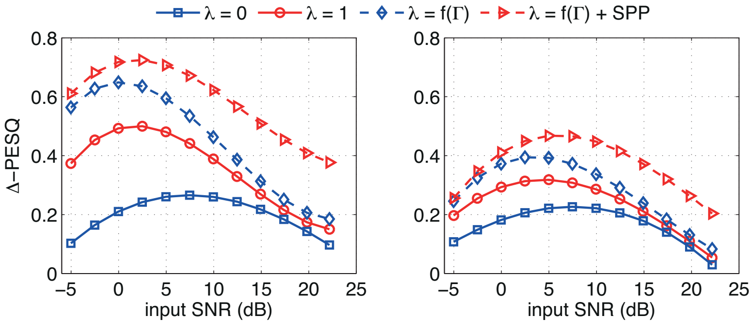


Figure: Performance evaluation: PESQ improvement for stationary diffuse noise (left) and diffuse babble speech (right) [Taseska and Habets, 2012].

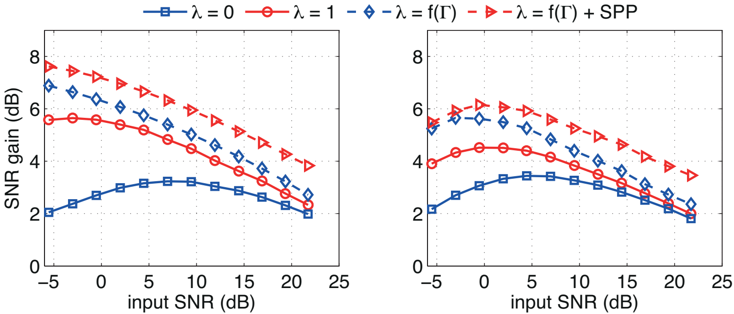
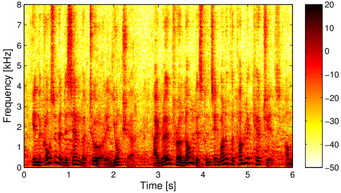
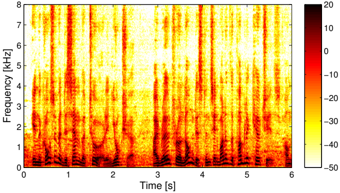


Figure: Performance evaluation: segmental SNR improvement for stationary diffuse noise (left) and diffuse babble speech (right) [Taseska and Habets, 2012].

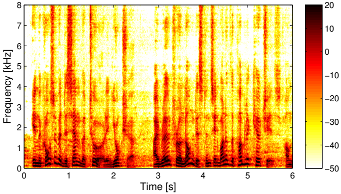
Results (3)



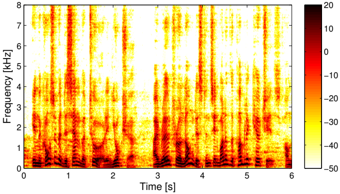
(a) First Microphone Signal



(b) MVDR



(c) Parametric MWF



(d) Parametric MWF with MC-SPP

Figure: Examples obtained using M=4 microphone signals corrupted by sensor noise and babble speech (input SNR = 10 dB).

3.2 Example B: Dereverberation in the SH Domain

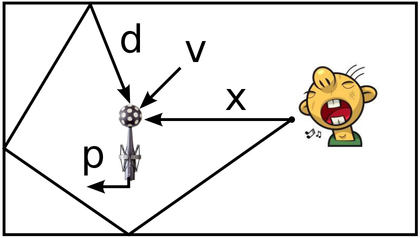
Assumed signal model with stacked spherical harmonic components:

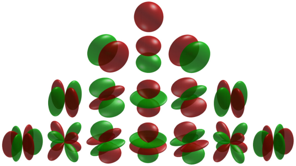
$$\widetilde{\mathbf{p}}(k,m) = \widetilde{\mathbf{x}}(k,m) \quad + \quad \underbrace{\widetilde{\mathbf{d}}(k,m) + \widetilde{\mathbf{v}}(k,m)}$$

$$= \boldsymbol{\gamma}(k,m)\widetilde{X}_{00}(k,m) \; + \; \widetilde{\mathbf{u}}(k,m)$$

$$\boldsymbol{\gamma}(k,m) = \frac{\widetilde{\mathbf{x}}(k,m)}{\widetilde{X}_{00}(k,m)} = \frac{\mathbf{y}(\Omega_{\mathrm{dir}})}{Y_{00}(\Omega_{\mathrm{dir}})} = \boldsymbol{\gamma}_{\mathrm{dir}},$$

where Y_{00} is the zero-order spherical harmonic and Ω_{dir} is the DOA.





Spherical Harmonics up to order 3

Proposed Solution [Braun et al., 2013]

- **Desired signal:** The direct signal component $\widetilde{X}_{00}(k,m)$ which corresponds to the sound pressure measured at the center of the array in the absence of the spherical microphone array.
- **Assumption:** Direct, diffuse and noise components are mutually uncorrelated.
- **Proposed solution:** The (rank-1) MWF provides an MMSE estimate of $\widetilde{X}_{00}(k,m)$. For practical reasons, we split the MWF into an MVDR filter followed by a single-channel Wiener filter:

$$
\begin{aligned}
\mathbf{h}_{\mathrm{MWF}}(k,m) &= \frac{\phi_{\widetilde{X}_{00}}(k,m)\,\mathbf{\Phi}_{\widetilde{\mathbf{u}}}^{-1}(k,m)\,\boldsymbol{\gamma}_{\mathrm{dir}}}{\phi_{\widetilde{X}_{00}}(k,m)\,\boldsymbol{\gamma}_{\mathrm{dir}}^{H}\,\mathbf{\Phi}_{\widetilde{\mathbf{u}}}^{-1}(k,m)\,\boldsymbol{\gamma}_{\mathrm{dir}}+1} \\
&= \underbrace{\frac{\mathbf{\Phi}_{\widetilde{\mathbf{u}}}^{-1}(k,m)\,\boldsymbol{\gamma}_{\mathrm{dir}}}{\boldsymbol{\gamma}_{\mathrm{dir}}^{H}\,\mathbf{\Phi}_{\widetilde{\mathbf{u}}}^{-1}(k,m)\,\boldsymbol{\gamma}_{\mathrm{dir}}}}_{\mathbf{h}_{\mathrm{MVDR}}(k,m)} \cdot \underbrace{\frac{\phi_{\widetilde{X}_{00}}}{\phi_{\widetilde{X}_{00}}+\left[\boldsymbol{\gamma}_{\mathrm{dir}}^{H}\,\mathbf{\Phi}_{\widetilde{\mathbf{u}}}^{-1}(k,m)\,\boldsymbol{\gamma}_{\mathrm{dir}}\right]^{-1}}}_{H_{\mathrm{W}}(k,m)}
\end{aligned}
$$

Parameter-based PSD Matrix Estimation

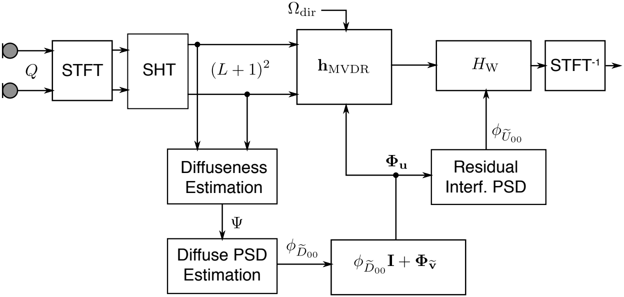
Required information:

- Direction of arrival (DOA) $\rightarrow \boldsymbol{\gamma}_{\text{dir}}$
- Interference PSD matrix:
  $\boldsymbol{\Phi}_{\tilde{\mathbf{u}}}(k,m) = \boldsymbol{\Phi}_{\tilde{\mathbf{d}}}(k,m) + \boldsymbol{\Phi}_{\tilde{\mathbf{v}}}(k,m)$

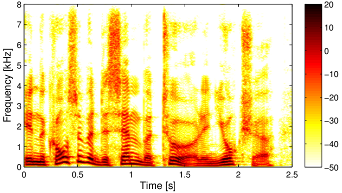
Diffuse PSD matrix estimation:

- Assume model for diffuse sound component:
  $\boldsymbol{\Phi}_{\tilde{\mathbf{d}}}(k,m) = \phi_{\widetilde{D}_{00}}(k,m)\,\mathbf{I}_{(L+1)^2}$
- Calculate diffuse sound PSD using an estimate of the diffuseness Ψ:

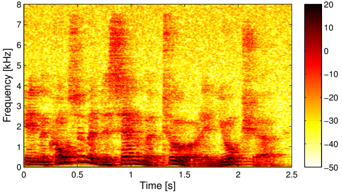
$$\phi_{\widetilde{D}_{00}}(k,m) = \frac{\phi_{\widetilde{P}_{00}}(k,m) - \phi_{\widetilde{V}_{00}}(k,m)}{\Psi^{-1}(k,m)}$$



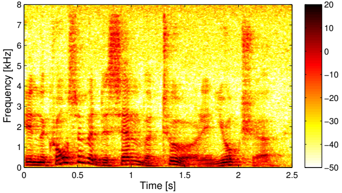
Results



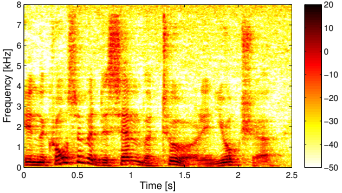
(a) Reference $\widetilde{X}_{00}(k,m)$



(b) Received $\widetilde{P}_{00}(k,m)$



(c) Processed: MVDR



(d) Processed: MWF

Figure: Examples obtained using simulated signals [Jarrett et al., 2012] (source-array distance is 2 m, SNR = 20 dB, T_{60}=400 ms).

3.3 Example C: Directional Filtering

- Flexible sound acquisition in noisy and reverberant environments with rapidly changing acoustic scenes is a common problem in modern communication systems.

- A spatial filter is proposed that provides an **arbitrary spatial response** for J sources being simultaneously active per time and frequency.

- The proposed filter provides an **optimal tradeoff** between the white noise gain (WNG) and the directivity index.

- The filter exploits instantaneous information on the spatial sound (narrowband DOAs, diffuse-to-noise ratio) which allows a nearly immediate adaption to changes in the acoustic scene.

AUDIO LABS

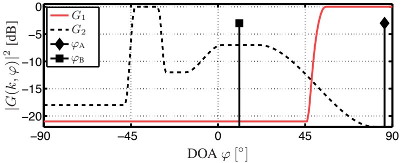
- **Signal model**: Based on a multi-wave sound field model, the M microphone signals can be expressed as:

$$\mathbf{y}(k,m) = \underbrace{\sum_{j=1}^{J} \mathbf{x}^{(j)}(k,m)}_{J \text{ plane waves}} + \underbrace{\mathbf{d}(k,m)}_{\text{diffuse sound}} + \underbrace{\mathbf{v}(k,m)}_{\text{sensor noise}}$$

- **Aim:** Capturing J plane waves ($J \leq M$) with desired arbitrary gain while attenuating the sensor noise and reverberation.

The desired signal is given by:

$$Z(k,m) = \sum_{j=1}^{J} G(k,\varphi_j) X_1^{(j)}(k,m)$$



- The desired signal is estimated using an informed LCMV filter:

$$\widehat{Z}(k,m) = \mathbf{h}_{\mathrm{iLCMV}}^{H}(k,m)\, \mathbf{y}(k,m)$$

Proposed Solution (1)

- The proposed informed LCMV filer is given by:

$$\mathbf{h}_{\mathrm{iLCMV}} = \underset{\mathbf{h}}{\operatorname{argmin}} \; \mathbf{h}^{H} \left[\boldsymbol{\Phi}_{\mathbf{d}}(k,m) + \boldsymbol{\Phi}_{\mathbf{v}}(k,m)\right] \mathbf{h}$$
$$\text{s.t.} \quad \mathbf{h}^{H}(k,m)\,\mathbf{a}(k,\varphi_j) = G(k,\varphi_j), \quad j \in \{1,2,\ldots,J\}$$

where $\mathbf{a}(k,\varphi_j)$ denotes the steering vector for the jth plane wave at time m and frequency k.

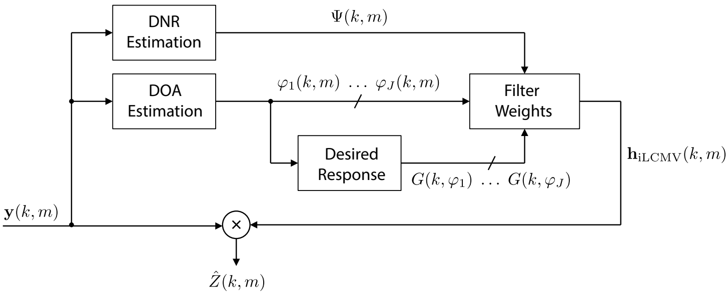
For the assumed signal model, we can alternatively minimize

$$\mathbf{h}^{H} \left[\Psi(k,m)\,\boldsymbol{\Gamma}_{\mathsf{d}}(k) + \mathbf{I}\right] \mathbf{h},$$

where $\Psi(k,m)$ denotes the instantaneous diffuse-to-noise ratio (DNR) and $\boldsymbol{\Gamma}_{\mathsf{d}}(k)$ denotes the spatial coherence matrix of the diffuse sound field.

- The filter is updated for each time and frequency given the instantaneous parametric information (DOAs, DNR).
- The filter requires knowledge of the DNR, which can be estimated using an auxiliary spatial filter (see poster session AASP-P8 on Friday or [Thiergart and Habets, 2013]).

Proposed Solution (2)



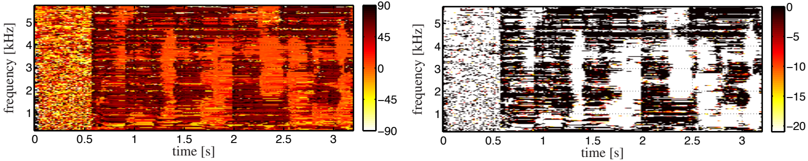
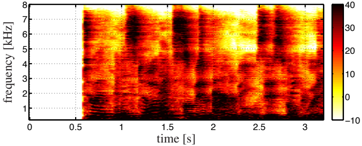
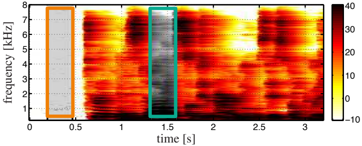


Figure: Left: DOA $\varphi_1(k,m)$ as a function of time and frequency. Right: Desired response $|G(k,\varphi_1)|^2$ in dB for DOA $\varphi_1(k,m)$ as a function of time and frequency.

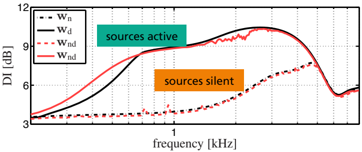
Results (1)



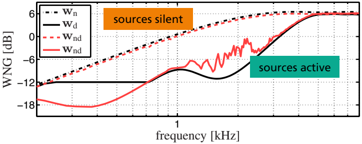
(a) True



(b) Estimated



(a) Mean DI



(b) Mean WNG

Figure: Top: True DNR in dB. Bottom: Estimated DNR in dB.

Figure: Top: Directivity index (DI) in dB. Bottom: White noise gain (WNG) in dB. $\mathbf{w}_{\mathrm{n}}$ minimizes the noise power, $\mathbf{w}_{\mathrm{d}}$ minimizes the diffuse power, $\mathbf{w}_{\mathrm{nd}}$ is the proposed iLCMV filter that minimizes the diffuse plus noise power [shown when the sources are active (red solid line) and silent (red dashed line)].

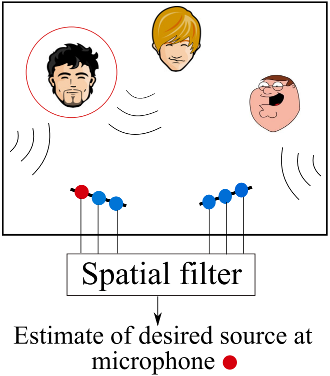
Results (2)

AUDIO LABS

- The proposed spatial filter provides a high DI when the sound field is diffuse and a high WNG when the sensor noise is dominant.
- Interfering sound can be strongly attenuated if desired.
- The proposed DNR estimator provides a sufficiently high accuracy and temporal resolution to allow signal enhancement under adverse conditions even in changing acoustic scenes.

	SegSIR [dB]		SegSRR [dB]		SegSNR [dB]		PESQ	
$*$	11	(11)	−7	(−7)	26	(26)	1.5	(1.5)
$\mathbf{w}_{\mathrm{n}}$	21	(32)	−2	(−3)	**33**	(**31**)	2.0	(1.7)
$\mathbf{w}_{\mathrm{d}}$	**26**	(**35**)	0	(−**1**)	22	(24)	**2.1**	(**2.0**)
$\mathbf{w}_{\mathrm{nd}}$	25	(**35**)	**1**	(−**1**)	28	(26)	**2.1**	(**2.0**)

Table: Performance of all spatial filters [$*$ unprocessed, first sub-column using true DOAs (of the sources), second sub-column using estimated DOAs (of the plane waves)].



Scenario

- Multiple talkers
- Additive background noise
- Distributed sensor arrays

Applications

- Teleconferencing systems
- Automatic speech recognition
- Spatial sound reproduction

- **Signal model:** $\mathbf{y}(k,m) = \mathbf{x}^{(d)}(k,m) + \sum_{i \neq d} \mathbf{x}^{(i)}(k,m) + \mathbf{v}(k,m)$.
- **Aim:** Obtain an MMSE estimate of $X_1^{(d)}(k,m)$.

Proposed Solution [Taseska and Habets, 2013]

- Hypotheses:

$$\mathcal{H}_{\mathbf{v}} : \mathbf{y}(k,m) = \mathbf{v}(k,m) \;\rightarrow \text{speech absent}$$

$$\mathcal{H}_{\mathbf{x}} : \mathbf{y}(k,m) = \mathbf{x}(k,m) + \mathbf{v}(k,m) \;\rightarrow \text{speech present}$$

$$\mathcal{H}_{\mathbf{x}}^{j} : \mathbf{y}(k,m) = \mathbf{x}^{(j)}(k,m) + \underbrace{\sum_{i\neq j}^{J} \mathbf{x}^{(i)}(k,m)}_{\approx 0} + \mathbf{v}(k,m) \qquad j = 1, 2, \ldots, J$$

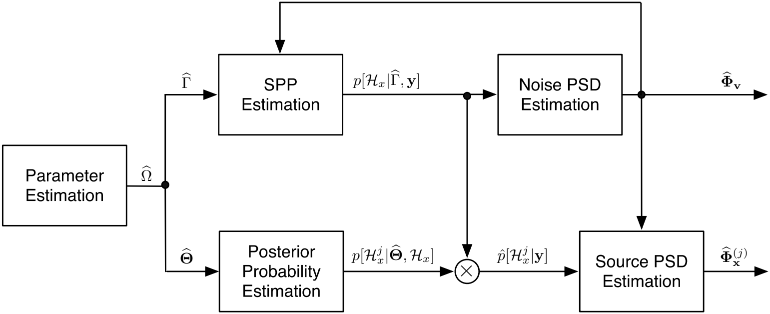
- Recursive estimation of the PSD matrices:

$$\widehat{\mathbf{\Phi}}_{\mathbf{x}}^{(j)}(m) = p[\mathcal{H}_{\mathbf{x}}^{j} \,|\, \mathbf{y}] \left(\alpha_x \, \widehat{\mathbf{\Phi}}_{\mathbf{x}}^{(j)}(m-1) + (1-\alpha_x)\, \mathbf{y}\mathbf{y}^{\mathrm{H}}\right) + \left(1 - p[\mathcal{H}_{\mathbf{x}}^{j} \,|\, \mathbf{y}]\right) \widehat{\mathbf{\Phi}}_{\mathbf{x}}^{(j)}(m-1)$$

- Signal-to-diffuse ratio (Γ) and position (Θ) -based posterior probabilities:

$$\boxed{p[\mathcal{H}_{\mathbf{x}}^{j} \,|\, \mathbf{y}] = p[\mathcal{H}_{\mathbf{x}}^{j} \,|\, \mathbf{y}, \mathcal{H}_{\mathbf{x}}] \cdot p[\mathcal{H}_{\mathbf{x}} \,|\, \mathbf{y}] \approx p[\mathcal{H}_{\mathbf{x}}^{j} \,|\, \Theta, \mathcal{H}_{\mathbf{x}}] \cdot p[\mathcal{H}_{\mathbf{x}} \,|\, \Gamma, \mathbf{y}]}$$

Parameter-based PSD Matrix Estimation

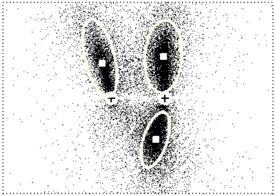


- The distribution $p[\widehat{\Theta}\,|\,\mathcal{H}_{\mathbf{x}}]$ is modelled as a Gaussian mixture (GM).
- GM parameters estimated by the Expectation-Maximization algorithm.

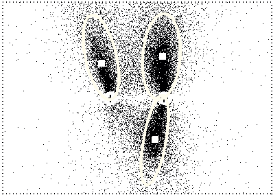
Results (1)

Setup:

- Three reverberant sources with approximately equal power, diffuse babble speech (SNR=22 dB), and uncorrelated sensor noise (SNR =50 dB). The reverberation time was T60 = 250 ms.
- Two uniform circular arrays were used with three omnidirectional microphones, a diameter 2.5 cm and an inter-array spacing of 1.5 m.



(a) Training during single-talk



(b) Training during triple-talk

Figure: Output of the EM algorithm (3 iterations) and 4.5 s of noisy speech data. The actual source positions are denoted by white squares. The array location is marked by a plus symbol. The interior of each ellipse contains 85% probability mass of the respective Gaussian.

Results (2)

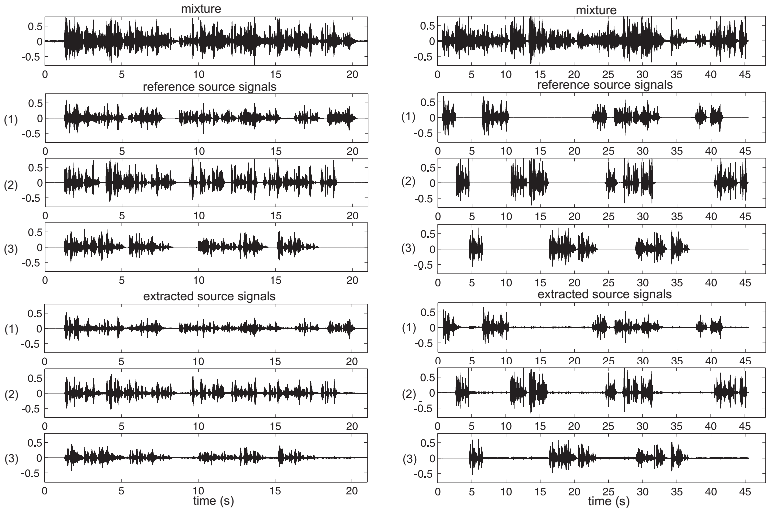


Figure: Left: constant triple-talk scenario. Right: mainly single-talk scenario.

Audio files available at http://home.tiscali.nl/ehabets/publications/Taseska2013.html.

Special thanks to Sebastian Braun, Maja Taseska,

Oliver Thiergart and Daniel Jarrett for their contributions.

References I

Benesty, J., Chen, J., and Habets, E. A. P. (2011).
Speech Enhancement in the STFT Domain.
SpringerBriefs in Electrical and Computer Engineering. Springer-Verlag.

Braun, S., Jarrett, D. P., Fischer, J., and Habets, E. A. P. (2013).
An informed spatial filter for dereverberation in the spherical harmonic domain.
In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada.

Jarrett, D. P., Habets, E. A. P., Thomas, M. R. P., and Naylor, P. A. (2012).
Rigid sphere room impulse response simulation: algorithm and applications.
J. Acoust. Soc. Am., 132(3):1462–1472.

Souden, M., Chen, J., Benesty, J., and Affes, S. (2010).
Gaussian model-based multichannel speech presence probability.
IEEE Trans. Audio, Speech, Lang. Process., 18(5):1072–1077.

Taseska, M. and Habets, E. (2013).
MMSE-based source extraction using position-based posterior probabilities.
In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*.

References II

Taseska, M. and Habets, E. A. P. (2012).

MMSE-based blind source extraction in diffuse noise fields using a complex coherence-based a priori SAP estimator.

In *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*.

Thiergart, O., Del Galdo, G., and Habets, E. A. P. (2012).

On the spatial coherence in mixed sound fields and its application to signal-to-diffuse ratio estimation.

J. Acoust. Soc. Am., 132(4):2337–2346.

Thiergart, O. and Habets, E. (2013).

Informed optimum spatial filtering using multiple instantaneous direction-of-arrival estimates.

In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*.

Thiergart, O. and Habets, E. A. P. (2012).

Sound field model violations in parametric spatial sound processing.

In *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*.