AUDIO LABS

Parametric Spatial Audio Processing

An Overview and Recent Advances

Emanuël Habets and Oliver Thiergart

140th AES Convention, Paris, France, June 7th, 2016

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

Fraunhofer IIS

Outline

Outline

Introduction

Applications and Motivation

- Many different devices with multiple microphones have emerged

Television screens	Mobile phones	Digital cameras
Up to 4 microphones, usually linear array	2 or more microphones, at different positions	2 omnidirectional microphones or stereo microphone
Voice-controlled television, teleconferencing	Hands-free communication, audio-video recording	Audio-video recording, pictures with sound
Speech enhancement and spatial filtering desired	Speech enhancement and spatial sound recording desired	Spatial sound and consistent audio-video capturing desired

Introduction

Applications and Motivation

- Many different devices with multiple microphones have emerged

Television screens

Up to 4 microphones, usually linear array

Voice-controlled television, teleconferencing

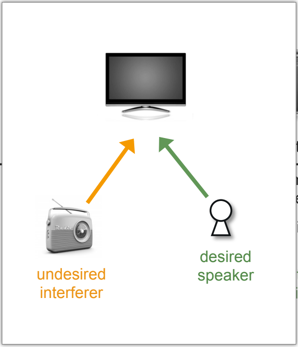
Speech enhancement and spatial filtering desired

tal cameras

nnidirectional microphones or eo microphone

io-video recording, pictures sound

tial sound and consistent io-video capturing desired



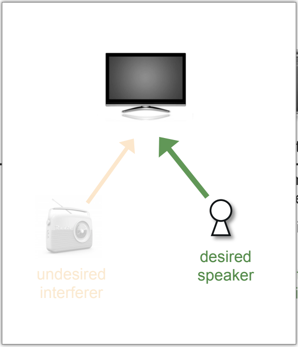
- Many different devices with multiple microphones have emerged

Television screens

Up to 4 microphones, usually linear array

Voice-controlled television, teleconferencing

Speech enhancement and spatial filtering desired



tal cameras

nnidirectional microphones or eo microphone

io-video recording, pictures sound

tial sound and consistent io-video capturing desired

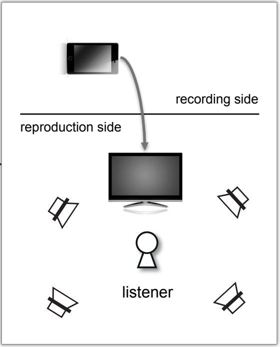
- Many different devices with multiple microphones have emerged



Television screens	Mobile phones	Digital cameras
Up to 4 microphones, usually linear array	2 or more microphones, at different positions	2 omnidirectional microphones or stereo microphone
Voice-controlled television, teleconferencing	Hands-free communication, audio-video recording	Audio-video recording, pictures with sound
Speech enhancement and spatial filtering desired	Speech enhancement and spatial sound recording desired	Spatial sound and consistent audio-video capturing desired

- Many different devices with multiple microphones have emerged

Television screens	Mobile phones
Up to 4 microphones, usually linear array	2 or more microphones, at different positions
Voice-controlled television, teleconferencing	Hands-free communication, audio-video recording
Speech enhancement and spatial filtering desired	Speech enhancement and spatial sound recording desired

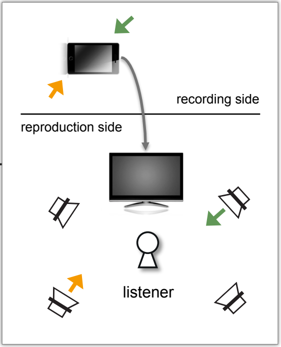


Introduction

Applications and Motivation

- Many different devices with multiple microphones have emerged

Television screens	Mobile phones
Up to 4 microphones, usually linear array	2 or more microphones, at different positions
Voice-controlled television, teleconferencing	Hands-free communication, audio-video recording
Speech enhancement and spatial filtering desired	Speech enhancement and spatial sound recording desired



Introduction

Applications and Motivation

- Many different devices with multiple microphones have emerged

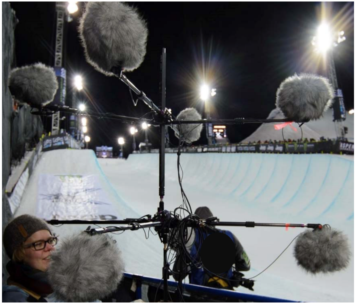


Television screens	Mobile phones	Digital cameras
Up to 4 microphones, usually linear array	2 or more microphones, at different positions	2 omnidirectional microphones or stereo microphone
Voice-controlled television, teleconferencing	Hands-free communication, audio-video recording	Audio-video recording, pictures with sound
Speech enhancement and spatial filtering desired	Speech enhancement and spatial sound recording desired	Spatial sound and consistent audio-video capturing desired

Introduction

Applications and Motivation

- One of the main challenges it to add (spatial) sound recording capabilities to consumer devices using a relatively compact and flexible microphone configuration
- In the context of virtual reality, microphones should be placed such that they are outside the camera's field-of-view (FOV). This can be challenging in particular when a full 3D FOV is required



Introduction

Applications and Motivation

- One of the main challenges it to add (spatial) sound recording capabilities to consumer devices using a relatively compact and flexible microphone configuration
- In the context of virtual reality, microphones should be placed such that they are outside the camera's field-of-view (FOV). This can be challenging in particular when a full 3D FOV is required

Introduction

Parametric Spatial Processing Concept

- A flexible processing scheme is required which can be used for different applications on the different devices
- Parametric-based spatial audio processing makes use of an efficient parametric representation of the sound-field. A major advantage compared to classical spatial processing is the limited number of parameters.

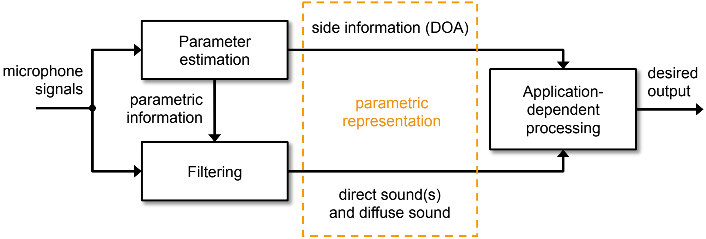


Figure: Parametric spatial audio processing scheme.

Introduction

Existing Parametric Spatial Processing Approaches

Some selected examples:

- 1993 : Computational Auditory Scene Analysis (CASA) c.f. [Kollmeier, Peissig, and V. Hohmann, 1993; Wittkop and V Hohmann, 2003]
- 2000 : Parametric Stereo Coding and Binaural Cue Coding
- 2006 : Using instantaneous TDOAs [Tashev and Acero, 2006]
- 2007 : Directional Audio Coding (DirAC) [Ville Pulkki, 2007]
- 2009 : Dereverberation techniques that make use of the reverberation time and direct-to-reverberation ratio [Habets, Gannot, and Cohen, 2009]
- 2010 : High Angular Resolution Planewave Expansion (HARPEX) [Berge and Barrett, 2010]
- 2015 : Using instantaneous phase differences [Sugiyama and Miyahara, 2015]

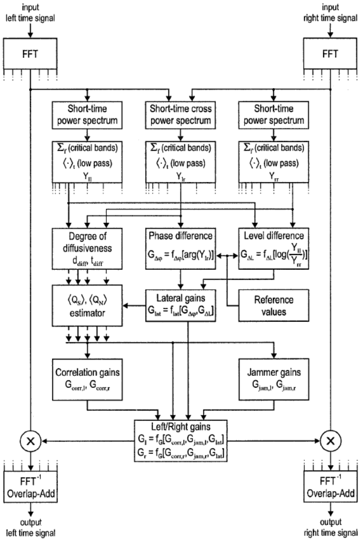


Figure: Block diagram of the strategy-selective algorithm for dereverberation and suppression of lateral noise sources [Wittkop and V Hohmann, 2003]

Introduction

Existing Parametric Spatial Processing Approaches

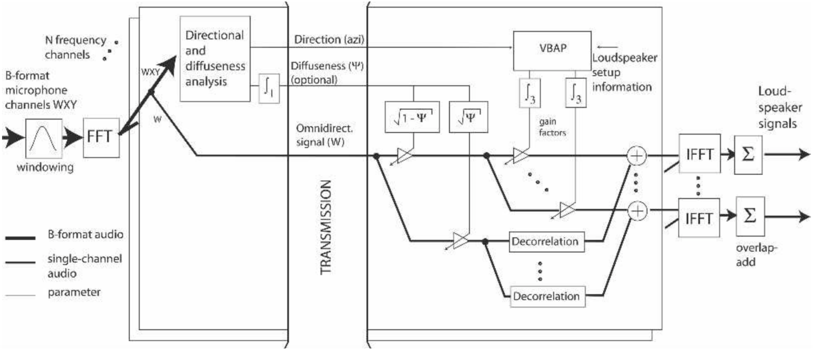


Figure: Block diagram of the original DirAC system [Ville Pulkki, 2007]

Introduction

Objectives of this Tutorial

- Provide an overview of parametric spatial audio processing
- Discuss the advantageous and disadvantages of parametric spatial audio processing
- Explain how the direct and diffuse sound components can be estimated
- Explain how some of the frequently used parameters can be estimated
- Provide some application examples:
  - Directional filtering
  - Acoustical Zoom
  - Spatial Sound Recording and Reproduction

- In practice, the short-time Fourier transform (STFT) is often used.
- STFT Analysis:

$$X(k,n) = \sum_{r=0}^{N-1} x(nR+r)w_{\mathrm{a}}(r)e^{-j\omega_k r} \quad \text{with} \quad \omega_k = \frac{2\pi k}{K},$$

$k = 0, 1, \ldots, K-1$ and $K \geq N$, and R denotes the number of samples between two successive frames.

Introduction

Time–Frequency Analysis and Synthesis - Window Functions

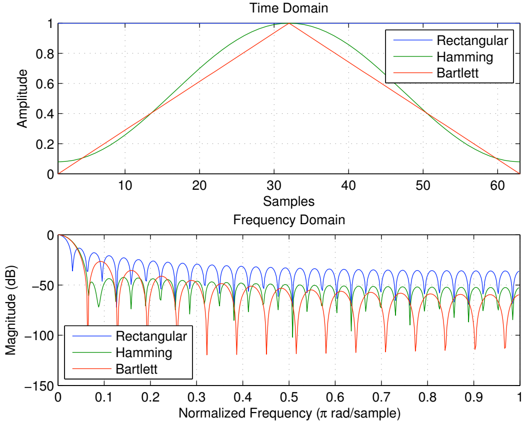


Figure: Rectangular, Hamming, and Bartlett windows. Note that an increased tapering of the window reduces the sidelobe level and increased the width of the main lobe.

- STFT Synthesis:

$$x(t) = \sum_{n} \sum_{k=0}^{K-1} X(k,n) w_{\mathrm{s}}(t - nR) e^{j\omega_k (t-nR)},$$

  where R denotes the number of samples between two successive frames.

- An overlap of 50% is obtained when $R = N/2$.

- The spectrogram is given by $|X(k,n)|^2$.

Introduction

Time–Frequency Analysis and Synthesis

- Completeness condition for analysis window (w_a) and synthesis window (w_s):

$$\sum_n w_\mathrm{a}(t - nR) w_\mathrm{s}(t - nR) = \frac{1}{N} \quad \text{for all } t. \tag{1}$$

- Given analysis and synthesis windows that satisfy (1) we can reconstruct $x(t)$ from its STFT coefficients $X(k, n)$.

- In practice, a Hamming window is often used for the synthesis window.

- The inverse STFT is efficiently implemented using the weighted overlap-add method [Crochiere and Rabiner, 1983].

Time–Frequency Analysis and Synthesis - Spectrogram

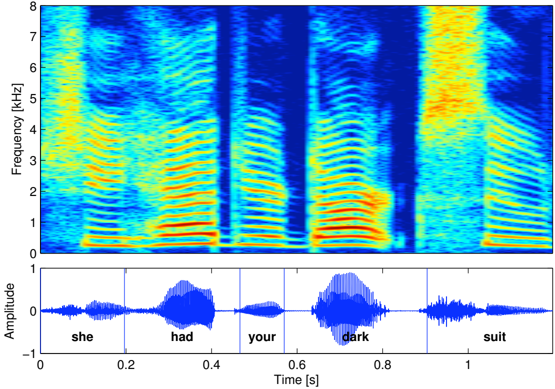


Figure: Spectrogram $(10\log(|X(k,n)|^2))$ of a speech signal (sample frequency 16 kHz, DFT length K $=$ 1024, window length N $=$ 512, hamming window).

Introduction

Time–Frequency Analysis and Synthesis - Spectrogram

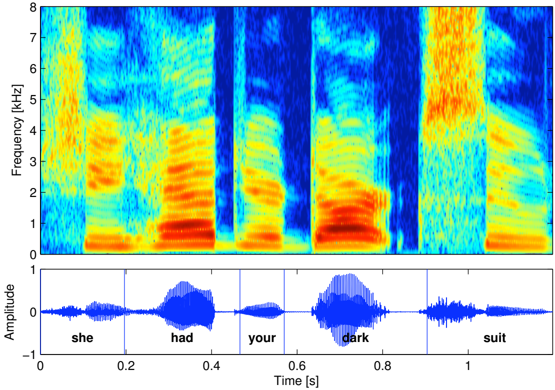


Figure: Spectrogram $(10\log(|X(k,n)|^2))$ of a speech signal (sample frequency 16 kHz, DFT length K = 1024, window length N = 64, hamming window).

Outline

Signal Model

- The sound field is modeled and processed in the time-frequency domain.
- The optimal time-frequency resolution depends an multiple aspects:
  - It should resample the spectral resolution of the human hearing.
  - It depends on the characteristics of the input signals.
  - It depends on the employed parameter estimators and filters.
- Therefore, the time-frequency resolution should be chosen carefully depending on the application and realized system.
- In the following, we consider setups with omnidirectional microphones. In many cases, an extension to directional setups is straight-forward.

Signal Model

Total Sound Field

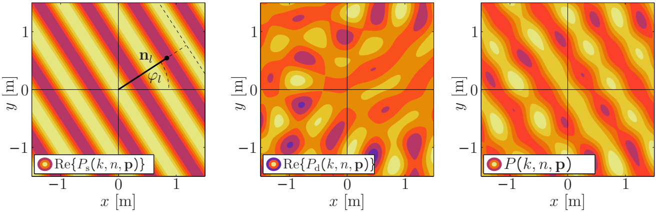
- To achieve the desired flexibility and efficiency, recent approaches use a parametric representation of the spatial sound at one position.
- The sound field in point $\mathbf{p}$ for time index n and frequency band k is modeled as a superposition of L direct sounds and a diffuse sound, i.e.,

$$P(k,n,\mathbf{p}) = \sum_{l=1}^{L} P_{\mathrm{s},l}(k,n,\mathbf{p}) + P_{\mathrm{d}}(k,n,\mathbf{p}).$$

- The direct sounds $P_{\mathrm{s},l}(k,n,\mathbf{p})$ model the direct sound of the sources. The diffuse sound $P_{\mathrm{d}}(k,n,\mathbf{p})$ models the reverberation or ambience.
- Well-known examples where a parametric signal model is employed: DirAC ($L=1$), HARPEX ($L=2$ direct sounds, no diffuse sound).

Signal Model

Total Sound Field



(a) Direct sound field (b) Diffuse sound field (c) Sum of both fields

Figure: Example of a single plane wave, a diffuse field, and the sum of both fields.

- Each direct sound $P_{\mathrm{s},l}(k, n, \mathbf{p})$ is represented as a single plane wave with DOA expressed by the unit-norm vector $\mathbf{n}_l(k, n)$.
- The DOA of the direct sounds can vary quickly in practice and represents a crucial parameter in parametric spatial sound processing.

Signal Model

Total Sound Field

- Given the sound field model, the microphone signals can be expressed as

$$\mathbf{x}(k,n) = \mathbf{x}_{\mathrm{s}}(k,n) + \mathbf{x}_{\mathrm{d}}(k,n) + \mathbf{x}_{\mathrm{n}}(k,n).$$

  $\mathbf{x}_{\mathrm{s}}$: microphone signals corresponding to the sum of the L direct sounds
  $\mathbf{x}_{\mathrm{d}}$: diffuse sound microphone signals
  $\mathbf{x}_{\mathrm{n}}$: stationary noise (e.g., microphone self-noise)

- Assuming mutually uncorrelated signal components, the microphone PSD matrix can be written as

$$\begin{aligned}\mathbf{\Phi}_x(k,n) &= \mathrm{E}\left\{\mathbf{x}(k,n)\mathbf{x}^{\mathrm{H}}(k,n)\right\} \\ &= \mathbf{\Phi}_{\mathrm{s}}(k,n) + \mathbf{\Phi}_{\mathrm{d}}(k,n) + \mathbf{\Phi}_{\mathrm{n}}(k).\end{aligned}$$

Signal Model

Direct Sound Model

- The microphone signals corresponding to the sum of the L direct sounds can be written as

$$\mathbf{x}_{\mathrm{s}}(k,n) = \mathbf{V}(k,n)\mathbf{s}(k,n,\mathbf{p}_1),$$

  where the vector $\mathbf{s}(k,n,\mathbf{p}_1)$ contains the L direct sounds $P_{\mathrm{s},l}(k,n,\mathbf{p}_1)$ at the position $\mathbf{p}_1$ of the reference microphone.

- The matrix $\mathbf{V}(k,n)$ contains the relative transfer functions between the M microphones and the reference microphone for each direct sound, i.e.,

$$V_{m,l}(k,n) = e^{-\jmath\kappa(\mathbf{p}_m - \mathbf{p}_1)^{\mathrm{T}}\mathbf{n}_l}.$$

- The expected powers of the direct sounds are given by

$$\Phi_{\mathrm{s},l}(k,n) = \mathrm{E}\left\{|P_{\mathrm{s},l}(k,n,\mathbf{p}_1)|^2\right\}.$$

Signal Model

Diffuse Sound Model

- The diffuse sound at the m-th microphone is a superposition of many plane waves with random phase and uniformly distributed DOAs, i.e.,

$$X_{\mathrm{d},m}(k,n) = \sqrt{\frac{\Phi_{\mathrm{d}}(k,n)}{N}} \sum_{i=1}^{N} e^{-\jmath\kappa \mathbf{p}_m^{\mathrm{T}} \mathbf{n}_i + \jmath\theta_i},$$

  where $\Phi_{\mathrm{d}}(k,n)$ is the expected power of the diffuse sound

- For this model, the diffuse sound PSD matrix is given by

$$\begin{aligned}\mathbf{\Phi}_{\mathrm{d}}(k,n) &= \mathrm{E}\left\{\mathbf{x}_{\mathrm{d}}(k,n)\mathbf{x}_{\mathrm{d}}^{\mathrm{H}}(k,n)\right\} \\ &= \Phi_{\mathrm{d}}(k,n)\mathbf{\Gamma}_{\mathrm{d}}(k),\end{aligned}$$

  where $\mathbf{\Gamma}_{\mathrm{d}}(k)$ is the diffuse coherence matrix.

Signal Model

Diffuse Coherence

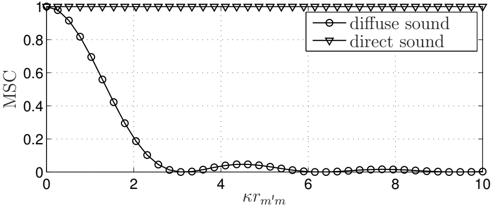


Figure: Magnitude-squared coherence between two omnidirectional microphones for a direct sound field a spherically isotropic diffuse sound field

- The (m, m')-th element of $\mathbf{\Gamma}_{\mathrm{d}}(k)$ is the diffuse sound coherence between microphone m and m', which is the well-known sinc-function depending on the wavenumber κ and microphone spacing $r_{m'm}$, i.e., [Cook et al., 1955]

$$\gamma_{\mathrm{d},m'm}(k) = \frac{\sin(\kappa r_{m'm})}{\kappa r_{m'm}}.$$

Signal Model

Diffuse Sound Relation between Different Microphones

- In the following, we introduce the definition

$$\mathbf{u}(k,n) \equiv \mathbf{x}_{\mathrm{d}}(k,n) P_{\mathrm{d}}^{-1}(k,n,\mathbf{p}_1),$$

which relates the diffuse sound at the M microphones to the diffuse sound at the first microphone.

- The vector $\mathbf{u}(k,n)$ is an unobservable random variable and its mean is the diffuse coherence vector, i.e., [Thiergart and Habets, 2014]

$$\mathrm{E}\{\mathbf{u}(k,n)\} = \boldsymbol{\gamma}_{\mathrm{d}}(k),$$

where $\boldsymbol{\gamma}_{\mathrm{d}}(k) = [1, \gamma_{\mathrm{d},12}(k), \ldots, \gamma_{\mathrm{d},1M}(k)]^{\mathrm{T}}$ is the first column of $\boldsymbol{\Gamma}_{\mathrm{d}}(k)$ containing the known diffuse sound coherences.

Signal Model

Noise Model and Useful Ratios

- The noise component is assumed to be stationary and independent and identically distributed (iid), i.e.,

$$\boldsymbol{\Phi}_{\mathrm{n}}(k) = \mathrm{E}\left\{\mathbf{x}_{\mathrm{n}}(k,n)\mathbf{x}_{\mathrm{n}}^{\mathrm{H}}(k,n)\right\} = \Phi_{\mathrm{n}}(k)\mathbf{I}_M .$$

- A useful ratio for later is the diffuse-to-noise ratio (DNR), defined as

$$\mathrm{DNR}(k,n) = \frac{\Phi_{\mathrm{d}}(k,n)}{\Phi_{\mathrm{n}}(k)},$$

  which is strongly time-varying in practice.

- Another useful ratio is the signal-to-diffuse ratio (SDR), which, for $L = 1$, is defined as

$$\mathrm{SDR}(k,n) = \frac{\Phi_{\mathrm{s}}(k,n)}{\Phi_{\mathrm{d}}(k,n)}.$$

Signal Model

Discussion of the Underlying Model Assumptions

- For $L = 1$ the source signals must be sparse (W-disjoint orthogonal), otherwise model violations occur when multiple sources are active.
- For instance in [Thiergart and Habets, 2012; Laitinen and V. Pulkki, 2012] the effects of such model violations are studied for the application of spatial sound reproduction.
- Assuming a multi-wave model ($L > 1$) greatly relaxes the sparsity requirement but also increases the complexity of the corresponding parameter estimators and filters.
- The plane wave model holds reasonably well in the far-field of the sources given that the inter-microphone distances are small compared to the distance to the sources.
- Assuming that the direct sound and diffuse sound are uncorrelated holds reasonably well for practical time-frequency resolutions.

Outline

Signal and Parameter Estimation

Overview

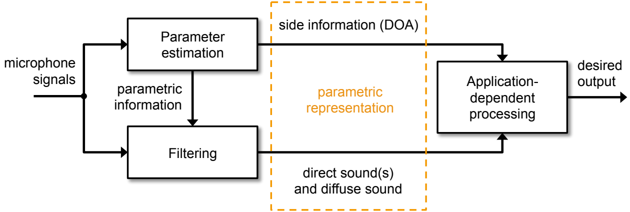


Figure: Parametric spatial audio processing scheme.

- Realizing applications with the parametric spatial audio processing requires
  - Estimating parameters of the underlying sound field model (e.g., DOA),
  - Extracting the direct sound(s) at the reference microphone,
  - Extracting the diffuse sound at the reference microphone.

Signal and Parameter Estimation

Overview

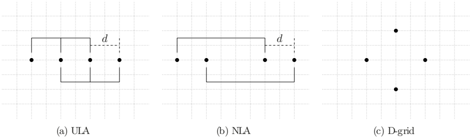


Figure: Typical microphone setups in practice.

- There exists a huge variety of parameter and signal estimators depending on the microphone setup and sound field model (single-wave, multi-wave).
- In the following, we discuss some selected estimators:
  - Direct and diffuse sound extraction with optimal single-channel filters,
  - Direct and diffuse sound extraction with optimal multi-channel filters,
  - SDR estimation based on the spatial coherence.

Signal and Parameter Estimation

Single-channel Direct Sound Extraction

- We assume the single-wave case $(L = 1)$ for the following single-channel filters. Applying the filter $W_\mathrm{s}(k,n)$ to the reference microphone provides an estimate of the direct sound, i.e.,

$$\widehat{P}_\mathrm{s}(k,n,\mathbf{p}_1) = W_\mathrm{s}(k,n) X_1(k,n).$$

- Without loss of generality, we consider an omnidirectional reference microphone in the following.
- To extract the direct sound from the microphone signals, we commonly make use of filters which are optimal in some specific sense.

Signal and Parameter Estimation

Single-channel Direct Sound Extraction: Wiener Filter

- The optimal single-channel Wiener filter minimizes the mean-square error (MSE) between the true and estimated direct sound, i.e.,

$$W_{\mathrm{s}}(k,n) = \arg\min_{W} \mathrm{E}\left\{ \left| W X_1(k,n) - P_{\mathrm{s}}(k,n) \right|^2 \right\}.$$

- One solution when substituting the signal model is given by

$$W_{\mathrm{s}}(k,n) = \left[\frac{\mathrm{SDR}(k,n)}{\mathrm{SDR}(k,n) + \mathrm{DNR}^{-1}(k,n) + 1} \right].$$

- In practice, $W_{\mathrm{s}}(k,n)$ should be limited to a specific lower bound to avoid musical tones. Moreover, spectral or temporal smoothing techniques can be applied (for instance, smoothing in ERB bands).

- The parametric Wiener filter includes additional weighting factors to control the trade-off between noise suppression and speech distortions, i.e.,

$$W_{\mathrm{s}}(k,n) = \left[\frac{\mathrm{SDR}(k,n)}{\mathrm{SDR}(k,n) + \alpha \mathrm{DNR}^{-1}(k,n) + \alpha}\right]^{\beta}.$$

- For $\beta = 0.5$ and $\alpha = 1$ we obtain the well-known square-root Wiener filter. Assuming $\Phi_{\mathrm{n}}(k) = 0$ (high SNR or DNR situations), this filter becomes

$$W_{\mathrm{s}}(k,n) = \sqrt{1 - \Omega(k,n)},$$

where

$$\Omega(k,n) = \frac{1}{1 + \mathrm{SDR}(k,n)}.$$

Signal and Parameter Estimation

Single-channel Direct Sound Extraction: Parametric Wiener Filter

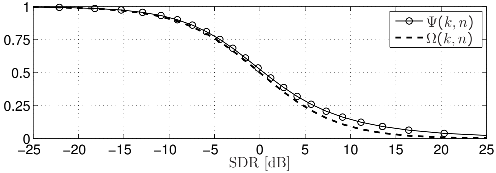


Figure: Comparison of $\Omega(k,n)$ to the intensity-based diffuseness $\Psi(k,n)$ [Del Galdo et al., 2012].

- The term $\Omega(k,n)$ is a very close approximation of the so-called diffuseness $\Psi(k,n)$, which was introduced in DirAC and which is defined based on the temporal variation of the active sound intensity vector.

- Hence, the diffuseness-based signal extraction in DirAC represents the single-channel square-root Wiener filter.

Single-channel Diffuse Sound Extraction: (Parametric) Wiener Filter

- The diffuse sound can be extracted using a single-channel filter similarly as for the direct sound, e.g.,

$$\widehat{P}_{\mathrm{d}}(k, n, \mathbf{p}_1) = W_{\mathrm{d}}(k, n) X_1(k, n).$$

- As for the direct sound, we can formulate for instance the Wiener filter (which here minimizes the MSE between the true and estimated diffuse sound) or the parametric Wiener filter.

- For example, in case of the square-root Wiener filter and noiseless assumption, we obtain

$$H_{\mathrm{d}}(k, n) = \sqrt{\Omega(k, n)}.$$

- This filter is used for example in DirAC (where $\Omega(k, n)$ is the diffuseness).

Signal and Parameter Estimation

Single-channel Sound Extraction: Conclusions

- Using single-channel filters for the sound extraction has specific advantages and disadvantages.
- Advantages:
  - Cheap: The filtering requires only a single microphone and estimating the filters and required parameters is usually not very complex.
  - Robust: For instance microphone positioning errors have no influence. Moreover, spectral and temporal smoothing strategies can be applied to reduce signal distortions and musical tones.
- Disadvantages:
  - In general rather poor performance in attenuating undesired signal components (e.g., direct sounds for the diffuse sound filter).

Signal and Parameter Estimation

Multi-channel Direct Sound Extraction

- A better performance compared to the single-channel direct sound extraction can be achieved using multiple microphones, for which different optimal multi-channel filters exists. For instance, for $L = 1$,

$$\widehat{P}_{\mathrm{s}}(k, n, \mathbf{p}_1) = \mathbf{w}_{\mathrm{s}}^{\mathrm{H}}(k, n)\mathbf{x}(k, n).$$

- As for the single-channel filters, the multi-channel filters are recomputed for each time and frequency with updated information on the DOA and second-order statistics (SOS) of the underlying sound field model.

- Thus, the filters can adapt fast to changing acoustics and provide a good trade-off between robustness and attenuation of undesired signals

Signal and Parameter Estimation

Multi-channel Direct Sound Extraction: Two Optimal Examples

- The linearly-constrained minimum variance (LCMV) filter minimizes the noise-plus-diffuse power and extracts the direct sound without distortion:

$$\mathbf{w}_{\mathrm{sLCMV}}(k,n) = \arg\min_{\mathbf{w}_{\mathrm{s}}} \mathbf{w}_{\mathrm{s}}^{\mathrm{H}} \left[\boldsymbol{\Phi}_{\mathrm{d}}(k,n) + \boldsymbol{\Phi}_{\mathrm{n}}(k)\right] \mathbf{w}_{\mathrm{s}}$$

$$\text{s.t.} \quad \mathbf{w}_{\mathrm{s}}^{\mathrm{H}}(k,n)\mathbf{v}(k,n) = 1.$$

- In contrast, the parametric multi-channel Wiener filter minimizes the MSE between the true and estimated direct sound subject to a distortion limit:

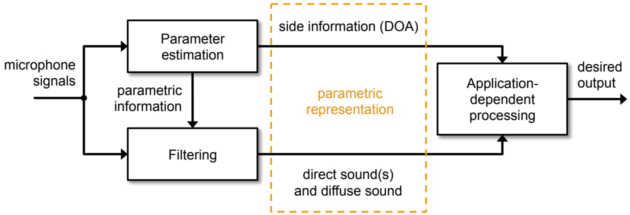
$$\mathbf{w}_{\mathrm{sPMW}}(k,n) = \arg\min_{\mathbf{w}_{\mathrm{s}}} \ \mathbf{w}_{\mathrm{s}}^{\mathrm{H}} \left[\boldsymbol{\Phi}_{\mathrm{d}}(k,n) + \boldsymbol{\Phi}_{\mathrm{n}}(k)\right] \mathbf{w}_{\mathrm{s}}$$

$$\text{s.t.} \quad \mathrm{E}\left\{ \left| \mathbf{w}_{\mathrm{s}}^{\mathrm{H}}(k,n)\mathbf{x}_{\mathrm{s}}(k,n) - P_{\mathrm{s}}(k,n,\mathbf{p}_1) \right|^2 \right\} \leq \sigma^2(k,n).$$

[Thiergart, Taseska, and Habets, 2014a]

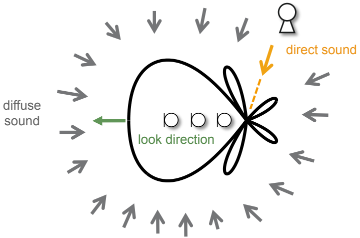
Signal and Parameter Estimation

Multi-channel Direct Sound Extraction: Automatic Trade-off



- Both filters can be computed in closed-form, which requires information on the DOA and SOS of the underlying sound field model.
- The LCMV filter provides a good trade-off between diffuse and noise attenuation depending on what undesired signal component is stronger.
- The parametric multi-channel Wiener filter provides a trade-off between signal distortions as well as noise and diffuse attenuation.

Signal and Parameter Estimation
Multi-channel Diffuse Sound Extraction



- To extract the diffuse sound, we use a spatial filter which cancels out the direct sound(s) while capturing the diffuse sound with a suitable response.
- State-of-the-art (SOA) approach: Using a spatial filter which nulls out the direct sound and captures the diffuse sound from a specific look direction.
- Advantage over single-channel filters: Instantaneous cancelation of the direct sound(s) due to the spatial null(s).

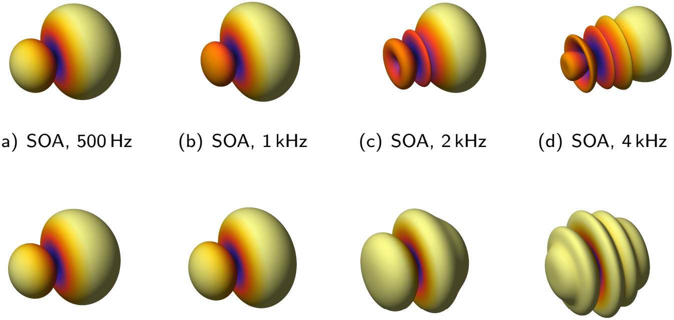
- An even better filter would capture the diffuse sound equally strong from all directions while canceling the direct sound(s).
- Such a filter can be formulated as an LCMV filter [Thiergart and Habets, 2014]:

$$\mathbf{w}_{\mathrm{dALCMV}}(k,n) = \arg\min_{\mathbf{w}} \mathbf{w}^{\mathrm{H}} \mathbf{\Phi}_{\mathrm{n}}(k) \mathbf{w}$$

$$\text{s.t.} \quad \mathbf{w}^{\mathrm{H}} \mathbf{v}(k,n) = 0 \quad \text{and} \quad \mathbf{w}^{\mathrm{H}} \, \mathrm{E}\left\{\mathbf{u}(k,n)\right\} = 1.$$

- Advantages:
  - Computing the filter requires only the DOA of the direct sound(s).
  - No (potentially sub-optimal) look direction needs to be specified.
  - The filter provides an almost omnidirectional directivity pattern with spatial nulls for the DOA of the direct sound(s).

Multi-channel Diffuse Sound Extraction



(a) SOA, 500 Hz (b) SOA, 1 kHz (c) SOA, 2 kHz (d) SOA, 4 kHz

(e) ALCMV, 500 Hz (f) ALCMV, 1 kHz (g) ALCMV, 2 kHz (h) ALCMV, 4 kHz

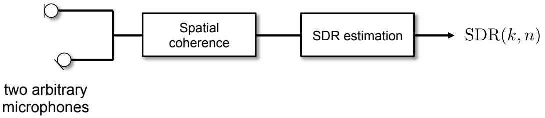
Signal and Parameter Estimation

Single/Multi-channel Sound Extraction: Conclusions

- Compared to single-channel filters, multi-channel filters can better attenuate undesired signal components (e.g., noise, undesired diffuse sounds, undesired direct sounds) while extracting the desired signal.
- The discussed multi-channel filters provide a good trade-off between signal distortions and attenuation of undesired signal components.
- Computing the filters requires the DOA of the direct sound(s) as well as SOS of the underlying parametric signal model (e.g., SDR, DNR, direct and diffuse PSDs).
- Recomputing the filters for each time and frequency with updated parametric information allows the filters to adapt quickly to changing acoustic scenes.

Signal and Parameter Estimation

Example SDR and DNR Estimator: Based on the Spatial Coherence



- One practical estimator for the SDR (assuming $L = 1$) is based on the spatial coherence between two arbitrary microphones [Thiergart, Galdo, and Habets, 2012].

- The (complex-valued) spatial coherence describes the correlation between two microphone signals in the frequency domain. It is computed as

$$\gamma_{12}(k, n) = \frac{\Phi_{x,12}(k, n)}{\sqrt{\Phi_{x,11}(k, n)}\sqrt{\Phi_{x,22}(k, n)}}.$$

$\Phi_{x,m'm}(k, n)$: cross and auto PSDs of the microphone signals

Signal and Parameter Estimation

Example SDR and DNR Estimator: Based on the Spatial Coherence

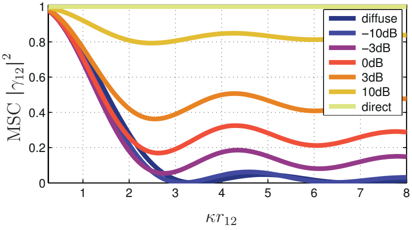


Figure: Spatial coherence (magnitude squared) as function of the SDR.

- Substituting the parametric sound field model leads to the following expression (in case of omnidirectional microphones):

$$\gamma_{12}(k,n) = \frac{\mathrm{SDR}(k,n)\gamma_{\mathrm{s},12}(k,n) + \gamma_{\mathrm{d},12}(k)}{\mathrm{SDR}(k,n) + 1}.$$

$\gamma_{\mathrm{s},12}(k,n)$: direct sound coherence, $\gamma_{\mathrm{d},12}(k)$: diffuse sound coherence

Signal and Parameter Estimation

Example SDR and DNR Estimator: Based on the Spatial Coherence

- A robust solution for the SDR is given by (omnidirectional microphones):

$$\widehat{\mathrm{SDR}}(k,n) = \mathrm{Re}\left\{ \frac{\gamma_{12}(k,n) - \gamma_{\mathrm{d},12}(k)}{e^{-\jmath \angle \Phi_{12}(k,n)} - \gamma_{12}(k,n)} \right\}.$$

- The estimator can be derived for arbitrary directional microphones as well.
- Note that the estimator is biased. Unbiased estimators which perform robust in practice were derived recently in [Schwarz and Kellermann, 2015].
- Once the SDR is estimated, it is straight-forward to compute the DNR by using the microphone signal PSD and noise PSD in the definition of the DNR presented before.

Signal and Parameter Estimation

Parameter Estimation: Examples of Further Estimators

- Estimators for the required DOA information and SOS (such as SDR, DNR, signal and diffuse PSDs) exist for almost any microphone setup.
- DOA:
  - Linear arrays: Narrowband estimators such as ESPRIT or Root MUSIC.
  - B-format microphone: Based on the active sound intensity vector as proposed in DirAC ($L = 1$), or as proposed in HARPEX ($L = 2$).
  - ...
- Direct sound PSDs and diffuse PSD:
  - Based on the power difference between multiple directional microphones ($L = 1$) [Thiergart, Ascherl, and Habets, 2014].
  - Using a quadratically-constrained null-beamformer and a least-squares approach ($L \geq 1$) [Thiergart, Taseska, and Habets, 2014a].
  - ...

Signal and Parameter Estimation

Parameter Estimation: Examples of Further Estimators

- Stationary noise PSDs: Estimated during speech pauses (detected using e.g. VAD or minimum statistics).
- Number of sources L: Assumed fixed or estimated based on the eigenvalues of the input PSD matrix (considering the minimum description length or eigenvalue ratios [Markovich, Gannot, and Cohen, 2009]).
- ...

Outline

General Overview

- The desired signal (loudspeaker or headphone signal) is defined as a weighted sum of the direct sound and diffuse sound

$$Y(k,n) = \underbrace{\sum_{l=1}^{L} G_{\mathrm{s}}(k,\varphi_l) P_{\mathrm{s},l}(k,n)}_{Y_{\mathrm{s}}(k,n)} + \underbrace{G_{\mathrm{d}}(k,n) P_{\mathrm{d}}(k,n)}_{Y_{\mathrm{d}}(k,n)}$$

- The direct weight and diffuse weight depend on the application

Application	Direct weight $G_{\mathrm{s}}(\varphi)$	Diffuse weight G_{d}
Speech enhancement	1	0
Spatial filtering	DOA-dependent spatial window	0
Spatial sound reproduction	DOA-dependent panning function for each loudspeaker	Constant factor > 0

Directional Filtering and Dereverberation

- Our goal is to provide a **desired directional gain** for L (simultaneously active) plane-waves per time and frequency while reducing both reverberation and sensor noise
- The spatial filter is controlled by nearly instantaneous information (i.e., narrowband DOAs and diffuse-to-noise ratio) to respond quickly to changes in the acoustic scene
- The proposed solution provides an **optimal tradeoff** between the white noise gain (WNG) and the directivity index
- Based on a multi-wave sound field model, the M microphone signals can be expressed as

$$\mathbf{x}(k,n) = \underbrace{\sum_{l=1}^{L} \mathbf{x}_{\mathrm{s},l}(k,n)}_{L \text{ plane waves}} + \underbrace{\mathbf{x}_{\mathrm{d}}(k,n)}_{\text{diffuse sound}} + \underbrace{\mathbf{x}_{\mathrm{n}}(k,n)}_{\text{sensor noise}}$$

Directional Filtering and Dereverberation

Problem Formulation

- The desired signal is given by

$$Y(k,n) = \sum_{l=1}^{L} G_{\mathrm{s}}(k,\varphi_l) P_{\mathrm{s},l}(k,n)$$

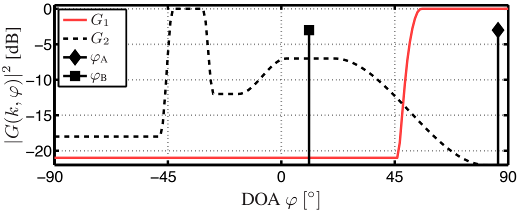


Figure: Two possible directional gain functions

- The desired signal is estimated using an informed linearly constraint minimum variance (LCMV) filter

$$\widehat{Y}(k,n) = \mathbf{w}_{\mathrm{LCMV}}^{\mathrm{H}}(k,n)\, \mathbf{x}(k,n)$$

Directional Filtering and Dereverberation

Proposed Solution (1)

- The proposed informed LCMV filter is given by

$$\mathbf{w}_{\mathrm{LCMV}} = \underset{\mathbf{w}}{\arg\min}\ \mathbf{w}^{\mathrm{H}} \left[\boldsymbol{\Phi}_{\mathrm{d}}(k,n) + \boldsymbol{\Phi}_{\mathrm{n}}(k)\right] \mathbf{w}$$
$$\text{s. t.} \quad \mathbf{w}^{\mathrm{H}}(k,n)\,\mathbf{v}(k,\varphi_l) = G_{\mathrm{s}}(k,\varphi_l), \quad l \in \{1,2,\ldots,L\}$$

where $\mathbf{v}(k,\varphi_l)$ denotes the steering vector for the lth plane wave at time m and frequency k.

Directional Filtering and Dereverberation

Proposed Solution (1)

- The proposed informed LCMV filter is given by

$$\mathbf{w}_{\mathrm{LCMV}} = \arg\min_{\mathbf{w}} \; \mathbf{w}^{\mathrm{H}} \left[\Phi_{\mathrm{d}}(k,n)\boldsymbol{\Gamma}_{\mathrm{d}}(k) + \Phi_{\mathrm{n}}(k)\,\mathbf{I}\right] \mathbf{w}$$
$$\text{s. t.} \quad \mathbf{w}^{\mathrm{H}}(k,n)\,\mathbf{v}(k,\varphi_l) = G_{\mathrm{s}}(k,\varphi_l), \quad l \in \{1,2,\ldots,L\}$$

where $\mathbf{v}(k,\varphi_l)$ denotes the steering vector for the lth plane wave at time m and frequency k.

Directional Filtering and Dereverberation

Proposed Solution (1)

- The proposed informed LCMV filter is given by

$$\mathbf{w}_{\mathrm{LCMV}} = \arg\min_{\mathbf{w}} \; \mathbf{w}^{\mathrm{H}} \left[\Phi_{\mathrm{d}}(k,n)\boldsymbol{\Gamma}_{\mathrm{d}}(k) + \Phi_{\mathrm{n}}(k)\,\mathbf{I}\right] \mathbf{w}$$
$$\text{s.t.} \quad \mathbf{w}^{\mathrm{H}}(k,n)\,\mathbf{v}(k,\varphi_l) = G_{\mathrm{s}}(k,\varphi_l), \quad l \in \{1,2,\ldots,L\}$$

  where $\mathbf{v}(k,\varphi_l)$ denotes the steering vector for the lth plane wave at time m and frequency k.

- For the assumed signal model, we can alternatively minimize

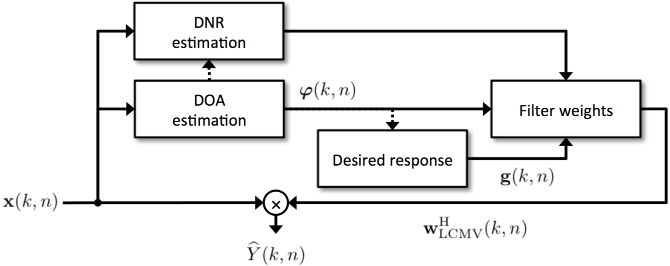
$$\mathbf{w}^{\mathrm{H}} \left[\mathrm{DNR}(k,n)\,\boldsymbol{\Gamma}_{\mathrm{d}}(k) + \mathbf{I}\right] \mathbf{w},$$

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

- The filter is computed for each time and frequency given the parametric information (i.e., DOAs and DNR). For more information see [Thiergart, Taseska, and Habets, 2014b]).

Directional Filtering and Dereverberation

Proposed Solution (2)



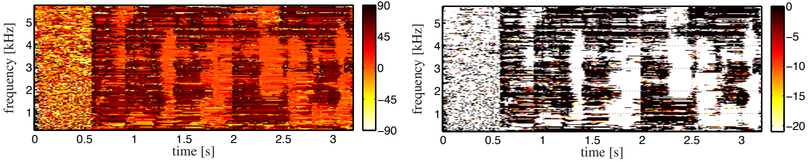
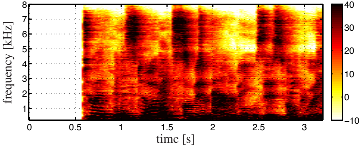


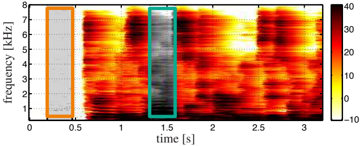
Figure: Left: DOA $\varphi_1(k,n)$ as a function of time and frequency. Right: Desired response $|G_{\mathrm{s}}(k,\varphi_1)|^2$ in dB for DOA $\varphi_1(k,n)$ as a function of time and frequency.

Directional Filtering and Dereverberation

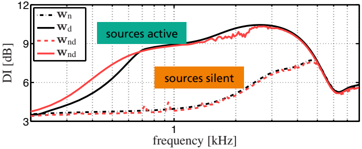
Results (1)



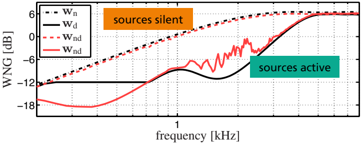
(a) True



(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 LCMV filter that minimizes the diffuse plus noise power [shown when the sources are active (red solid line) and silent (red dashed line)].

Directional Filtering and Dereverberation

Results (2)

- 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**)

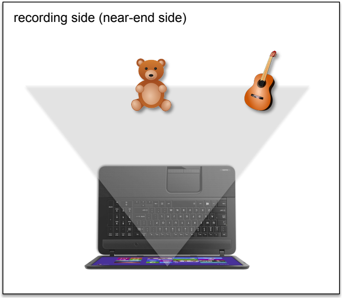
Table: Performance of all spatial filters [$*$ unprocessed, first sub-column using true DOAs (of the sources), second sub-column using estimated DOAs (of the plane waves)].

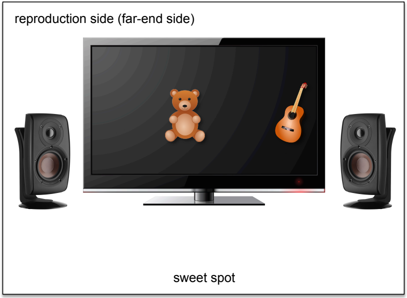
Audiovisual Demo

https://www.audiolabs-erlangen.de/fau/professor/habets/demos

Acoustical Zoom

Motivation

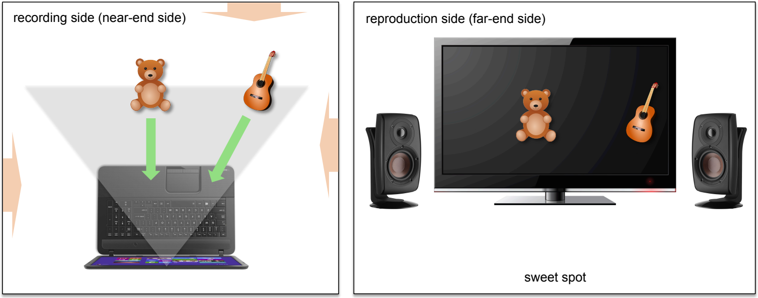




- Aim of the proposed approach: Recording and reproduction of the original spatial sound such that it is consistent with the video

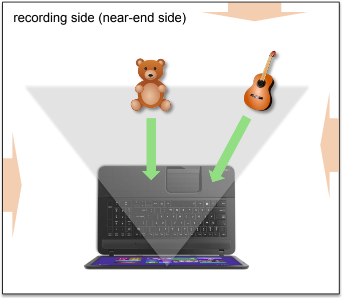
Acoustical Zoom

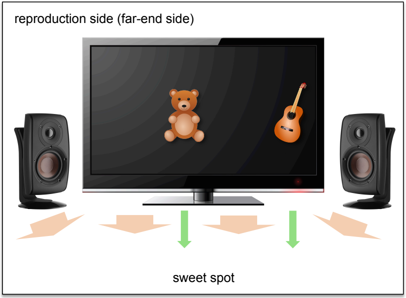
Motivation



- Aim of the proposed approach: Recording and reproduction of the original spatial sound such that it is consistent with the video

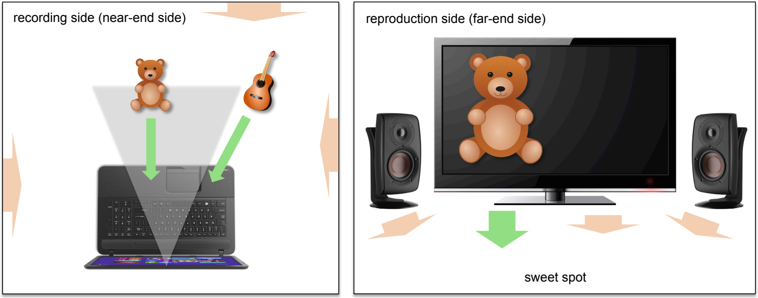
Acoustical Zoom
Motivation





- Aim of the proposed approach: Recording and reproduction of the original spatial sound such that it is consistent with the video

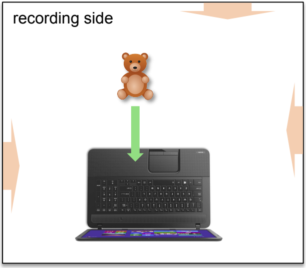
Acoustical Zoom
Motivation



- Aim of the proposed approach: Recording and reproduction of the original spatial sound such that it is consistent with the video

Acoustical Zoom

Sound Field Model



- We assume that for each time-frequency instant the sound is composed of a single plane wave plus diffuse sound

$$\mathbf{x}(k,n) = \underbrace{\mathbf{x}_{\mathrm{s}}(k,n)}_{\text{plane wave}} + \underbrace{\mathbf{x}_{\mathrm{d}}(k,n)}_{\text{diffuse sound}} + \underbrace{\mathbf{x}_{\mathrm{n}}(k,n)}_{\text{sensor noise}}$$

- The plane wave models the direct sound of the sources while the diffuse sound models the reverberation

Acoustical Zoom

Sound Field Model

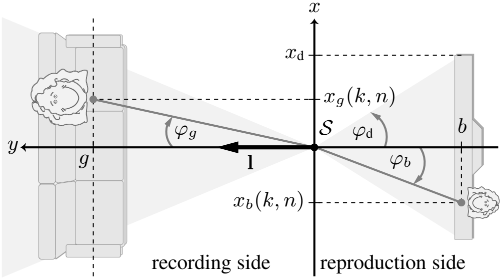
- The desired signal is a weighted sum of the direct sound and diffuse sound at the reference microphone

$$Y_q(k,n) = G_{\mathrm{s},q}(k,\varphi)P_{\mathrm{s}}(k,n) + Q(k)\,P_{\mathrm{d}}(k,n)$$

- The gains $G_{\mathrm{s},q}$ and Q are used to align the visual and acoustical image and to create the acoustical zoom. The gains are adjusted when zooming.

- The gain $G_{\mathrm{s},q}$ assures that the direct sound is reproduced from the correct direction, while the gain Q controls the output signal-to-diffuse ratio.

Acoustical Zoom
Zoom Parameters



- Since the optical system is linear, the DOA of the direct sound at the recording side is related to the DOA at the reproduction side by

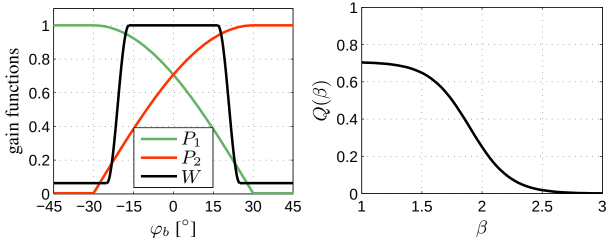
$$\tan \varphi_b(k,n) = \beta\, c \tan \varphi_g(k,n)$$

- Here, β is the zoom factor and c is a constant depending on the screen size, position of the sweet spot, and distance of the source

Acoustical Zoom

Zoom Parameters

- Example gain functions for the direct sound and diffuse sound:



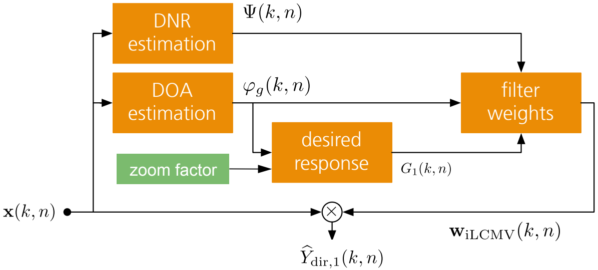
(a) direct sound (b) diffuse sound

- The direct sound gain is the product of a panning function and window function and depends on the DOA of the direct sound at the reproduction side. The diffuse sound gain depends on the zoom factor.

Acoustical Zoom

Direct and Diffuse Sound Estimation

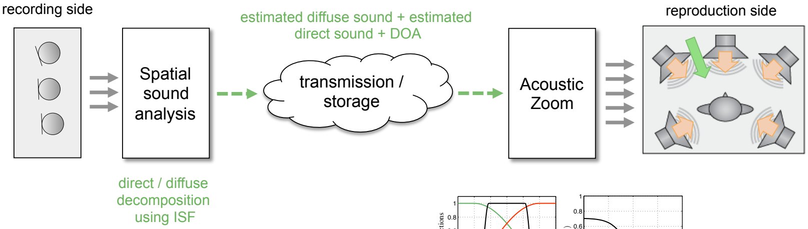
- The desired signals Y_{dir} and Y_{diff} are estimated using so-called informed spatial filters (ISFs). The desired direct signal can be estimated as follows:

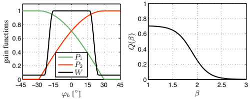


- The diffuse sound can be estimated using the diffuse beamformer explained earlier. More details can also be found in [Thiergart, Kowalczyk, and Habets, 2014].

Acoustical Zoom

System Overview



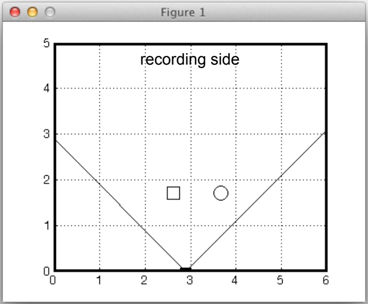


(a) direct sound (b) diffuse sound

Acoustical Zoom

Demo

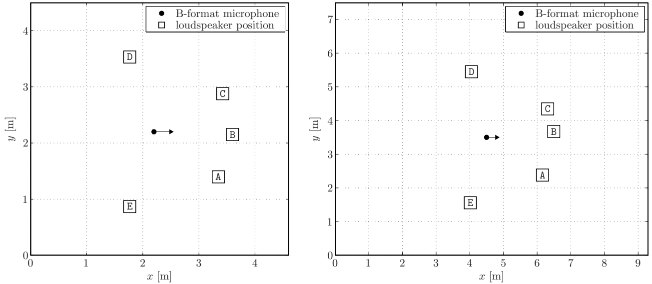
- Simulated shoebox room (RT60 = 270ms), single talk and double talk
- Sound captured with a linear array with 6 noisy microphones
- DOAs estimated with ESPRIT



Application Examples

Spatial Sound Recording and Reproduction

- Several scenarios were recorded using a B-format microphones



(a) Room 1 ($RT_{60} \approx 110\,\mathrm{ms}$)   (b) Room 2 ($RT_{60} \approx 390\,\mathrm{ms}$)

- Processing using the informed spatial filtering scheme
- Sound reproduction using a 5.1 surround sound setup

- We aim at reproducing the sound at the reproduction side with the same spatial impression as on the recording side

- The q-th loudspeaker signal are given by

$$\begin{aligned} Y_q(k,n) &= \sum_{l=1}^{L} G_{\mathrm{s},q}(k,\varphi_l) P_{\mathrm{s},l}(k,n) + G_{\mathrm{d},q}(k,n) P_{\mathrm{d}}(k,n) \\ &= Y_{\mathrm{s},q}(k,n) + Y_{\mathrm{d},q}(k,n) \end{aligned}$$

- The weights for the direct sound are selected from a panning function

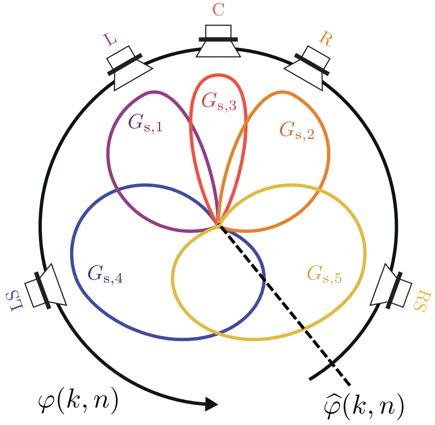
- The weights for the diffuse sound are fixed

$$G_{\mathrm{d},q}(k,n) = \sqrt{\frac{1}{Q}}$$

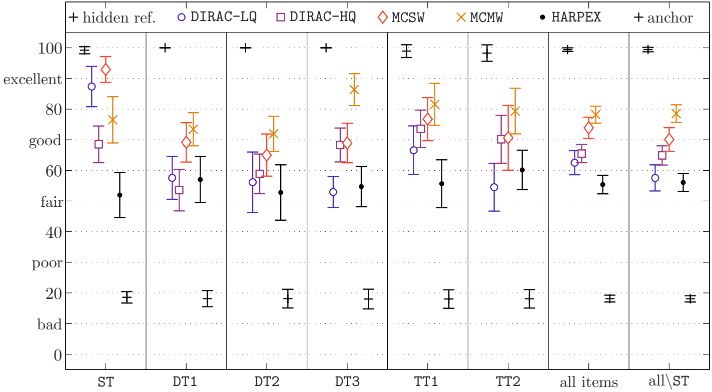
where Q is the number of loudspeakers.

Application Examples
Spatial Sound Recording and Reproduction

- We consider the vector-base amplitude panning (VBAP) function to select the direct sound weights depending on the estimated DOA

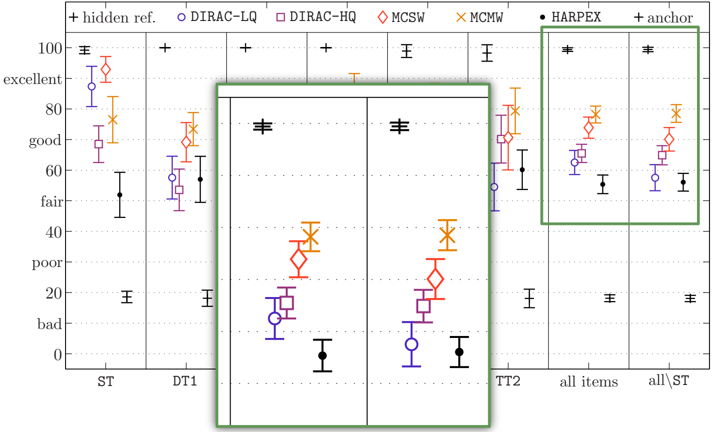


Application Examples
Spatial Sound Recording and Reproduction



Application Examples

Spatial Sound Recording and Reproduction



Outline

Summary and Outlook

Summary

- Parametric spatial audio processing relies on a simple yet powerful description of the sound-field.
- Accurate estimation of the parameters as well as the estimation of the direct and diffuse sound signal is of paramount importance.
- Several applications have been developed over the last few years.
- Using this approach we were able to perform robust, flexible and efficient spatial audio processing.

Summary and Outlook
Outlook

- In some cases the sound field model is violated, for example due to early reflections. Research towards more sophisticated models is ongoing.
- Especially in adverse environments (low SNR and low SDR) the parameter estimation remains a challenging task. Further research is needed to develop estimators that are even more accurate in such challenging scenarios.
- The framework allows to include additional perceptual information into the design of the desired spatial response.
- We are exploiting new applications, for example, in the areas of virtual and augmented reality.
- Be creative...

Acknowledgments

- Giovanni Del Galdo
- Konrad Kowalczyk
- Maja Taseska
- Sebastian Braun

References I

Berge, Svein and Natasha Barrett (2010). "A new method for B-format to binaural transcoding". In: **Audio Engineering Society Conference: 40th International Conference: Spatial Audio**. Tokyo, Japan.

Cook, Richard K. et al. (1955). "Measurement of Correlation Coefficients in Reverberant Sound Fields". In: **The Journal of the Acoustical Society of America** 27.6, pp. 1072–1077.

Crochiere, R. E. and L. R. Rabiner (1983). **Multirate Digital Signal Processing**. Englewood Cliffs, New Jersey, USA: Prentice-Hall.

Del Galdo, G. et al. (2012). "The diffuse sound field in energetic analysis". In: **The Journal of the Acoustical Society of America** 131.3, pp. 2141–2151.

Habets, E. A. P., S. Gannot, and I. Cohen (2009). "Late Reverberant Spectral Variance Estimation Based on a Statistical Mode". In: **IEEE Signal Process. Lett.** 16.9, pp. 770–774. DOI: 10.1109/LSP.2009.2024791.

Kollmeier, B., J. Peissig, and V. Hohmann (1993). "Binaural noise-reduction hearing aid scheme with real-time processing in the frequency domain". In: **Scandinavian Audiology** 22, pp. 28–38.

References II

Laitinen, M.-V. and V. Pulkki (2012). "Utilizing Instantaneous Direct-to-Reverberant Ratio in Parametric Spatial Audio Coding". In: **Audio Engineering Society Convention 133**.

Markovich, S., S. Gannot, and I. Cohen (2009). "Multichannel Eigenspace Beamforming in a Reverberant Noisy Environment With Multiple Interfering Speech Signals". In: **Audio, Speech, and Language Processing, IEEE Transactions on** 17.6, pp. 1071–1086.

Pulkki, Ville (2007). "Spatial Sound Reproduction with Directional Audio Coding". In: **J. Audio Eng. Soc** 55.6, pp. 503–516.

Schwarz, A. and W. Kellermann (2015). "Coherent-to-Diffuse Power Ratio Estimation for Dereverberation". In: **Audio, Speech, and Language Processing, IEEE/ACM Transactions on** 23.6, pp. 1006–1018.

Sugiyama, A. and R. Miyahara (2015). "A directional noise suppressor with a specified beamwidth". In: **Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)**.

References III

Tashev, I. and A. Acero (2006). "Microphone Array Post-Processor Using Instantaneous Direction of Arrival". In: **Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)**.

Thiergart, O., T. Ascherl, and E. A. P. Habets (2014). "Power-based Signal-to-Diffuse Ratio Estimation using Noisy Directional Microphones". In: **Acoustics Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on**.

Thiergart, O., G. Del Galdo, and E. A. P. Habets (2012). "On the spatial coherence in mixed sound fields and its application to signal-to-diffuse ratio estimation". In: **The Journal of the Acoustical Society of America** 132.4, pp. 2337–2346.

Thiergart, O. and E. A. P. Habets (2012). "Sound field model violations in parametric spatial sound processing". In: **Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)**. Aachen, Germany.

– (2014). "Extracting Reverberant Sound Using a Linearly Constrained Minimum Variance Spatial Filter". In: **Signal Processing Letters, IEEE** 21.5, pp. 630–634.

References IV

Thiergart, O., K. Kowalczyk, and E. A. P. Habets (2014). "An acoustical zoom based on informed spatial filtering". In: **Proc. of the International Workshop on Acoustic Signal Enhancement (IWAENC)**. Juan-les-Pins, France: IEEE, pp. 109–113. DOI: 10.1109/IWAENC.2014.6953348.

Thiergart, O., M. Taseska, and E. A. P. Habets (2014a). "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates". In: **IEEE/ACM Trans. Acoust., Speech, Signal Process.** 22.12.

– (2014b). "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates". In: **IEEE Trans. Audio, Speech, Lang. Process.** 22.12, pp. 2182–2196. DOI: 10.1109/TASLP.2014.2363407.

Wittkop, T and V Hohmann (2003). "Strategy-selective noise reduction for binaural digital hearing aids". In: **Speech Communication** 39, pp. 111–138.