Incorporating sparsity into multi-microphone speech dereverberation techniques

Prof. Dr. Simon Doclo

University of Oldenburg, Germany

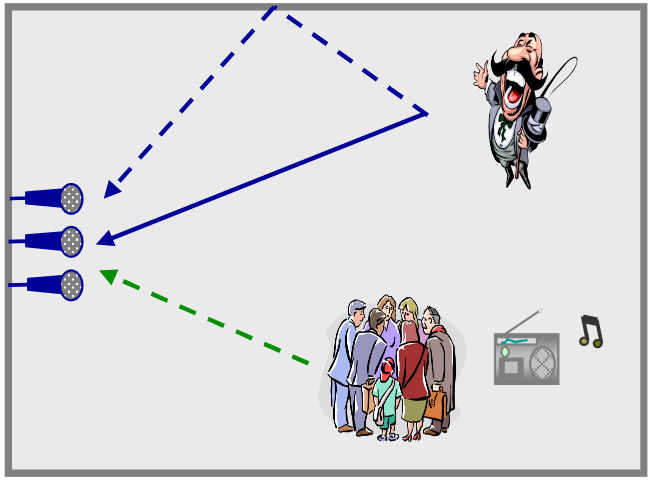
Dept. of Medical Physics and Acoustics, Cluster of Excellence Hearing4All

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

ICSEE 2016 Symposium on Speech and Audio Processing

Dereverberation and noise reduction

- **Problem**
  - Noise and reverberation jointly present in typical acoustic environments
  - Speech quality and intelligibility degradation
  - Performance degradation of ASR systems



- **Objectives**
  - Develop single- and microphone joint dereverberation and noise reduction algorithms
  - Exploit knowledge / statistical models of room acoustics and speech signals
- **This presentation:**
  - Focus on **multi-microphone dereverberation**
  - Two classes of techniques:
    - Acoustic multi-channel equalization (*non-blind, time-domain*)
    - Multi-channel linear prediction (*blind, frequency-domain*)
  - **Incorporate sparsity of clean speech TF coefficients into both techniques**

Signal model

- **Scenario:** speech source in noisy and reverberant environment, *M* microphones
- **Time-domain model:** "perfect" model

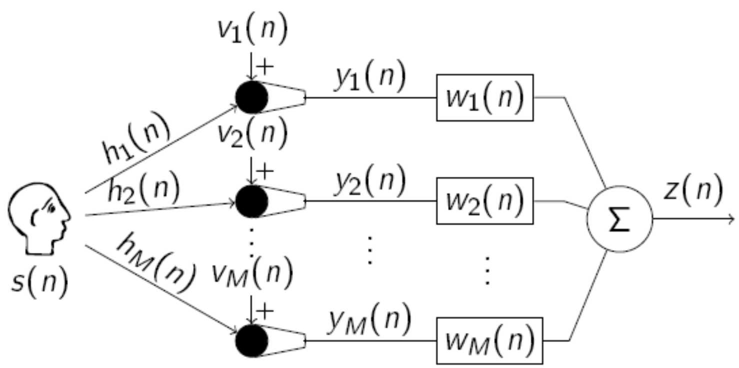
$$y_m(n) = x_m(n) + v_m(n) = s(n) * h_m(n) + v_m(n)$$

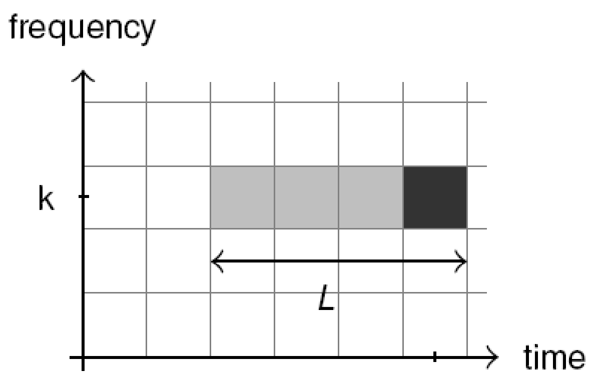
$h_m(n)$ = room impulse response (RIR), typically long and difficult to blindly estimate

- **STFT-domain model:** approximation of time-domain model

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

$h_m(k,n)$ = convolutive transfer function (CTF) in frequency bin *k* and time frame *n*





Acoustic multi-channel equalization

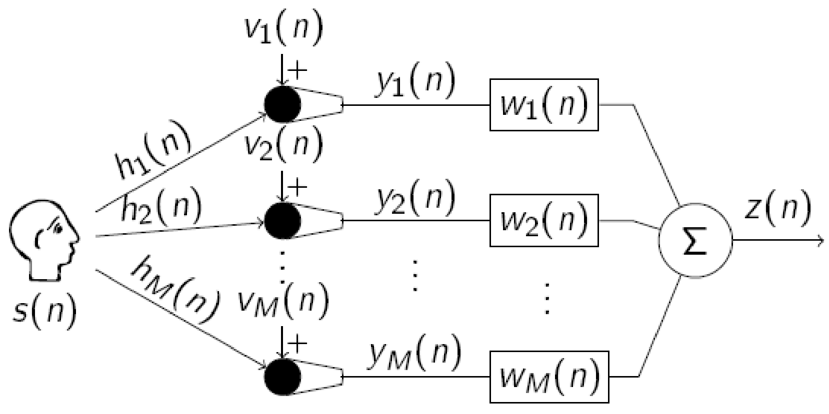
Outline

- **Acoustic multi-channel equalization** for speech dereverberation:
  - State-of-the-art time-domain approaches (RMCLS, P-MINT)
  - Very sensitive to RIR perturbations

- **Increase robustness by:**
  1. Decreasing filter length
  2. Signal-independent regularization
  3. **Signal-dependent regularization, enforcing sparsity of output signal**

Acoustic multi-channel equalization

- **Time-domain approach** *(although frequency-domain versions possible)*
- **Indirect approach*:***
  1. estimate/measure RIRs
  2. Estimate the clean speech signal by inverting/equalizing the acoustic system + suppressing noise



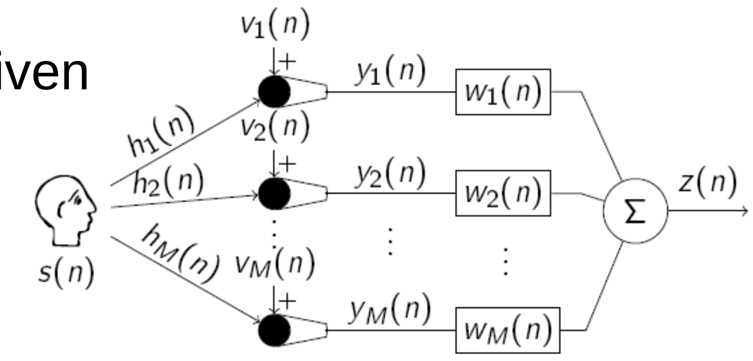
$$z(n) = \underbrace{\mathbf{w}^T \mathbf{H}^T}_{\mathbf{c}^T} \mathbf{s}(n) + \mathbf{w}^T \mathbf{v}(n)$$

Speech enhancement objectives

- Dereverberation: Optimize $\mathbf{c}$
- Noise reduction: Minimize the noise output power while controlling the speech distortion
- Joint dereverberation and noise reduction: Optimize $\mathbf{c}$ and minimize the noise output power

Acoustic multi-channel equalization

- Disregard additive noise and aim **only at dereverberation**
- Assumptions:
  - Measurements or estimates of RIRs **H** are given
  - Reshaping filter length is $L_w \geq \lceil \frac{L_h-1}{M-1} \rceil$
  - RIRs do not share any common zeros



In theory perfect dereverberation performance

Optimize the **true** equalized impulse response

$$\mathbf{H}\mathbf{w} = \mathbf{c}_t$$

$\mathbf{c}_t$ = user-defined dereverberated target response (delayed impulse, early reflections, …)

In practice large distortions due to RIR perturbations

Optimize the **perturbed** equalized impulse response

$$\hat{\mathbf{H}}\mathbf{w} = \mathbf{c}_t$$

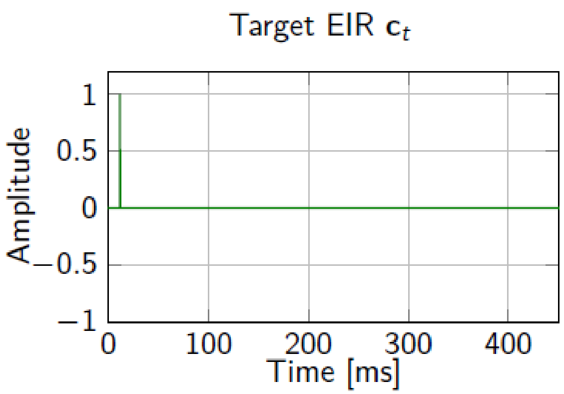
State-of-the-art acoustic multi-channel equalization

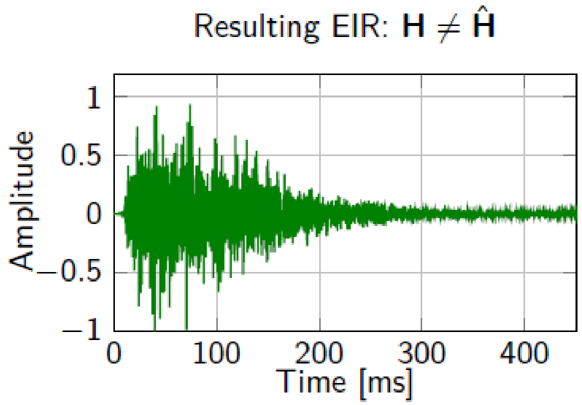
Optimize the equalized impulse response by minimizing

$$J_{LS} = \|\mathbf{W}(\hat{\mathbf{H}}\mathbf{w} - \mathbf{c}_t)\|_2^2 \qquad \mathbf{w}_{LS} = (\mathbf{W}\hat{\mathbf{H}})^+(\mathbf{W}\mathbf{c}_t)$$

Multiple-input/output inverse theorem (MINT)

Aim: Suppress all reflections





- Analytical solution
- Perceptual speech quality preservation
- Sensitivity to RIR perturbations

[Myoshi and Kaneda, IEEE ASSP, 1998]

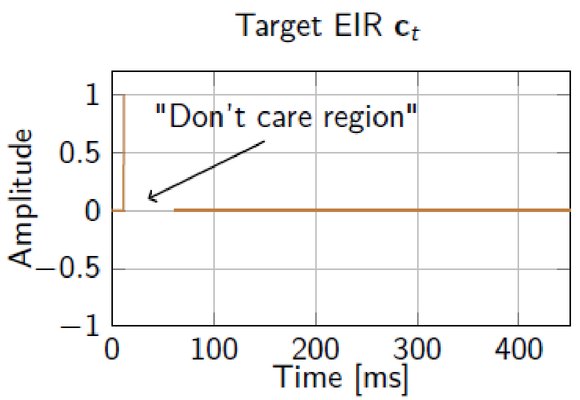
State-of-the-art acoustic multi-channel equalization

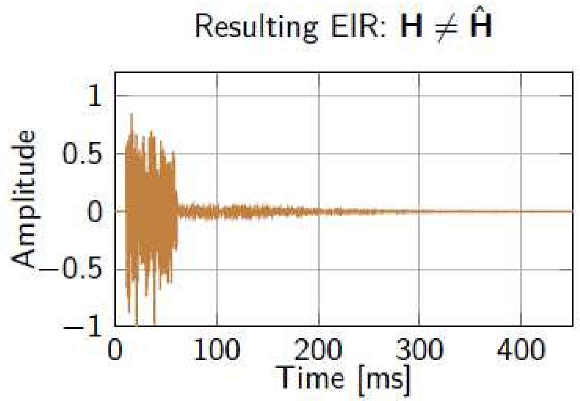
Optimize the equalized impulse response by minimizing

$$J_{LS} = \|\mathbf{W}(\hat{\mathbf{H}}\mathbf{w} - \mathbf{c}_t)\|_2^2 \qquad \mathbf{w}_{LS} = (\mathbf{W}\hat{\mathbf{H}})^+(\mathbf{W}\mathbf{c}_t)$$

Relaxed multi-channel least-squares (RMCLS)

Aim: Suppress only late reflections while not constraining early reflections





- Analytical solution
- No guaranteed perceptual speech quality preservation
- Lower sensitivity to RIR perturbations

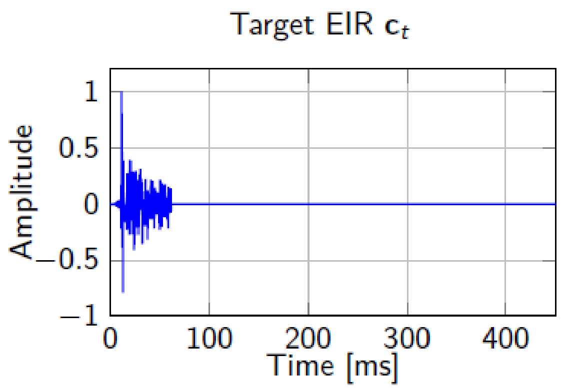
State-of-the-art acoustic multi-channel equalization

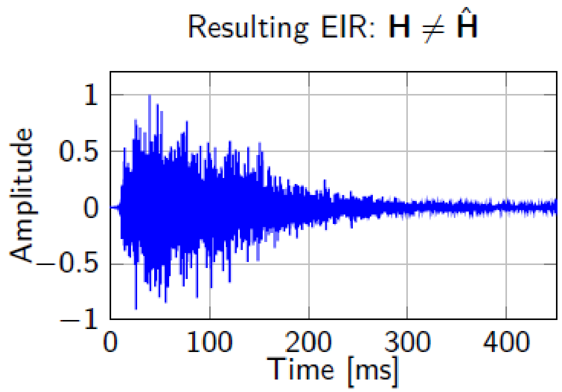
Optimize the equalized impulse response by minimizing

$$J_{LS} = \|\mathbf{W}(\hat{\mathbf{H}}\mathbf{w} - \mathbf{c}_t)\|_2^2 \qquad \mathbf{w}_{LS} = (\mathbf{W}\hat{\mathbf{H}})^+(\mathbf{W}\mathbf{c}_t)$$

Partial multi-channel equalization based on MINT (PMINT)

Aim: Suppress only late reflections while constraining early reflections





- Analytical solution
- Perceptual speech quality preservation
- Sensitivity to RIR perturbations

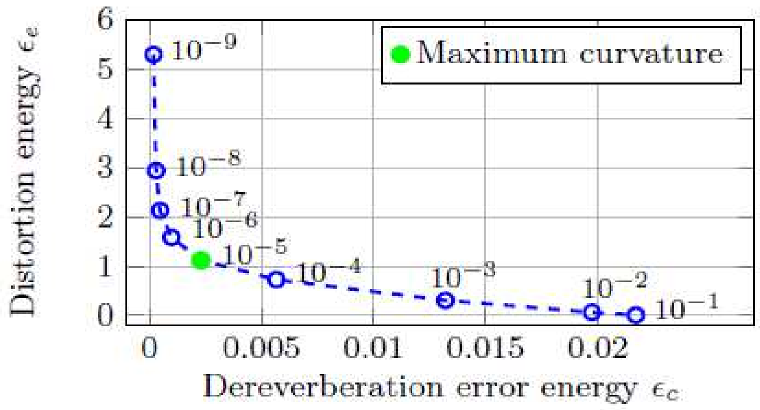
Robust acoustic multi-channel equalization

- **Increase robustness by:**
  1. *Decreasing filter length:* better conditioned optimization criterion
  2. *Signal-independent regularization*: control distortion energy due to RIR perturbations

$$J = \underbrace{\|\mathbf{W}(\hat{\mathbf{H}}\mathbf{w} - \mathbf{c}_t)\|_2^2}_{\epsilon_c} + \delta \underbrace{\mathbf{w}^T \mathbf{R}_e \mathbf{w}}_{\epsilon_e}$$

with $\mathbf{E} = \hat{\mathbf{H}} - \mathbf{H}$ and $\mathbf{R}_e = \mathcal{E}\{\mathbf{E}^T\mathbf{E}\}$ constructed using a statistical model

- **Automatic procedure for selecting regularization parameter** δ (based on L-curve), yielding both low dereverberation error energy and distortion energy

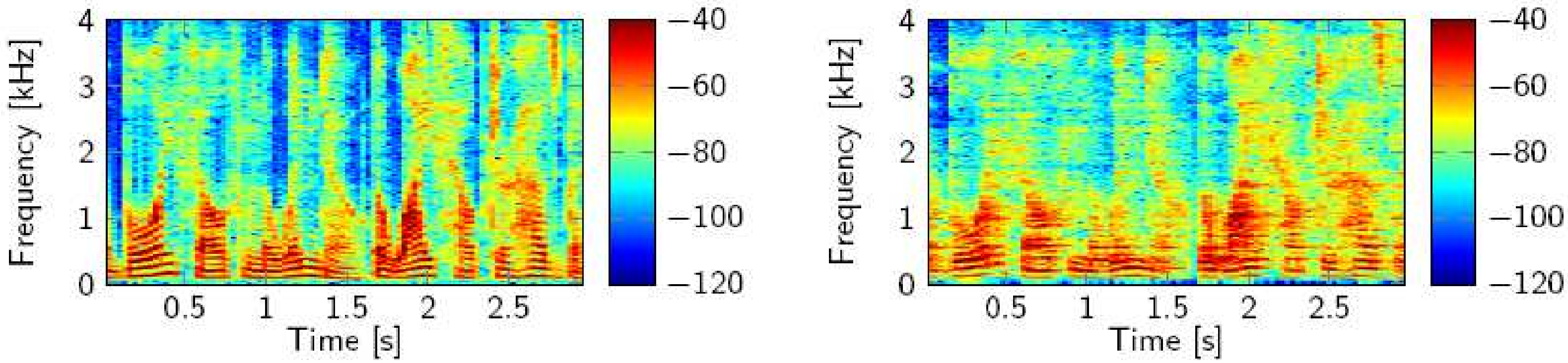


  3. ***Signal-dependent regularization*: enforce output signal to exhibit characteristics of clean signal (e.g., sparsity)**

$$\min_{\mathbf{w}} \left[\|\mathbf{W}(\hat{\mathbf{H}}\mathbf{w} - \mathbf{c}_t)\|_2^2 + \eta f_{\text{sp}}(\mathbf{z}(n))\right]$$

Sparsity-promoting multi-channel equalization

- **STFT-domain:** clean speech is more sparse than reverberant speech



- **Aim:** optimize the equalized impulse response and enforce sparsity on the output signal STFT coefficients

$$\min_{\mathbf{w}} \left[\|\mathbf{W}(\hat{\mathbf{H}}\mathbf{w} - \mathbf{c}_t)\|_2^2 + \eta f_{\text{sp}}(\mathbf{z}(n)) \right]$$

- **Select** f_{sp} as a function which promotes sparsity of the STFT coefficients of the output signal, i.e.

$$\tilde{\mathbf{z}} = \boldsymbol{\Psi}\mathbf{z} = \boldsymbol{\Psi}\mathbf{X}\mathbf{w}$$

with $\boldsymbol{\Psi}$ denoting STFT operator

Sparsity-promoting multi-channel equalization

$$\min_{\mathbf{w}} \left[\|\mathbf{W}(\hat{\mathbf{H}}\mathbf{w} - \mathbf{c}_t)\|_2^2 + \eta f_{\mathrm{sp}}(\mathbf{z}(n)) \right]$$

- Commonly used **sparsity-promoting norms**

l_0-norm:	$\|\tilde{\mathbf{z}}\|_0 = \lvert q : \tilde{z}(q) \neq 0 \rvert$
l_1-norm:	$\|\tilde{\mathbf{z}}\|_1 = \sum_{q=0}^{L_{\tilde{z}}-1} \lvert \tilde{z}(q) \rvert$
weighted l_1-norm:	$\|\mathrm{diag}\{\mathbf{u}\}\tilde{\mathbf{z}}\|_1 = \sum_{q=0}^{L_{\tilde{z}}-1} \lvert u(q)\tilde{z}(q) \rvert$

- Selecting weights $u(q)$
  - *Ideally*: STFT coefficients of clean speech signal $u(q) = \frac{1}{|\tilde{s}(q)|+\zeta}$
  - *In practice*: STFT coefficients of a reverberant microphone signal $u(q) = \frac{1}{|\tilde{x}_1(q)|+\zeta}$
- No closed-form analytical solution
- Iterative optimization using the alternating direction method of multipliers (ADMM)

Experimental results

Simulation parameters

- $T_{60} \approx 610$ ms, M = 4, fs :
- RIR perturbation levels:

$$\mathrm{NPM} = 10\log_{10}\frac{\left|\mathbf{h}_m - \frac{\mathbf{h}_m^T\hat{\mathbf{h}}_m}{\hat{\mathbf{h}}_m^T\hat{\mathbf{h}}_m}\hat{\mathbf{h}}_m\right|_2^2}{\|\mathbf{h}_m\|_2^2}\,[\mathrm{dB}]$$

ADMM parameters

- STFT: 32 ms Hamming window with 50% overlap
- Initialization $\mathbf{w}^{(0)} = [1\ 0 \ldots 0]^T$ (first microphone signal)
- Number of iterations: 500

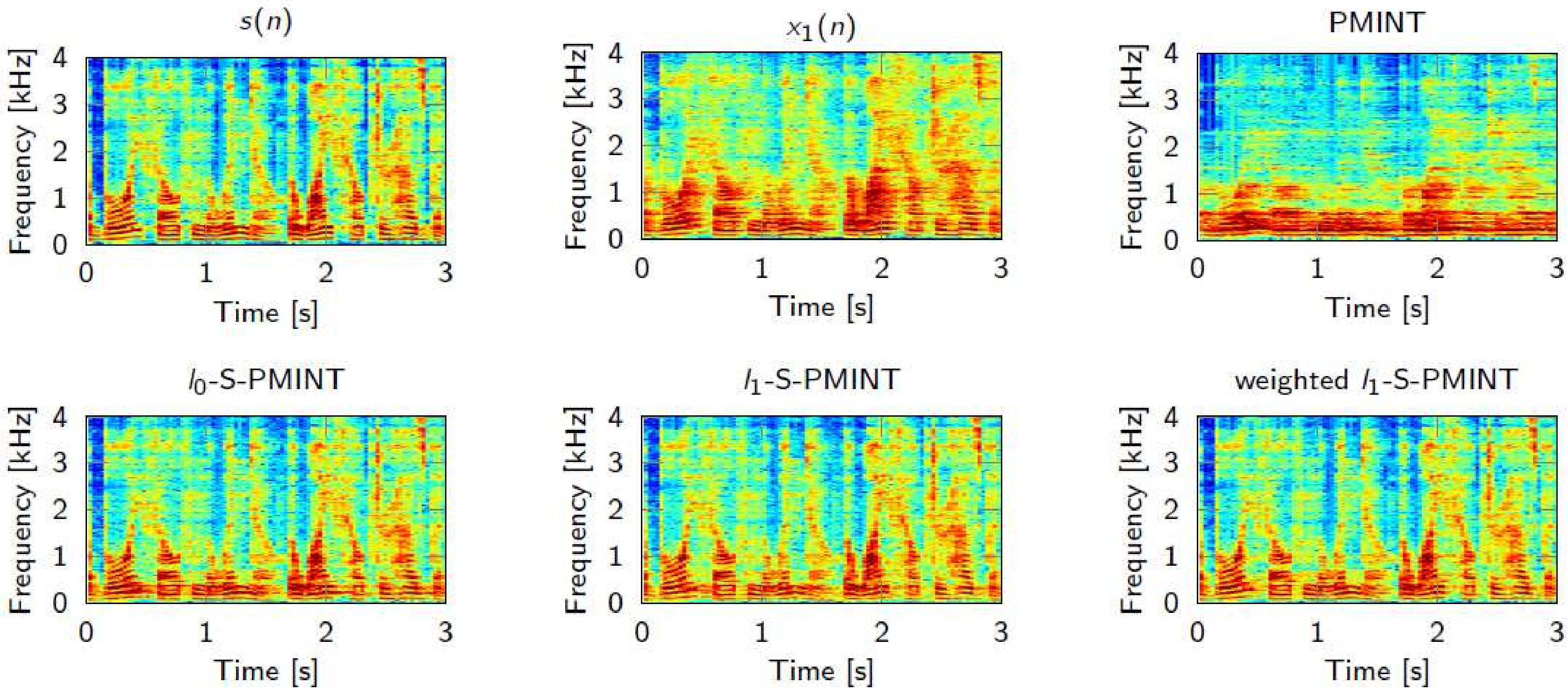
Performance measures

- Direct-to-reverberant ratio (DRR)
- Cepstral distance (CD)
- Perceptual evaluation of speech quality (PESQ)

Regularization parameters (ρ,η) intrusively selected as the parameters minimizing cepstral distance

Experimental results

Exemplary spectrograms (NPM = -33 dB)

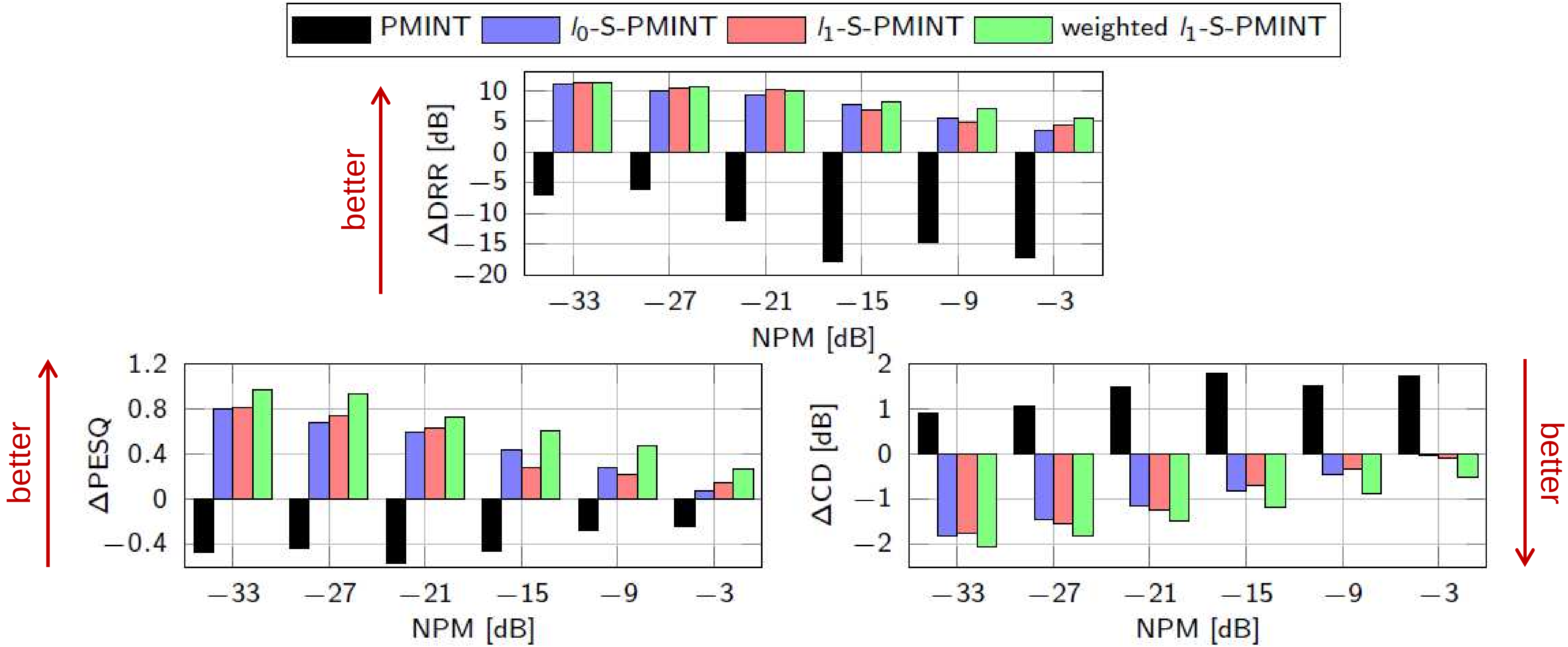


Sparsity-promoting penalty functions suppress

- reverberant energy
- distortions introduced by the non-robust PMINT technique

Experimental results

Performance measures (different NPMs)



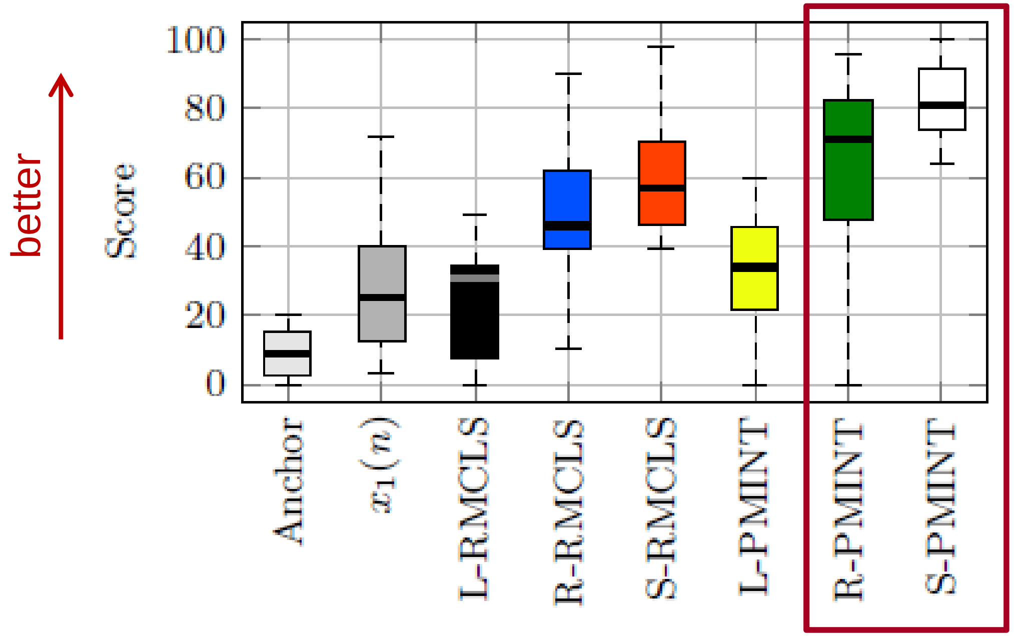
- All sparsity-promoting norms increase robustness against RIR perturbations
- **Weighted l_1–norm yields best performance (especially for large NPM)**

[Kodrasi, Jukić, Doclo, ICASSP 2016]

Experimental results

Perceptual validation (NPM = -33 dB)

- 13 self-reported normal hearing subjects
- MUSHRA test, evaluating "overall speech quality" on a scale from 0 to 100



Clean
Reverberant
R-RMCLS
R-PMINT
S-RMCLS
S-PMINT

- Robust PMINT extensions outperform robust RMCLS extensions
- **Sparsity-promoting PMINT best speech quality for moderate NPMs**

[Kodrasi, Cauchi, Goetze, Doclo, JAES, in press]

Joint dereverberation and noise reduction

- Equalization techniques for dereverberation lead to **noise amplification**
- **Cost functions for joint dereverberation and noise reduction:**
  1. Incorporate **noise statistics** into regularized P-MINT (RPM-DNR)

$$J = \underbrace{\|\hat{\mathbf{H}}\mathbf{w} - \hat{\mathbf{h}}_1^d\|_2^2}_{\epsilon_c} + \delta \underbrace{\mathbf{w}^T\mathbf{R}_e\mathbf{w}}_{\epsilon_e} + \mu \underbrace{\mathbf{w}^T\mathbf{R}_v\mathbf{w}}_{\epsilon_v}$$

  2. Incorporate **speech statistics** → Multi-channel Wiener Filter, using dereverberated output signal of regularized P-MINT as reference signal (MWF-DNR)

$$J = \mathcal{E}\{(\mathbf{w}^T\mathbf{x}(n) - \mathbf{w}_{RP}^T\mathbf{x}(n))^2\} + \mu\mathcal{E}\{(\mathbf{w}^T\mathbf{v}(n))^2\}$$

- Automatic selection of trade-off parameter(s)

$y_1(n)$	PMINT	R-PMINT	RPM-DNR	MWF-DNR

Measure	PMINT	RPMINT	RPM-DNR	MWF-DNR
ΔDRR [dB]	−3.3	**9.9**	9.8	9.1
ΔPESQ	−0.4	**0.7**	**0.7**	0.6
ψ_{NR} [dB]	−26.8	1.9	3.2	**13.0**
ΔfwSSNR [dB]	−3.0	0.9	1.1	**3.2**

M=4, T60=610 msec, DRR=-2 dB, fs=8 kHz, NPM=-33 dB, SIR=0 dB, SNR=10 dB (diffuse noise), no estimation errors in correlation matrices

Blind probabilistic model-based approach

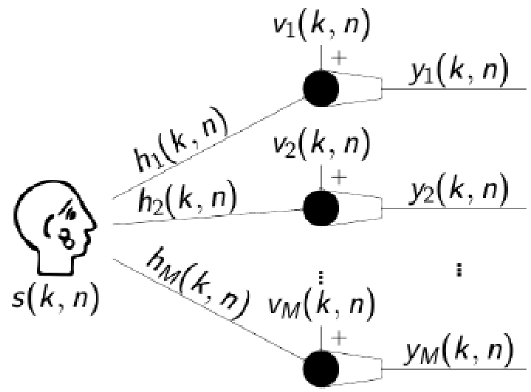
Outline

- **Multi-channel Linear Prediction** (MCLP) for speech dereverberation:
  - Conventional approach using time-varying Gaussian (TVG) model
  - Generalization using **circular sparse prior**
  - (Batch processing, single output signal, frequency-independent processing)
- **Extensions:**
  1. Exploit **low-rank structure** of speech spectrogram (NMF)
  2. MIMO speech dereverberation based on **group sparsity**
  3. **Adaptive MCLP** with robustness constraints
  4. **General framework** for incorporating time-frequency domain sparsity

Multi-channel linear prediction (MCLP)

- **STFT-domain approach** *(although time-domain versions possible)*
  - Speech properties (e.g., sparsity) can be modelled more naturally in STFT-domain
  - Low computational complexity (independent frequency bin processing)

- **Direct approach:** directly estimate clean speech STFT coefficients *s(k,n)* from reverberant (and noisy) STFT coefficients $y_m(k,n)$



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

1. *Directly using CTF model → sparse Bayesian deconvolution based on variational Bayesian inference*

2. Transform to equivalent AR model → **multi-channel linear prediction (MCLP)**

$$x_1(k,n) = d(k,n) + \sum_{m=1}^{M} \sum_{l=0}^{L_g-1} g_m(k,l) x_m(k,n-\tau-l)$$

clean signal (incl. early reflections) **prediction filters** **delay (early reflections)**

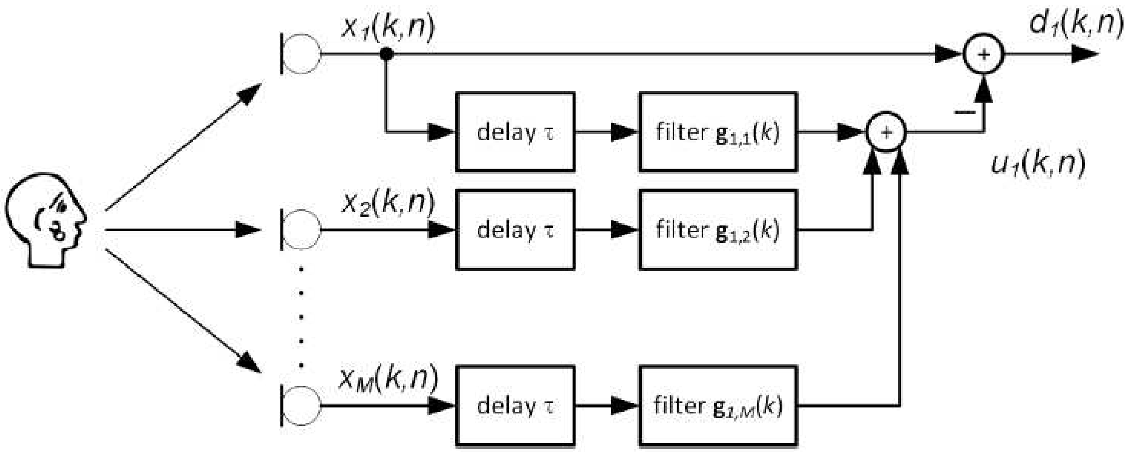
Multi-channel linear prediction (MCLP)

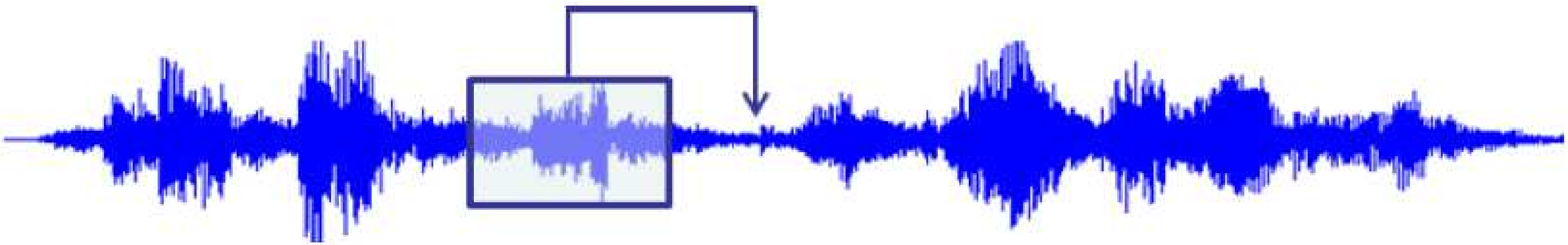
- **AR model of reverberant speech**

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

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

predicted reverberation





How to select suitable cost function for prediction filters ?

Multi-channel linear prediction (MCLP)

- **Conventional approach:**
  - STFT coefficients of desired signal are assumed to be independent and modelled using **circular complex Gaussian distribution with time-varying variance** $\lambda(k,n)$

$$\mathcal{N}_{\mathbb{C}}(d(k,n);0,\lambda(k,n)) = \frac{1}{\pi\lambda(k,n)} e^{-\frac{|d(k,n)|^2}{\lambda(k,n)}}$$

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

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

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

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

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

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

Multi-channel linear prediction (MCLP)

- **Generalization:**
  - STFT coefficients of desired signal are assumed to be independent and modelled using **circular sparse/super-Gaussian prior with time-varying variance** $\lambda(n)$

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

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

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

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

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

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

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

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

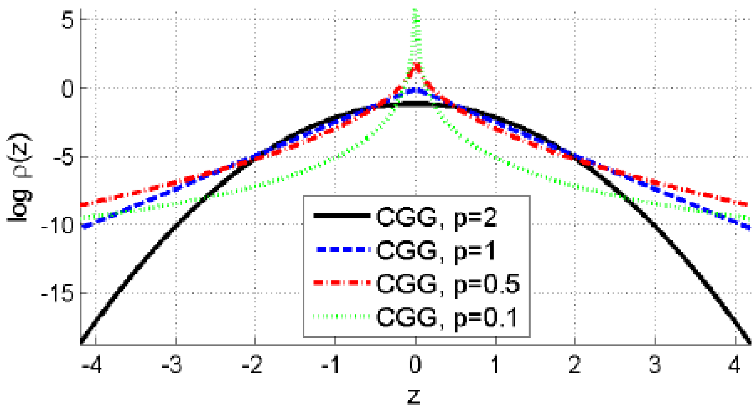
[Jukić, van Waterschoot, Gerkmann, Doclo, IEEE TASLP, 2015]

Multi-channel linear prediction (MCLP)

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

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

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



- **Remarks:**

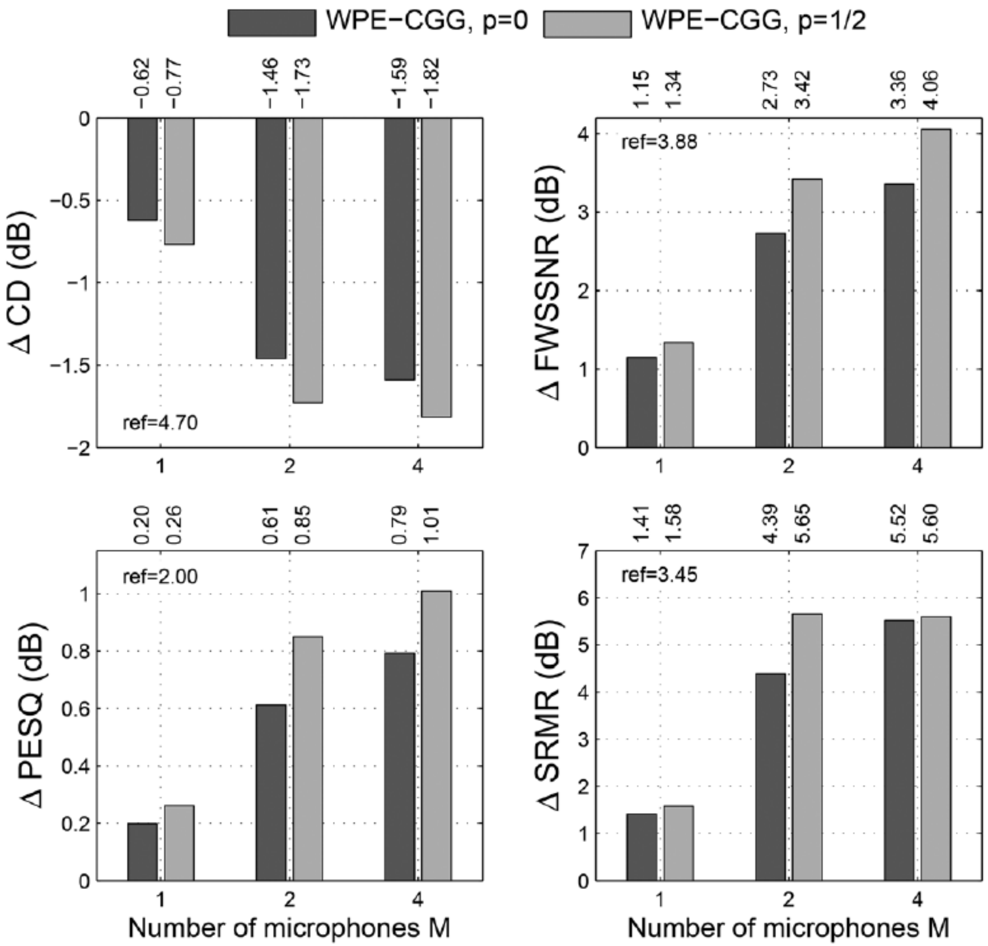
1. ML estimation using CGG prior is equivalent to ***l_p*-norm minimization**
   → **promotes sparsity of TF-coefficients across time** (for p < 2)

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

   Solved using (regularized) iteratively reweighted least-squares (IRLS) procedure
2. Conventional approach (TVG model) corresponds to **p=0:**
   - **Strong sparse prior,** strongly favoring values of desired signal close to zero
   - Hyper-prior on variance equal to constant value

Multi-channel linear prediction (MCLP)

- **Instrumental validation (noiseless, batch)**



clean **microphone**

WPE-CGG p=0 **WPE-CGG p=0.5**

Performance depends on p, with p=0.5 consistently yielding (small) improvements

T_{60} ≈ 700ms, M={1,**2**,4}, distance 2 m, fs=16 kHz; STFT: 64ms (overlap 16ms); MCLP: L_g={35,**15**,8}, τ=2

[Jukić, van Waterschoot, Gerkmann, Doclo, IEEE TASLP, 2015]

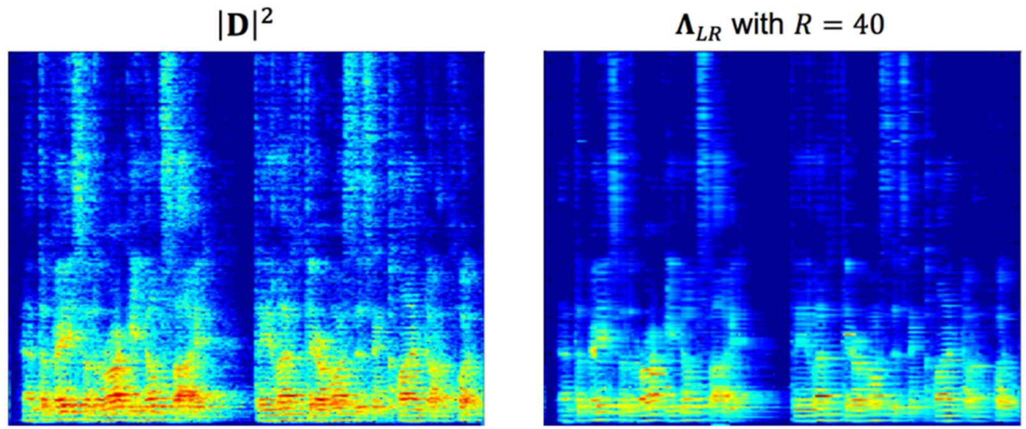
MCLP extensions (low-rank structure)

- **Incorporate additional knowledge of speech spectrogram**
  - Exploit time-frequency structure of spectrogram (no frequency-independent processing)
  - Speech spectrogram exhibits low-rank structure [Smaragdis 2006] → non-negative matrix factorization (NMF)

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

→ Improved preservation of time-frequency structure

→ Increased sparsity



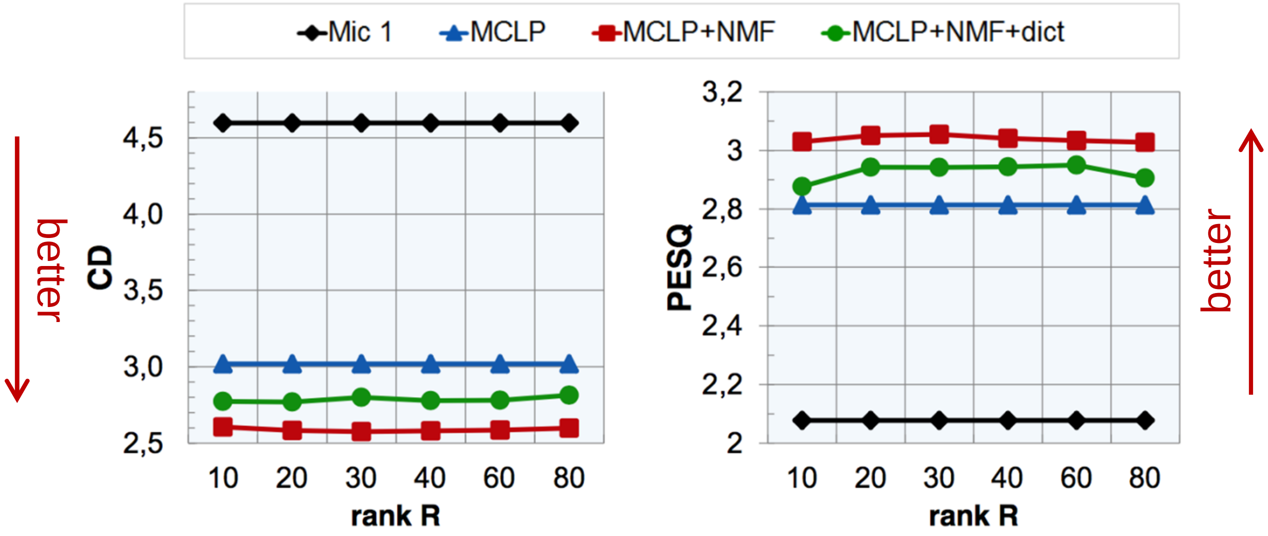
  - **Incorporate NMF in MCLP-based dereverberation**
    - Variances estimated as $\mathbf{\Lambda}_{LR} = \text{low\_rank\_approximation}(|\mathbf{D}|^2)$
    - Either unsupervised or supervised (using pre-trained dictionary)

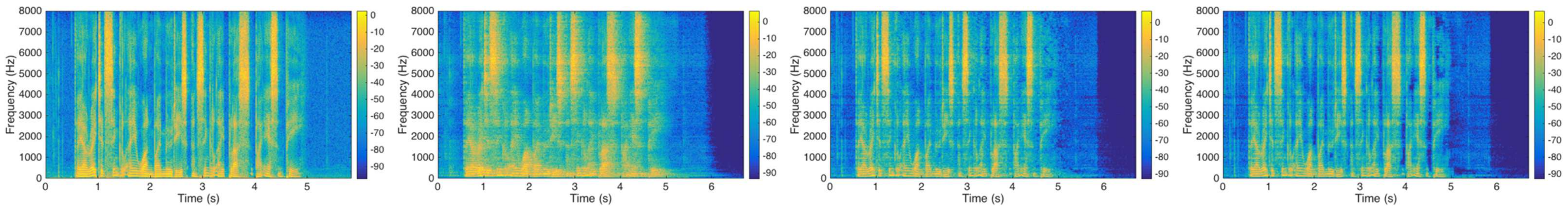
MCLP extensions (low-rank structure)

- **Instrumental validation (noiseless, batch)**

1. *unsupervised*: dictionary learned from spectrogram $|\mathbf{D}|^2$ (MCLP+NMF)

2. *supervised*: pretrained dictionary (MCLP+NMF+dict)





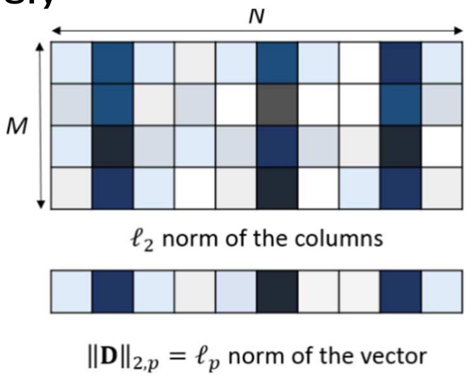
Clean | Microphone | MCLP | MCLP+NMF

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

[Jukić, Mohammadiha, van Waterschoot, Gerkmann, Doclo, ICASSP 2015]

MCLP extensions (group sparsity)

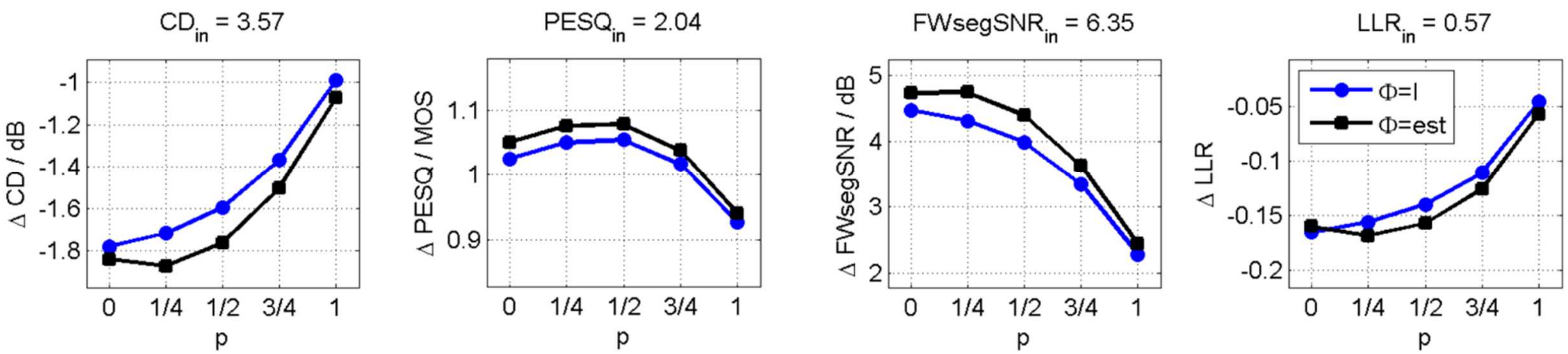
- **Group sparsity** for MIMO speech dereverberation:
  - Maximize sparsity of TF-coefficients across time + simultaneously keep/discard TF-coefficients across microphones (= groups)
  - → **Mixed $l_{2,p}$-norm**



$$\|\mathbf{D}\|_{\Phi;2,p} = \left(\sum_{n=1}^{N} \|\mathbf{d}_{n,:}\|_{\Phi;2}^{p}\right)^{1/p} \qquad \sum_{n=1}^{N} \|\mathbf{d}_{n,:}\|_{\Phi;2}^{p} \approx \sum_{n=1}^{N} w_n^{(i)} \|\mathbf{d}_{n,:}\|_{\Phi;2}^{2}$$

- **Remarks:**
  - Multiple outputs → possibility to apply spatial filtering (e.g., MVDR beamforming)

- **Instrumental validation (noiseless, batch)**



T_{60} ≈ 700ms, M=4, distance 2m, fs=16 kHz; STFT: 64ms (overlap 16ms); MCLP: L_g=10, τ=2

MCLP extensions (adaptive MCLP)

- **Batch processing → adaptive processing**
  - Incorporate exponential weighting in cost function (iteratively reweighted l_2-norm)
    → **RLS-based algorithm**

$$\hat{\mathbf{G}} = \arg\min_{\mathbf{G}} \sum_{n=1}^{N} w(n)\|\mathbf{d}(n)\|_2^2 \quad \Rightarrow \quad \hat{\mathbf{G}}(n) = \arg\min_{\mathbf{G}(n)} \sum_{t=1}^{n} \gamma^{n-t} w(t)\|\mathbf{d}(t)\|_2^2$$

  - **Problem:** overestimation of undesired component (late reverberation) for small forgetting factors γ (dynamic scenarios) → severe distortion in output signal

- **Constrained adaptive MCLP**
  - **Idea**: constrain MCLP-based estimate of undesired component using estimate of late reverberant PSD (e.g., based on statistical model [Polack, Lebart])

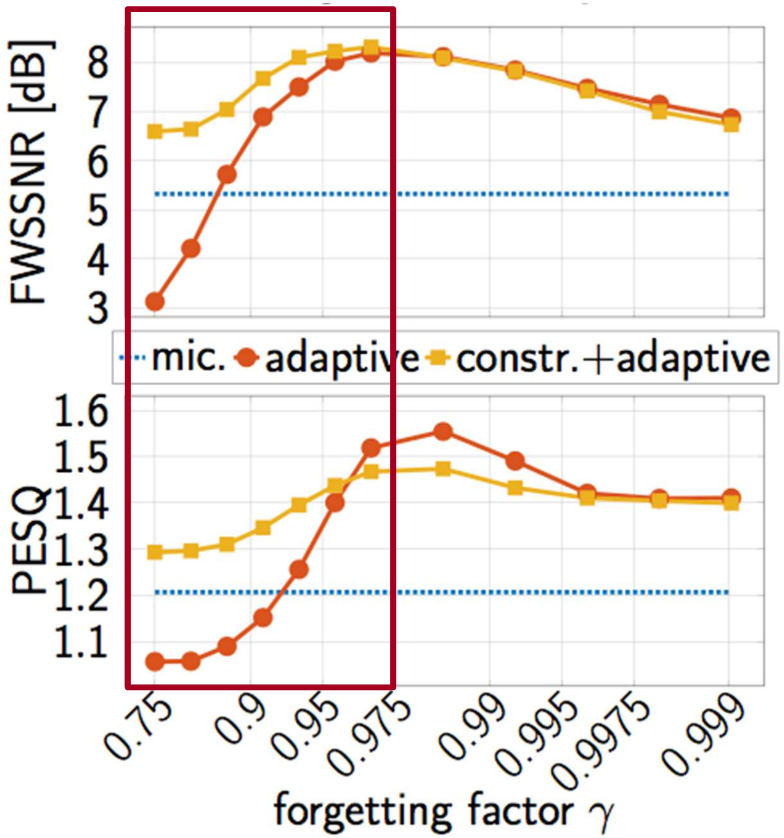
$$\check{\mathbf{G}}(n) = \arg\min_{\mathbf{G}(n)} \sum_{t=1}^{n} \gamma^{n-t} w(t)\|\mathbf{d}(t)\|_2^2 \quad \boxed{\text{subject to} \quad |\mathbf{G}^{\mathsf{H}}(n)\tilde{\mathbf{x}}_{\tau}(n)|^2 \leq \hat{\boldsymbol{\sigma}}_u^2(n)}$$

  - Constraint ensures stability and prevents overestimation
  - Optimization method: ADMM – results in RLS-like updates

[Yoshioka and Nakatani, EUSIPCO 2013] [Jukić, van Waterschoot, Doclo, IEEE SPL 2017]

MCLP extensions (adaptive MCLP)

- **Instrumental validation (noiseless, adaptive)**



clean **microphone**

	ADA	Constr. +ADA
γ=0.98		
γ=0.88		

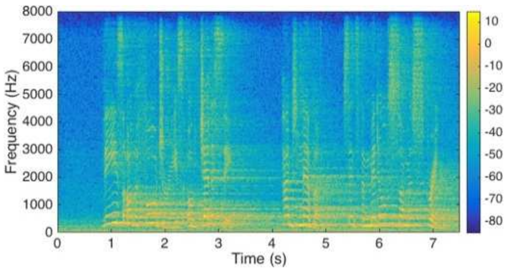
Constrained MCLP much less sensitive to forgetting factor (especially for small values)

T_{60} ≈ 700ms, M=2, distance 2m, **source switching between +45 and -45,** fs=16 kHz; STFT: 64ms (overlap 16ms); L_g=20, τ=2, p=0

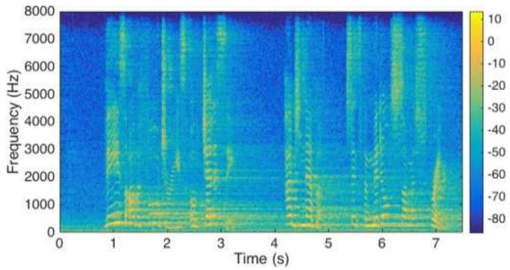
MCLP extensions (adaptive MCLP)

- **Instrumental validation (high reverberation + noisy, adaptive)**

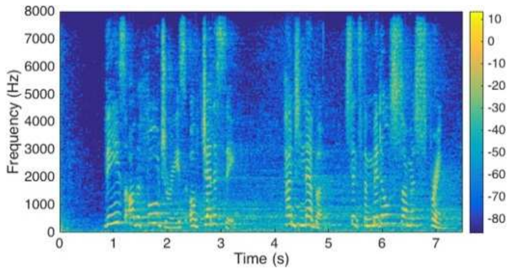
d ~ 2m



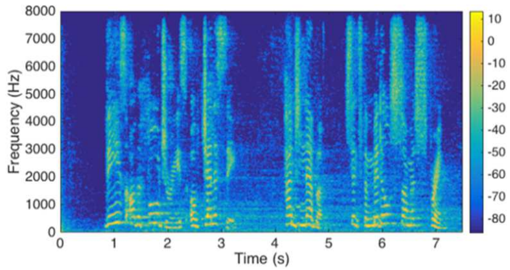
Microphone



1ch SE [REVERB]



Adaptive MCLP



Adaptive MCLP + SE

T60 ~ 6s (St Alban The Martyr Church, London), M=2 (spacing~1m), fs=16 kHz, **real recordings**
STFT: 64ms (overlap 16ms); MCLP: L_g=30, τ=2, p=0, adaptive (γ=0.96)

MCLP extensions (general framework)

- **General framework:**
  - **Wideband (WB) signal model:** $\mathbf{x}_{\text{ref}} = \mathbf{d} + \mathbf{X}\mathbf{g}$

    Narrowband (NB) signal model: $\tilde{\mathbf{x}}_{\text{ref},k} = \tilde{\mathbf{d}}_k + \tilde{\mathbf{X}}_k \tilde{\mathbf{g}}_k$
  - **Sparsity of STFT coefficients** of desired speech signal:
    - *Synthesis sparsity*: time-domain signal $\mathbf{d}$ can be represented using sparse estimated STFT coefficients $\tilde{\mathbf{d}}$
    - *Analysis sparsity*: STFT coefficients $\tilde{\mathbf{d}}$ of estimated time-domain signal $\mathbf{d}$ are sparse

Wideband-Analysis (WB-A)

$$\begin{aligned} \min_{\mathbf{d},\mathbf{g}} \quad & P\left(\mathbf{\Psi}^{\mathrm{H}}\mathbf{d}\right) \\ \text{subject to} \quad & \mathbf{d} + \mathbf{X}\mathbf{g} = \mathbf{x}_{\text{ref}} \end{aligned}$$

Wideband-Synthesis (WB-S)

$$\begin{aligned} \min_{\tilde{\mathbf{d}},\mathbf{g}} \quad & P\left(\tilde{\mathbf{d}}\right) \\ \text{subject to} \quad & \mathbf{\Psi}\tilde{\mathbf{d}} + \mathbf{X}\mathbf{g} = \mathbf{x}_{\text{ref}} \end{aligned}$$

Narrowband (NB)

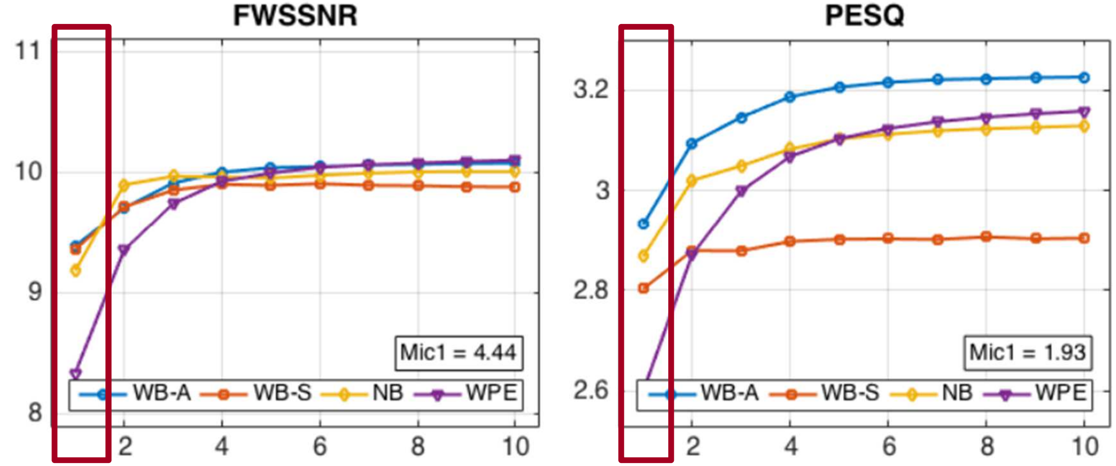
$$\begin{aligned} \min_{\tilde{\mathbf{d}}_k,\tilde{\mathbf{g}}_k} \quad & P\left(\tilde{\mathbf{d}}_k\right) \\ \text{subject to} \quad & \tilde{\mathbf{d}}_k + \tilde{\mathbf{X}}_k \tilde{\mathbf{g}}_k = \tilde{\mathbf{x}}_{\text{ref},k} \end{aligned}$$

- $\mathbf{\Psi}$ denotes TF transform (e.g. STFT), P denotes sparsity-promoting function (e.g. weighted l_1-norm), possibly including structured sparsity (e.g. NMF weights)
- Optimization method: ADMM
- Wideband model: more flexibility (selection of TF transform), but much larger complexity

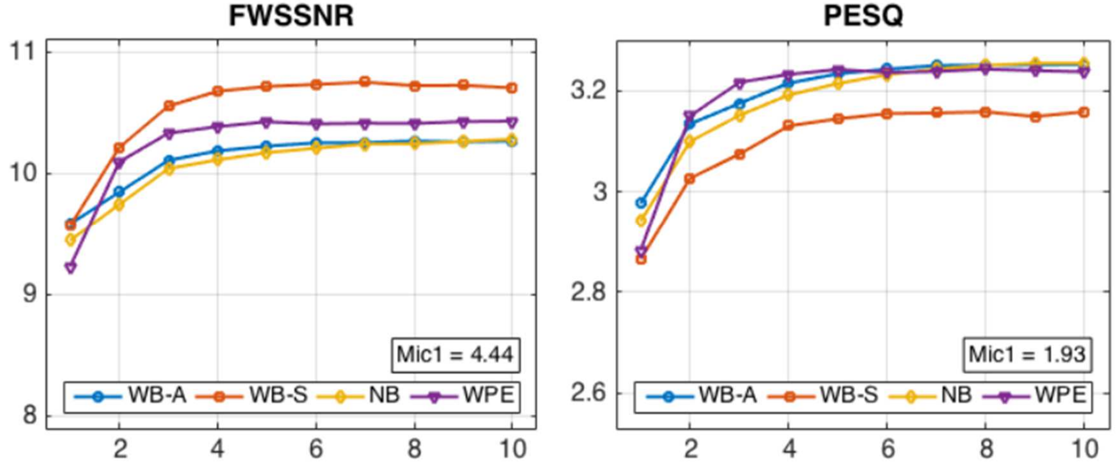
MCLP extensions (general framework)

- **Instrumental validation**
  - **ADMM-based methods (l_1-norm) perform better than WPE (l_2-norm)** for single reweighting iteration
  - Similar performance for multiple iterations
  - **Structured weights result in improved performance** (especially for WPE)

Local weights



NMF weights



T_{60} ≈ 700ms, M=2, distance 2m, fs=16 kHz; STFT: 64ms (overlap 16ms); MCLP: L_g=5120, τ=20 (WB), L_g=20, τ=2 (NB)

[Jukić, van Waterschoot, Gerkmann, Doclo, JAES, in press]

Conclusions

- **Incorporate sparsity of clean speech TF coefficients into multi-microphone speech dereverberation**

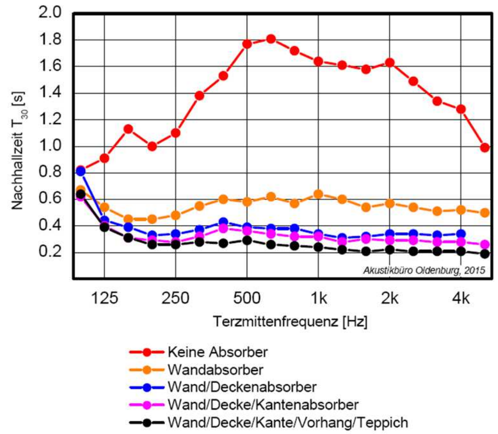
- **Acoustic multi-channel equalization:**
  - Signal-dependent regularization with sparsity-promoting penalty function (weighted l_1-norm) **increases robustness against RIR perturbations**

- **Multi-channel linear prediction:**
  - Role of sparsity: ML estimation using CGG prior is equivalent to **l_p-norm minimization → promotes sparsity of TF-coefficients across time**
  - Extensions by using time-frequency structure (NMF) and group sparsity
  - **General framework** (wideband + narrowband)

Current / future work

- **Blind probabilistic model-based approach**
  - Joint dereverberation and noise reduction based on sparsity-promoting cost functions
  - Comparison of CTF model vs. AR model
- **Distributed MCLP** for acoustic sensor networks
- **Instrumental measures**: prediction of perceived level of reverberation and speech quality for speech dereverberation algorithms
- Inaugurate new **varechoic lab**

Abbildung 1: In Raum E10 in den in Tabelle 1 angegebenen Raumzuständen gemessenen Nachhallzeiten in Terzbändern im Vergleich



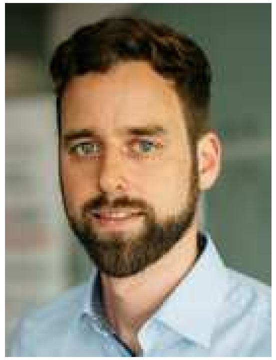
Acknowledgments

Dr. Ina Kodrasi

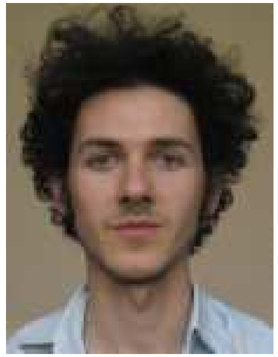
Ante Jukić

Benjamin Cauchi

Dr. Nasser Mohammadiha



Prof. Timo Gerkmann



Prof. Toon van Waterschoot

- **Funding:**
  - Cluster of Excellence Hearing4All (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)

DFG Deutsche Forschungsgemeinschaft

DREAMS Dereverberation and Reverberation of Audio, Music, and Speech

GIF German-Israeli Foundation for Scientific Research and Development

Recent publications

- A. Jukić, T. van Waterschoot, T. Gerkmann, S. Doclo, "A general framework for incorporating time-frequency domain sparsity in multi-channel speech dereverberation," *Journal of the Audio Engineering Society*, in press.
- I. Kodrasi, S. Doclo, *Joint Dereverberation and Noise Reduction Based on Acoustic Multichannel Equalization*, *IEEE/ACM Trans. Audio, Speech and Language Processing*, vol. 24, no. 4, pp. 680-693, Apr. 2016.
- I. Kodrasi, A. Jukic, S. Doclo, *Robust sparsity-promoting acoustic multi-channel equalization for speech dereverberation*, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016.
- I. Kodrasi, S. Goetze, S. Doclo, *Regularization for Partial Multichannel Equalization for Speech Dereverberation*, *IEEE Trans. Audio, Speech and Language Processing*, vol. 21, no. 9, pp. 1879-1890, Sep. 2013.
- A. Jukić, T. van Waterschoot, T. Gerkmann, S. Doclo, *Multi-channel linear prediction-based speech dereverberation with sparse priors*, *IEEE/ACM Trans. Audio, Speech and Language Processing*, vol. 23, no. 9, pp. 1509-1520, Sep. 2015.
- A. Jukić, T. van Waterschoot, T. Gerkmann, S. Doclo, *Group sparsity for MIMO speech dereverberation*, in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, USA, Oct. 2015, pp. 1-5.
- A. Jukić, N. Mohammadiha, T. van Waterschoot, T. Gerkmann, S. Doclo, *Multi-channel linear prediction-based speech dereverberation with low-rank power spectrogram approximation*, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 96-100.
- B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, S. Goetze, *Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech*, *EURASIP Journal on Advances in Signal Processing*, 2015:61, pp. 1-12.
- N. Mohammadiha, S. Doclo, *Speech Dereverberation Using Non-negative Convolutive Transfer Function and Spectro-temporal Modeling*, *IEEE/ACM Trans. Audio, Speech and Language Processing*, vol. 24, no. 2, pp. 276-289, Feb. 2016.

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

Questions ?

House of Hearing, Oldenburg