Speaker Adapted Beamforming for Multi-Channel Automatic Speech Recognition

Tobias Menne, **Ralf Schlüter**, **Hermann Ney**

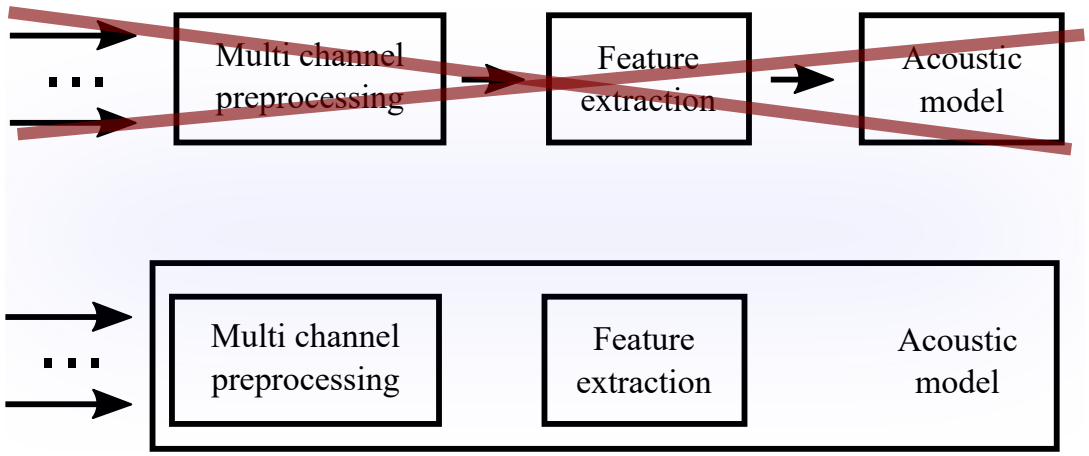
menne@cs.rwth-aachen.de

Content

- Introduction
- Mask based beamforming
- Integration into the acoustic model
- Speaker adaptation
- Results (on CHiME-4)

Motivation

- Integration of signal preprocessing into the acoustic model
- Joint optimization towards ASR objective
- Here:
  adjusting preprocessing to specific acoustic scenario (e.g. speaker of interest)



Previous Work

- Learning beamforming from the (raw) input signal
  [Sainath & Weiss$^+$ 15, Sainath & Weiss$^+$ 16, Li & Sainath$^+$ 16]

- Estimating the filter matrix of filter and sum beamforming
  [Xiao & Watanabe$^+$ 16, Meng & Watanabe$^+$ 17]

- Incorporation of statistically optimal beamforming into the acoustic model
  [Heymann & Durde$^+$ 17]

Mask based beamforming

Recorded noisy multi-channel signal:

$$\mathbf{Y}_{t,f} = \mathbf{X}_{t,f} + \mathbf{N}_{t,f} \tag{0.1}$$

$\mathbf{Y}_{t,f}, \mathbf{X}_{t,f}, \mathbf{N}_{t,f} \in \mathbb{C}^M$

t - time frame index

f - frequency bin index

M - nr of microphones

Mask based beamforming

Recorded noisy multi-channel signal:

$$\mathbf{Y}_{t,f} = \mathbf{X}_{t,f} + \mathbf{N}_{t,f} \tag{0.2}$$

$\mathbf{Y}_{t,f}, \mathbf{X}_{t,f}, \mathbf{N}_{t,f} \in \mathbb{C}^M$

t - time frame index

f - frequency bin index

M - nr of microphones

Filter and sum beamforming:

$$\hat{S}_{t,f} = \mathbf{w}_f^H \cdot \mathbf{Y}_{t,f} \tag{0.3}$$

$\hat{S}_{t,f} \in \mathbb{C}$ - estimate of the speech component

$\mathbf{w}_f \in \mathbb{C}^M$ - beamforming vector

Mask based beamforming

Beamforming vector by speech and noise mask:

$$\mathbf{w}_f^H = g(\lambda_{t,f}^{(X)}, \lambda_{t,f}^{(N)}, \mathbf{Y}_{\cdot,f}) \tag{0.4}$$

$\lambda^{(\nu)} = f_\Theta^{(\nu)}(\mathbf{Y})$ - Mask estimated by neural network

Mask based beamforming

Beamforming vector by speech and noise mask:

$$\mathbf{w}_f^H = g(\lambda_{t,f}^{(X)}, \lambda_{t,f}^{(N)}, \mathbf{Y}_{\cdot,f}) \tag{0.5}$$

$\lambda^{(\nu)} = f_\Theta^{(\nu)}(\mathbf{Y})$ - Mask estimated by neural network

GEV Beamformer

$\mathbf{w}_f^{(\mathrm{GEV})} = \mathcal{P}(\mathbf{\Phi}_{NN,f}^{-1}\mathbf{\Phi}_{XX,f})$

[Warsitz & Haeb-Umach 07]

MVDR Beamformer

$\mathbf{w}_f^{(\mathrm{MVDR})} = \mathcal{P}(\mathbf{\Phi}_{XX,f})$

[Higuchi & Ito$^+$ 16]

Mask based beamforming

Beamforming vector by speech and noise mask:

$$\mathbf{w}_f^H = g(\lambda_{t,f}^{(X)}, \lambda_{t,f}^{(N)}, \mathbf{Y}_{\cdot,f}) \tag{0.6}$$

$\lambda^{(\nu)} = f_\Theta^{(\nu)}(\mathbf{Y})$ - Mask estimated by neural network

GEV Beamformer

$\mathbf{w}_f^{(\mathrm{GEV})} = \mathcal{P}(\mathbf{\Phi}_{NN,f}^{-1}\mathbf{\Phi}_{XX,f})$

[Warsitz & Haeb-Umach 07]

MVDR Beamformer

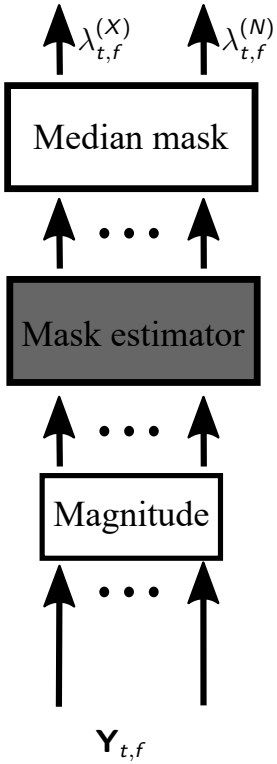
$\mathbf{w}_f^{(\mathrm{MVDR})} = \mathcal{P}(\mathbf{\Phi}_{XX,f})$

[Higuchi & Ito$^+$ 16]

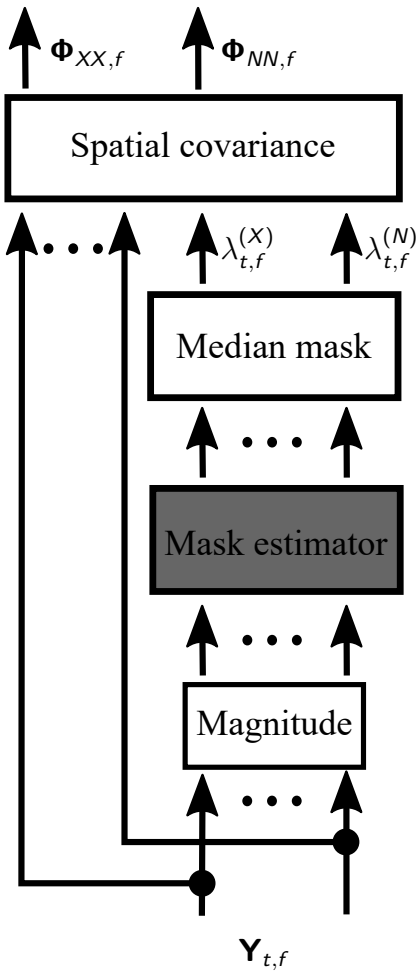
$\mathcal{P}(\cdot)$ - Principal eigenvector

$$\mathbf{\Phi}_{\nu\nu,f} = \frac{1}{\sum_{t=1}^{T} \lambda_{t,f}^{(\nu)}} \sum_{t=1}^{T} \lambda_{t,f}^{(\nu)} \mathbf{Y}_{t,f}\mathbf{Y}_{t,f}^H \tag{0.7}$$

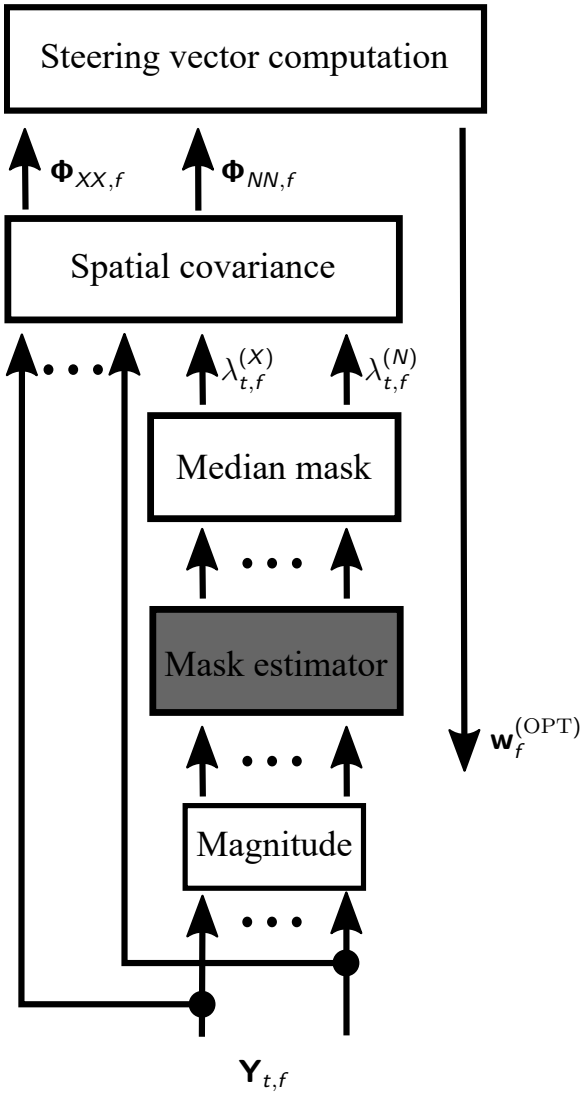
Integration



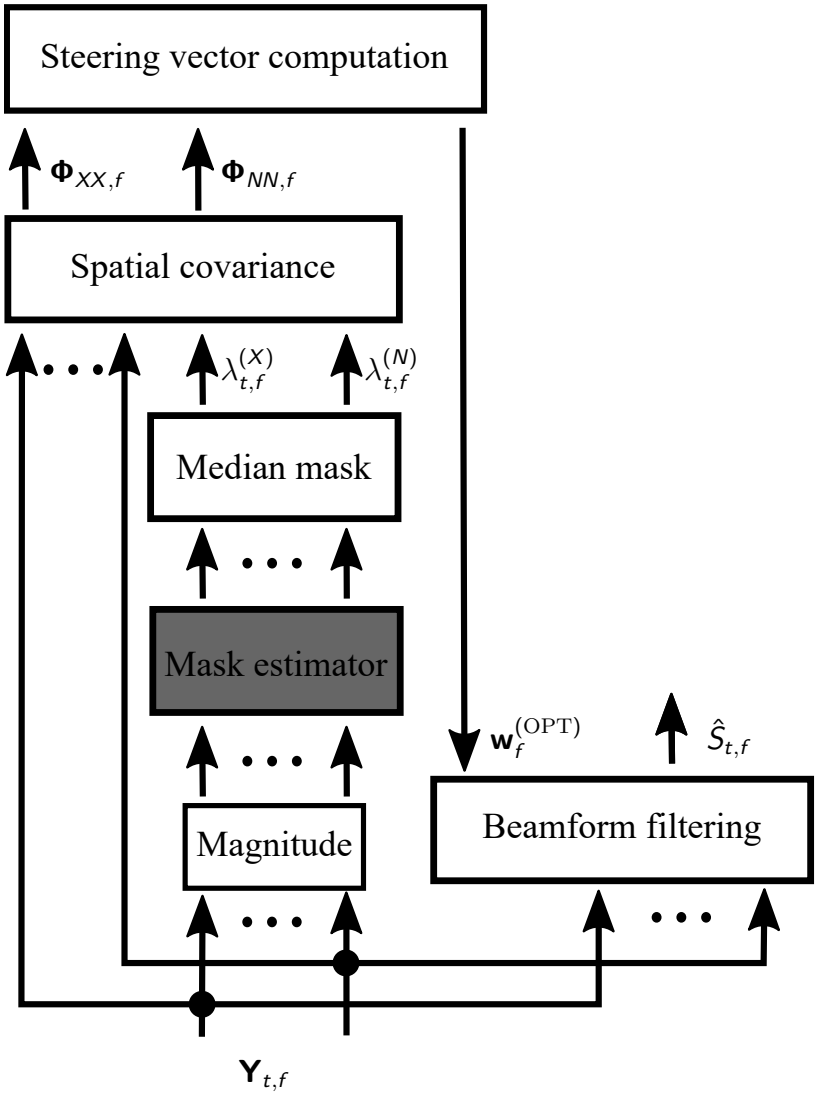
Integration



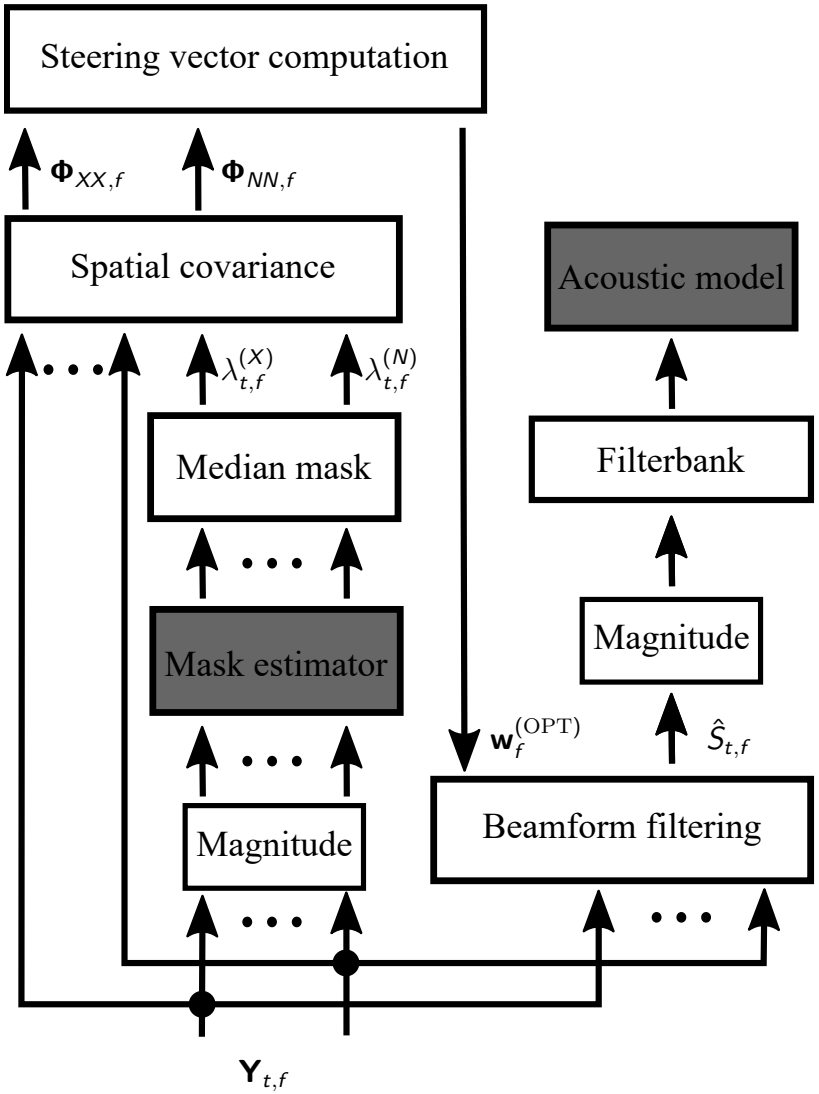
Integration



Integration



Integration



QR algorithm

[Francis 61]

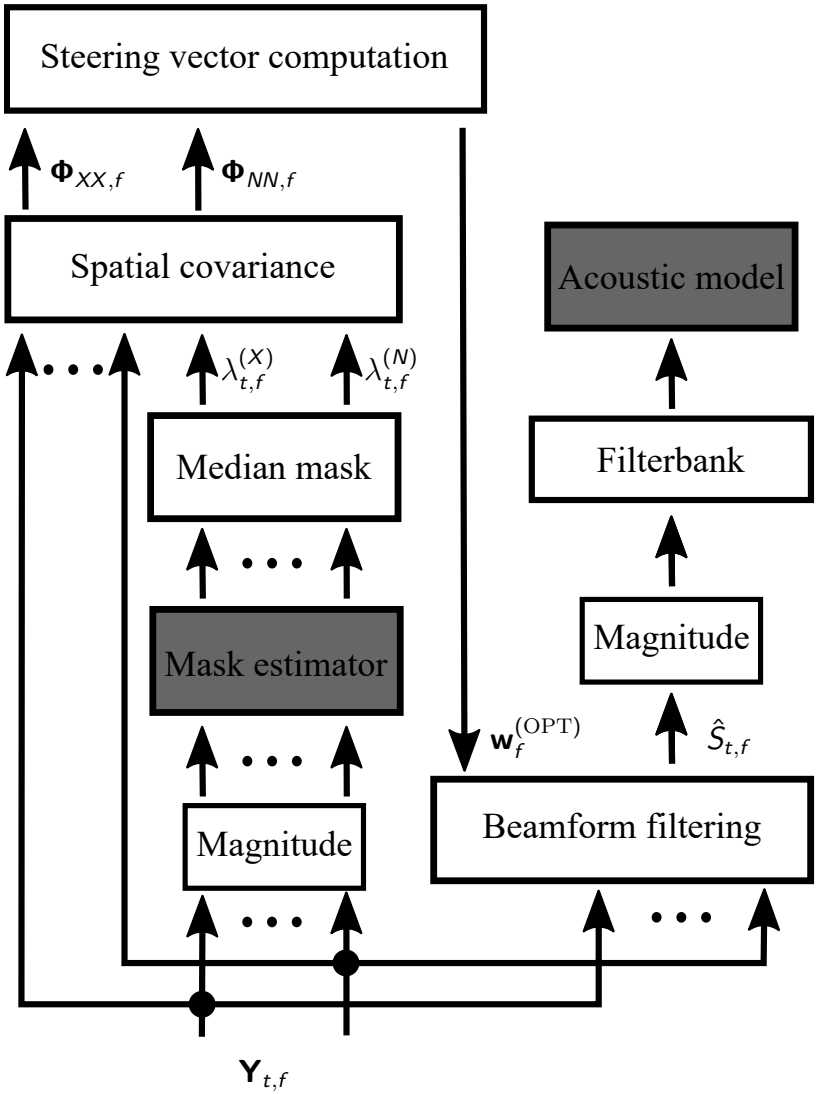
Objective: compute $\mathcal{P}(A_0)$

1. QR-decomposition of $A_k = Q_k R_k$
2. set $A_{k+1} = R_k Q_k$

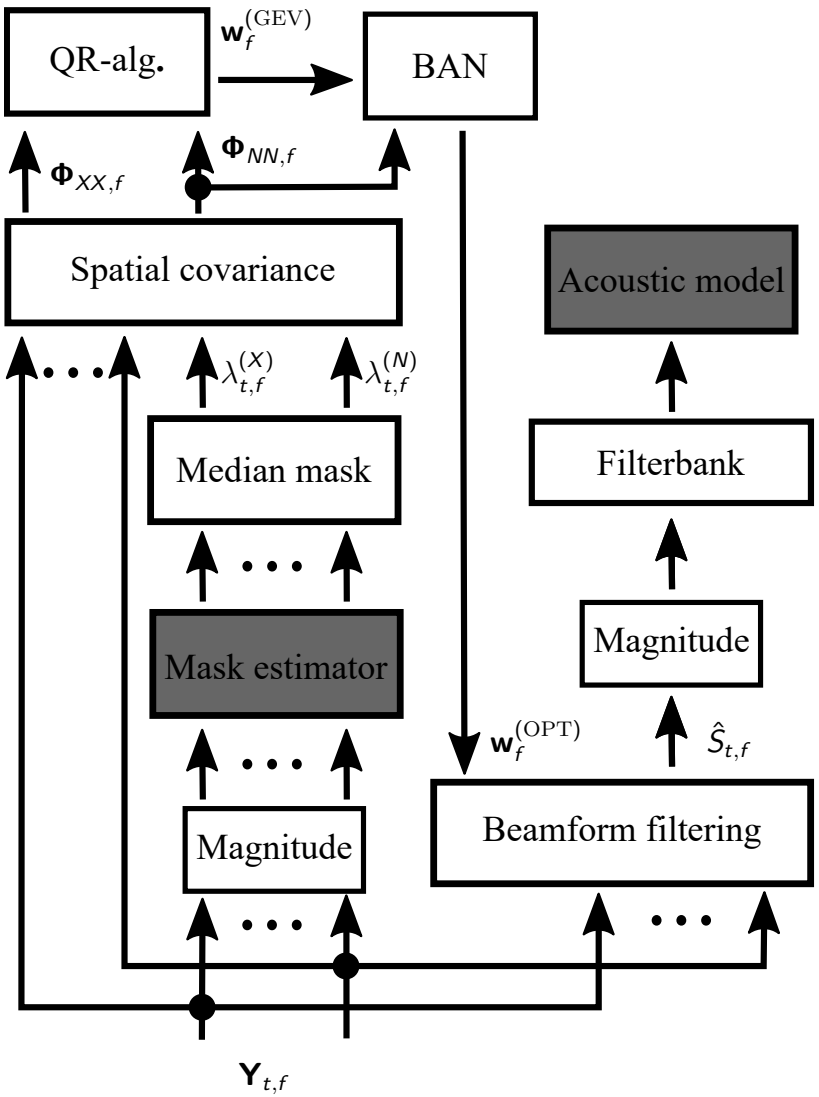
For $K \to \infty$ the diagonal of A_K contains eigenvalues of A_0
With $\prod_{k=0}^{K} Q_k$ containing the respective eigenvectors

QR-decomposition is differentiable as shown in [Walter & Lehmann$^+$ 12].

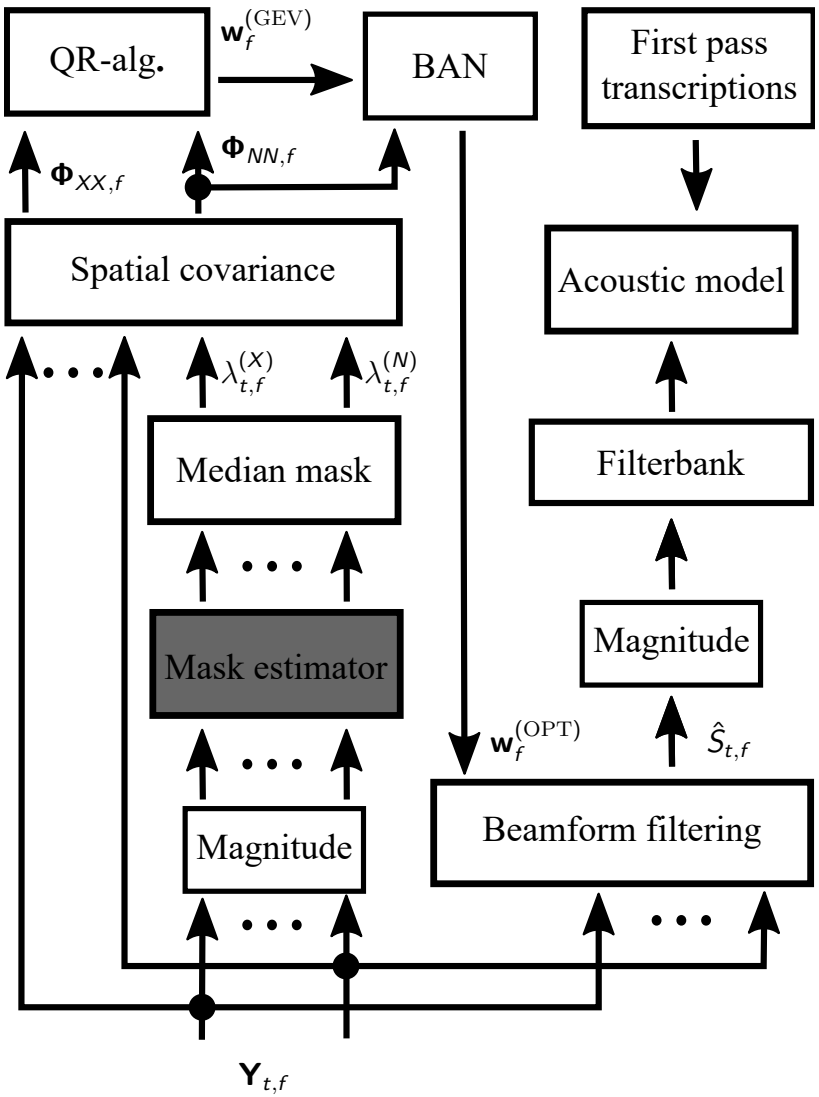
Integration



Integration



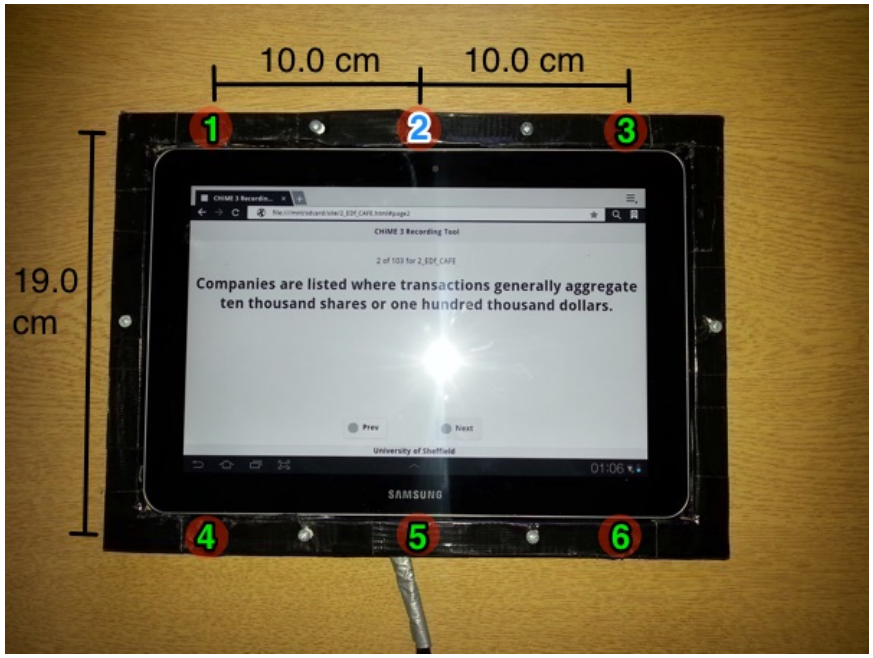
Speaker adaptation



CHiME-4 dataset

- 18 h of simulated and real 6-channel data (87 speakers)
- recorded with handheld device
- noise scenarios: caffee, bus, pedestrian area, street junction

Set	female	male	time (h) per speaker
Dev	2	2	0.7
Eval	2	2	0.5



System

Mask estimator

- 1 BLSTM (256 units)A
- 2 fully connected ReLU (512 units)
- 5 iterations of QR-algorithm
- hyper parameters of speaker adaptation tuned on dev set

Acoustic model

- unnormalized 80 dimensional log mel filterbank features
- linear layer (80 units) with batch normalization
- 5 BLSTM layers (600 units each)

System performance at different training stages

Average WER (%) for the described systems for different stages of the integrated training.

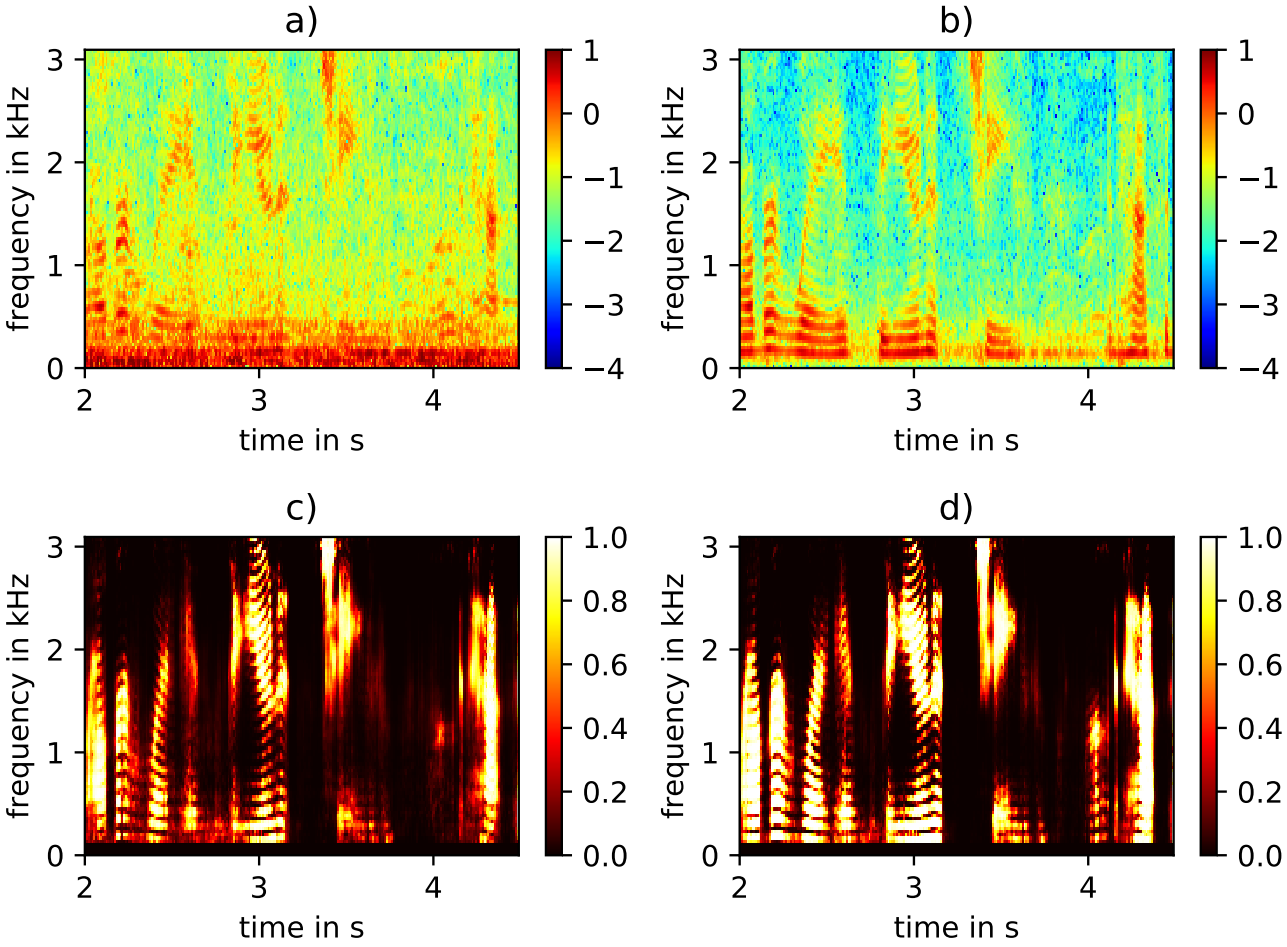
System				Dev	Eval
System id	Front-end	Joint training	Speaker adapted		
0	BFIT	-	-	4.36	7.17
1	GEV			3.46	5.18
2		+		3.32	4.84
3			+	**3.09**	**4.58**

Speaker specific performance

WER (%) of separate speakers for the jointly trained system and the speaker adapted system

System id	Dev				Eval			
	F01	F04	M03	M04	F05	F06	M05	M06
2	4.19	3.23	2.77	**3.07**	6.88	**4.09**	3.83	4.58
3	**3.55**	**3.20**	**2.48**	3.14	**6.35**	**4.09**	**3.38**	**4.48**

Mask estimations



Conclusion

- Joint optimization of preprocessing and acoustic modeling is beneficial
- Speaker adaptation of mask estimator for beamforming can be beneficial depending on target speaker
- Adaptation does not decrease performance otherwise
- Masks after adaptation show stronger emphasis of speakers fundamental frequency and harmonics

Thank you for your attention

Any questions?

- J. G. Francis.
  The QR transformation a unitary analogue to the LR transformation - part 1.
  The Computer Journal, Vol. 4, No. 3, pp. 265–271, Jan 1961.
- J. Heymann, L. Durde, C. Boeddeker, P. Hanebrink, R. Haeb-Umbach.
  Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system.
  In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5325–5329, New Orleans, LA, March 2017.
- T. Higuchi, N. Ito, T. Yoshioka, T. Nakatani.
  Robust MVDR beamforming using time frequency masks for online offline ASR in noise.
  In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5210–5214, Shanghai, China, March 2016.
- B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, M. Bacchiani.

Neural network adaptive beamforming for robust multichannel speech recognition.
In *Proc. Interspeech*, pp. 1976–1980, San Francisco, CA, Sep 2016.

- Z. Meng, S. Watanabe, J. R. Hershey, H. Erdogan.
  Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition.
  In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 271–275, New Orleans, LA, Mar 2017.

- T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani, A. Senior.
  Speaker location and microphone spacing invariant acoustic modeling from raw multichannel waveforms.
  In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 30–36, Scottsdale, AZ, Dec 2015.

- T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani.
  Factored spatial and spectral multichannel raw waveform cldnns.

In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5075–5079, Shanghai, China, Mar 2016.

- S. F. Walter, L. Lehmann, R. Lamour.
  On evaluating higher-order derivatives of the QR decomposition of tall matrices with full column rank in forward and reverse mode algorithmic differentiation.
  Optimization Methods and Software, Vol. 27, No. 2, pp. 391–403, 2012.

- E. Warsitz, R. Haeb-Umach.
  Blind acoustic beamformting based on generalized eigenvalue decomposition.
  IEEE Transactions on audio, speech, and language processing, Vol. 15, pp. 1529–1539, June 2007.

- X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, D. Yu.
  Deep beamforming networks for multi-channel speech recognition.
  In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5745–5749, Shanghai, China, Mar 2016.