UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Nachrichtentechnik

Neural Network Supported Acoustic Beamforming and Source Separation for ASR

Reinhold Häb-Umbach

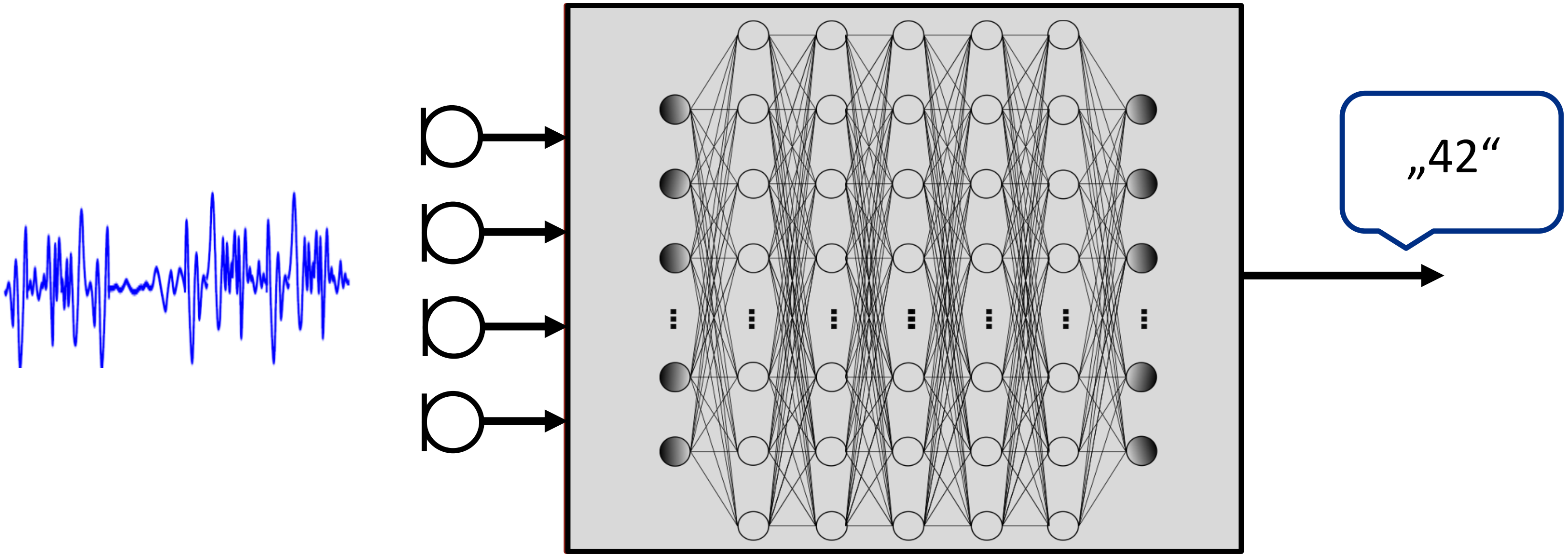
Paderborn University

Contributors

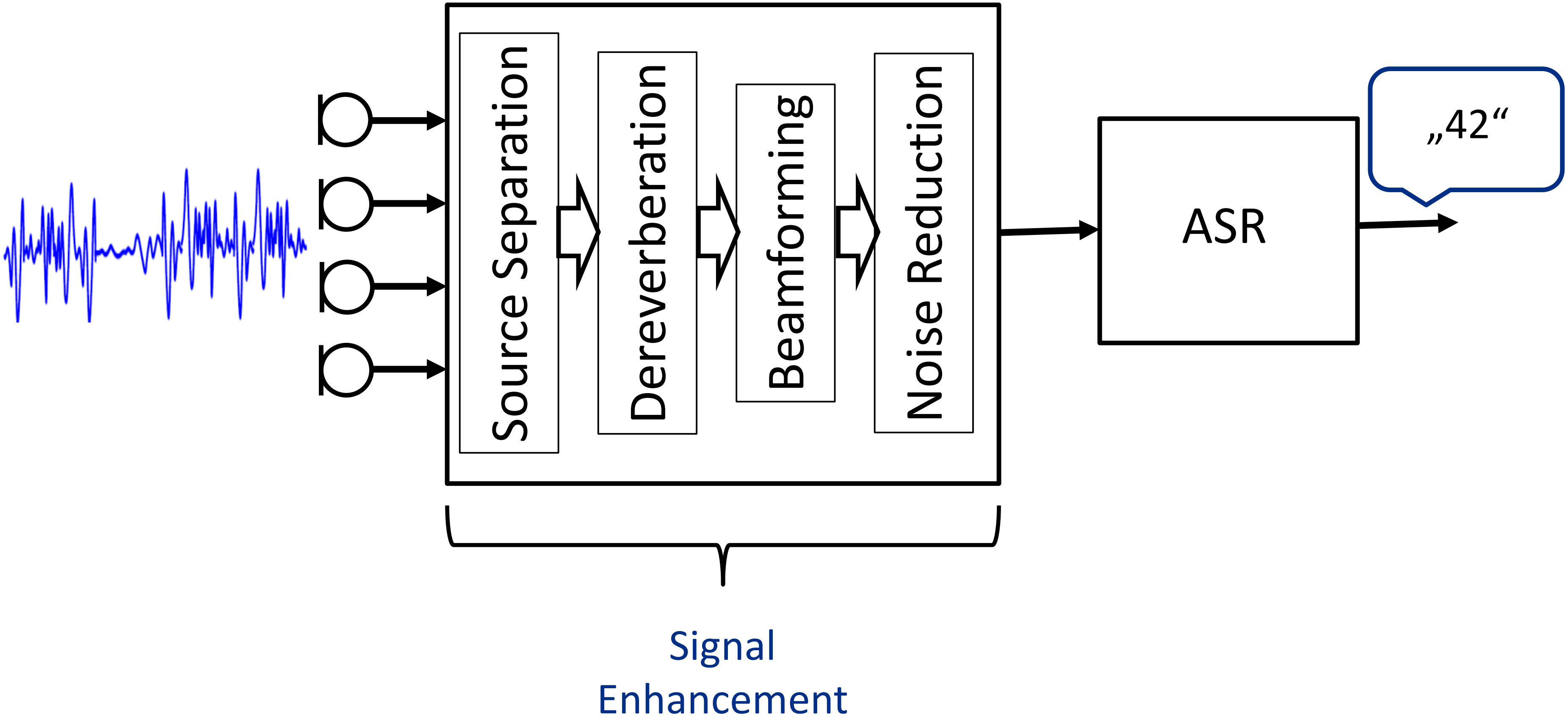
- Joint work with

Jens Heitkämper	Lukas Drude	Jahn Heymann	Christoph Boeddeker

ASR – the Modern Approach



The Old-Fashioned Approach



Integrated vs Modular (1/2)

Integrated (no explicit enhancement stage):

- \+ Common objective function (discrim. training)
- \+ Avoids premature decisions
- ± Robust?
  - – Irrelevant variations left in signal
  - \+ Acoustic model is exposed to large variability in training
- – Large network, requires lots of training data, large computational and memory demands
- – Cannot easily exploit phase (spatial) information

Integrated vs Modular (2/2)

Modular (explicit enhancement stage):

+ Statistically optimum solutions known (for some tasks)
+ Can efficiently treat phase (spatial) information
+ Parsimonious w.r.t. parameters, computing power
- Separately optimized, and hence suboptimal

Our Conclusion

For some tasks
(beamforming, dereverberation, source separation)
an explicit enhancement stage is (still?) advantageous

How to Do Signal Enhancement? (1/2)

Model-based:

+ Can incorporate prior knowledge
  (physical constraints, findings from psychoacoustics, …)
+ Easier to adapt in dynamic acoustic scenarios

± Unsupvervised learning, if any

– But the model is only as good as the model is – and its parameters

How to Do Signal Enhancement? (2/2)

Neural Networks:

- \+ Can model arbitrary mapping
- \+ Not limited by (often simplifying) model assumptions
- \+ Have shown to be superior on several tasks
- ± Supervised learning
- – Difficult to incorporate prior knowledge, constraints

Our Conclusion

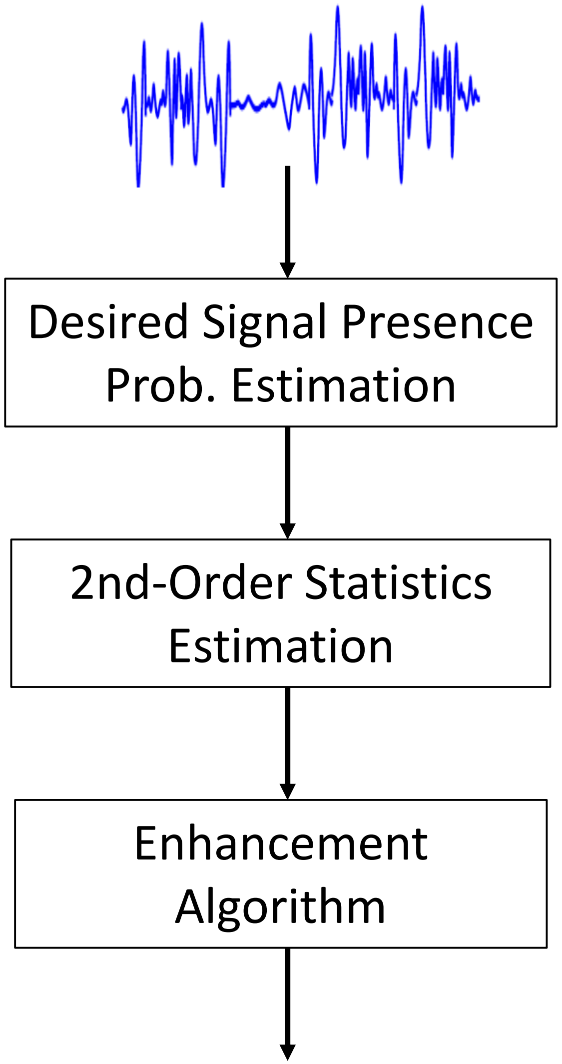
A clever combination of neural networks with model-based approaches can combine the advantages of both worlds:

Neural network Supported Signal Enhancement

Table of Contents

- Neural network for „desired signal presence“ probability estimation,
  to support
  - Acoustic beamforming
  - Dereverberating beamforming
  - (Noise tracking)
  - Blind source separation
- Integration with backend ASR

Structure



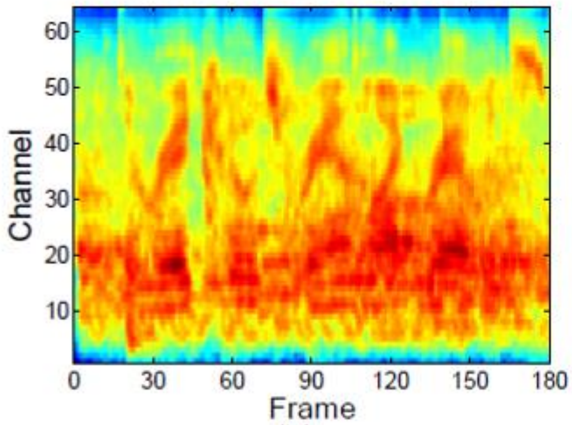
Desired signal presence probability (DSPP) estimation

(= mask estimation)

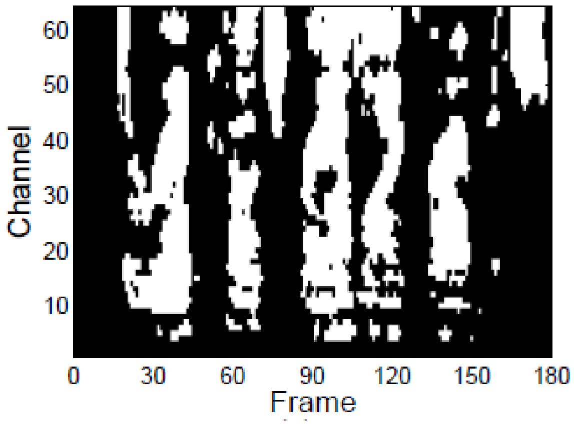
- Power spectral density
- Spatial covariance matrix
- Speech presence prob.
- Noise presence prob.
- Dominant speaker index
- Beamforming
- Dereverberation
- Noise reduction
- Source extraction

Speech Presence Probability (SPP) Estimation

Given:



Wanted:



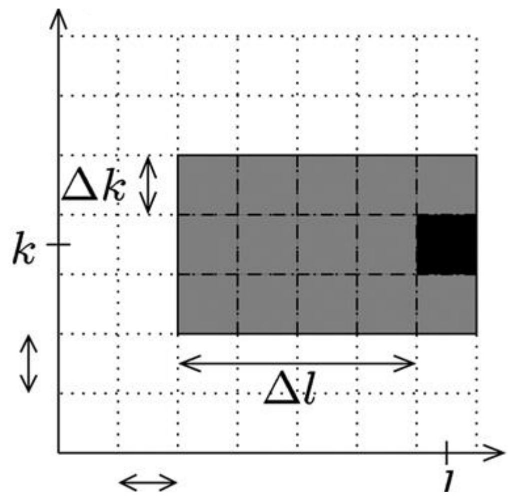
- Decide for each tf-bin if it contains speech or noise only, using
  - ➢ spectral information
  - ➢ or spatial information
  - ➢ or both

Options for SPP Estimation

- Spectro-temporal smoothing
- Formulated as unsupervised problem
- Formulated as supervised learning problem

SPP via Spectro-Temporal Smoothing

[Raj 2002, Gerkmann&Martin 2008, Momeni&Habets 2014, …]



from
[Gerkmann & Martin, 2008]

- Discussion
  - Mostly single-channel
  - Suitable for online processing

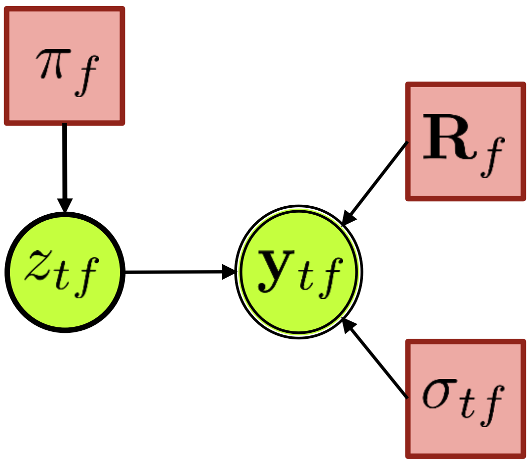
SPP as Unsupervised Learning Problem: EM

[Souden 2010, Tran & Haeb-Umbach 2010/2012, Ito 2014, …]

- Generative model:

$$\mathbf{y}_{tf} = \begin{cases} \mathbf{n}_{tf} & z_{tf} = 0 \\ \mathbf{a}_{tf} s_{tf} + \mathbf{n}_{tf} & z_{tf} = 1 \end{cases}$$

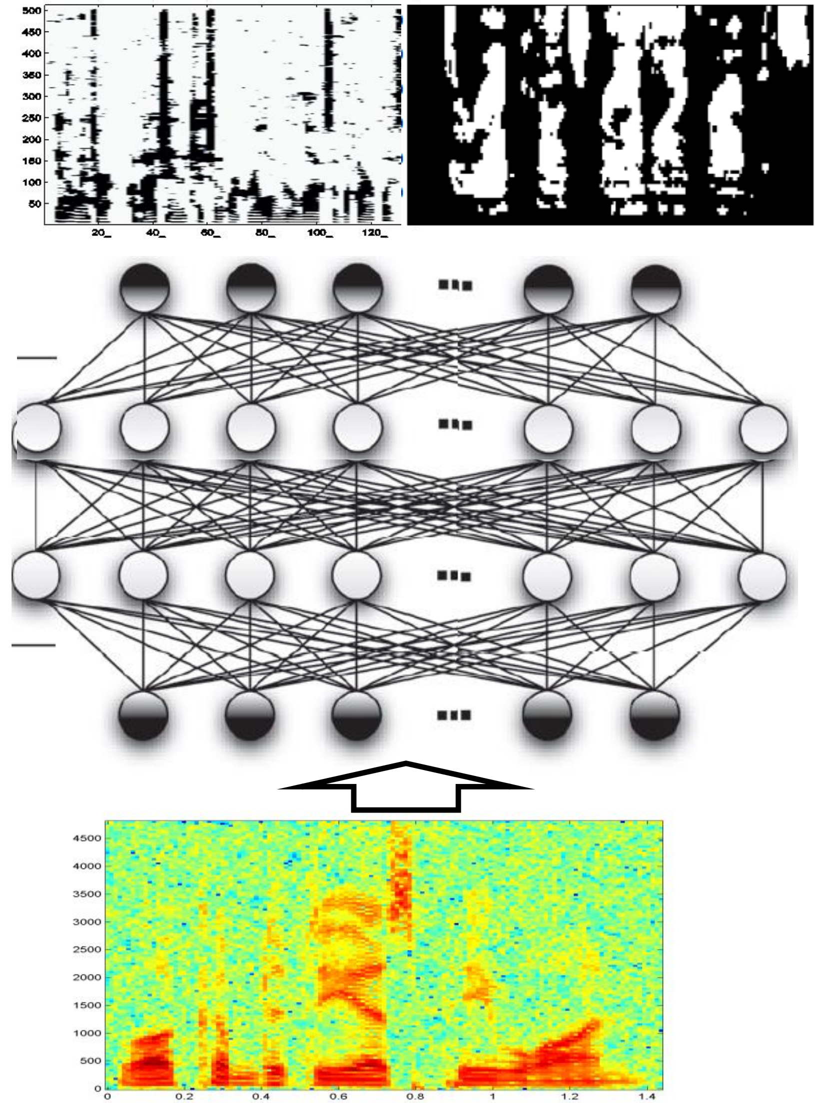
- EM Algorithm:
  - E-Step
    - Estimate SPP: $\gamma_{tf} = \Pr(z_{tf} = 1 | \mathbf{y}_{tf})$
  - M-Step
    - Estimate source/signal parameters



- Discussion
  - Mostly multi-channel, exploiting spatial information
  - i.i.d.
  - Frequencies treated independently
  - Offline block processing

SPP as Supervised Learning Problem: NN

[Wang 2013, Heymann 2015 & 2016, …]



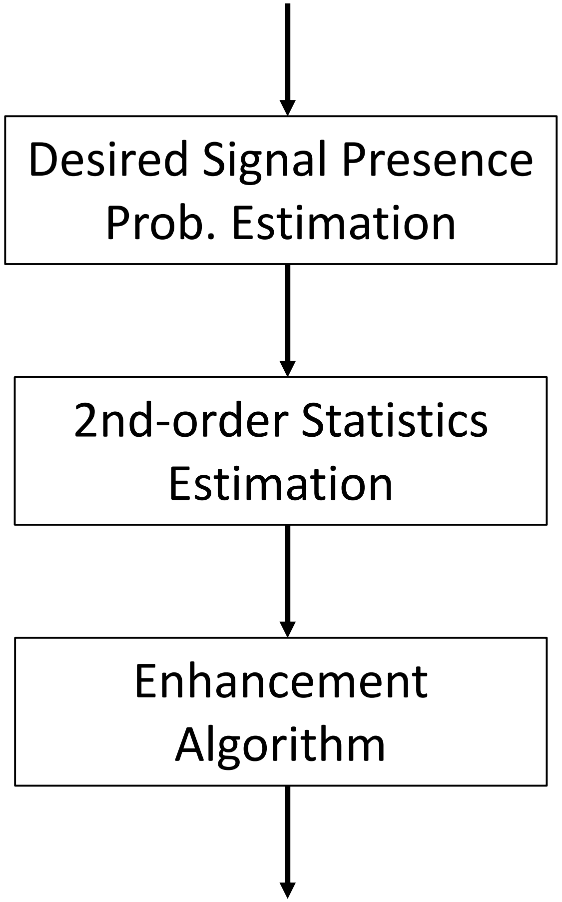
- NN as classifier

- Discussion:
  - Single channel or cross-channel features
  - Can capture temporal **and** spectral correlations
  - Offline / block online / online

Table of Contents

- Neural network for „desired signal presence“ probability estimation,
  to support
  - Acoustic beamforming
  - Dereverberating beamforming
  - (Noise tracking)
  - Blind source separation
- Integration with backend ASR

Acoustic Beamforming

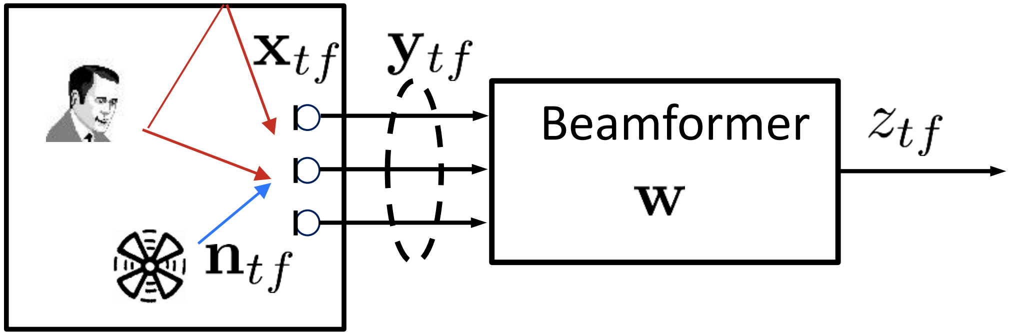


Speech & Noise presence prob. estimation

Spatial covariance matrices of speech, noise

Beamforming

Statistically Optimum Beamforming

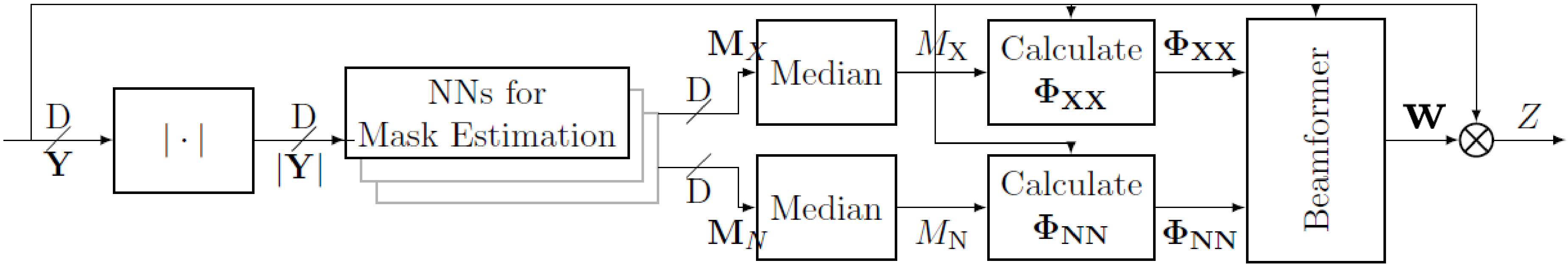


- 2nd-order statistics: $\mathbf{\Phi}_{\mathbf{xx},f}$, $\mathbf{\Phi}_{\mathbf{nn},f}$

- Beamforming: $z_{tf} = \mathbf{w}_f^H \mathbf{y}_{tf}$

e.g. MVDR: $\mathbf{w}_f^{\mathrm{MVDR}} = \frac{\mathbf{\Phi}_{\mathbf{nn},f}^{-1}\mathbf{a}_f}{\mathbf{a}_f^H \mathbf{\Phi}_{\mathbf{nn},f}^{-1}\mathbf{a}_f}$ where $\mathbf{a}_f \propto \mathcal{P}(\mathbf{\Phi}_{\mathbf{xx},f})$

NN Supported Beamformer



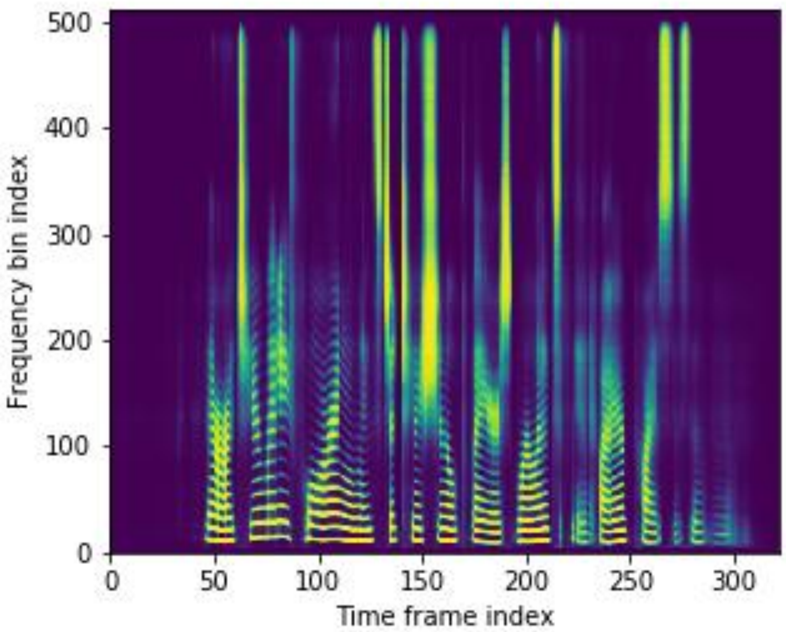
$$\boldsymbol{\Phi}_{\mathbf{xx},f} = \sum_{t=1}^{T} M_{\mathbf{x},tf} \mathbf{y}_{tf} \mathbf{y}_{tf}^{H} / \sum_{t=1}^{T} M_{\mathbf{x},tf}$$

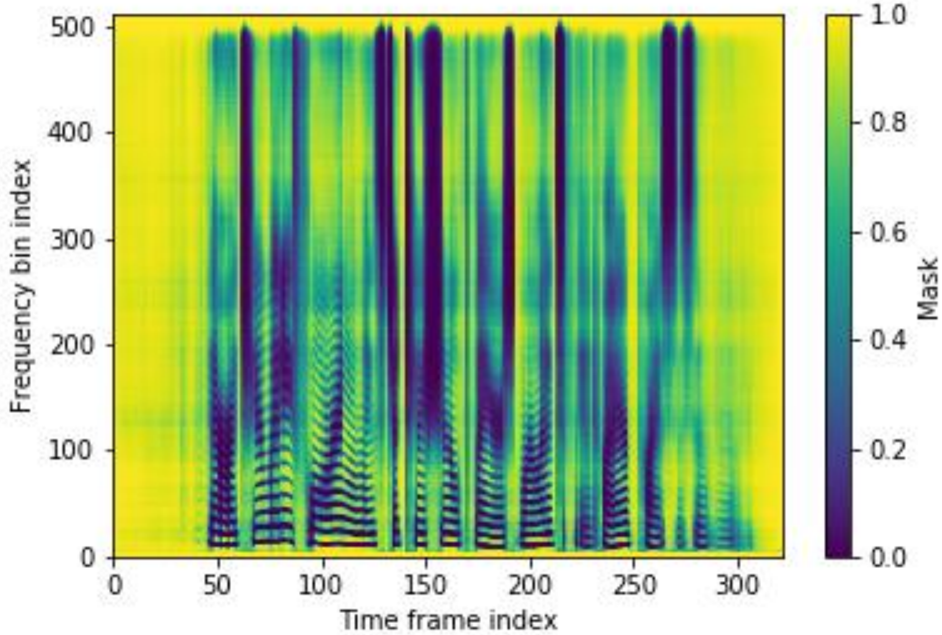
$$\boldsymbol{\Phi}_{\mathbf{nn},f} = \sum_{t=1}^{T} M_{\mathbf{n},tf} \mathbf{y}_{tf} \mathbf{y}_{tf}^{H} / \sum_{t=1}^{T} M_{\mathbf{n},tf}$$

Example Masks (CHiME-3)

Utterance ID: f04_051c0112_str

Speech mask | Noise mask





WER Results (1/2)

[Heymann et al. 2015]

CHiME-3:

- WSJ utterances
- „Fixed“ speaker positions
- Low reverberation
- Noisy environment: bus, café, street, pedestrian
- Trng set size: 18 hrs
- Offline processing

WER [%]	Eval Simu	Eval Real
Baseline	12.7	40.2
BeamformIt	23.5	22.6
Spatial mixture model [Tran & Haeb-Umbach, 2010]	20.6	22.1
NN supported Beamformer	9.7	15.4

WER Results (2/2)

[Heymann et al. 2018]

Google Voice Search data:

- Short utterances
- No prior on speaker position
- Reverberation: T_{60} = 400 … 900 ms (600 ms avg)
- Cross Talk (CT): SNR = 0 … 20 dB (12 dB avg)
- Trng set size: 150 hrs
- Online processing

	# channels			
	1	**2**	**4**	**8**
Baseline	30.6			
NN Beamformer		28.4	27.3	27.4
Baseline CT	34.8			
NN Beamformer CT		29.6	29.1	28.6

Discussion

- Neural Network supported beamforming is powerful:
  - On CHiME-4 challenge all leading groups used NN-supported beamforming

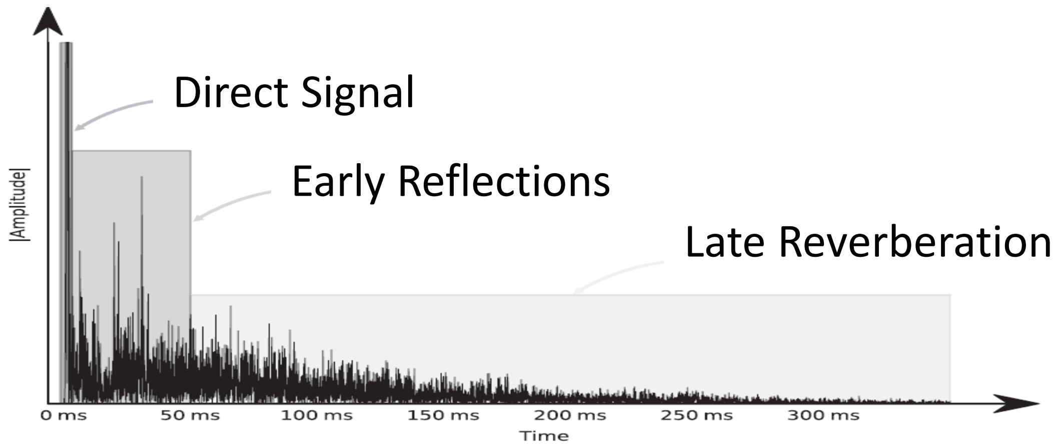
- NN independent of array configuration
- Some performance loss from offline to online (ca. 10%)
- Requires parallel (stereo) data

Table of Contents

- Neural network for „desired signal presence“ probability estimation,
  to support
  - Acoustic beamforming
  - Dereverberating beamforming
  - (Noise tracking)
  - Blind source separation
- Integration with backend ASR

Room Impulse Response

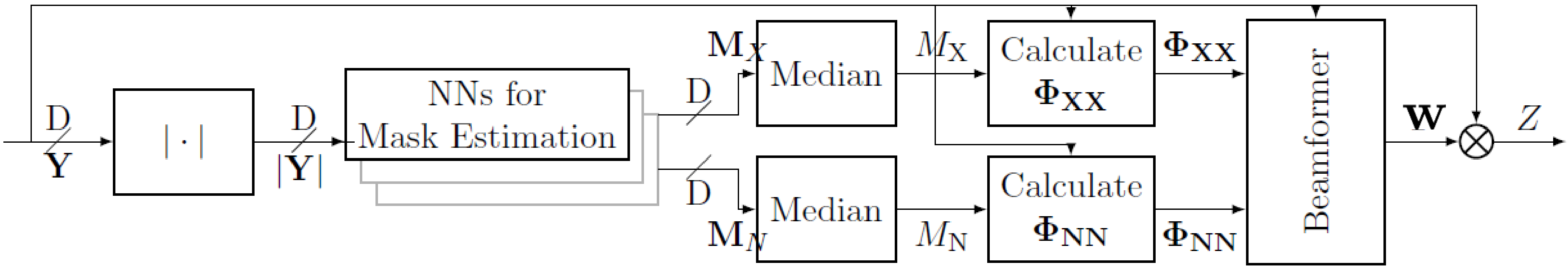
- Room impulse response
  - Desired signal: Direct signal + early reflections (50ms)
  - Distortion: late reverberation (> 50ms)



$$\mathbf{y}_{tf} = \mathbf{x}_{tf}^{\text{early}} + \mathbf{x}_{tf}^{\text{late}} + \mathbf{n}_{tf}$$

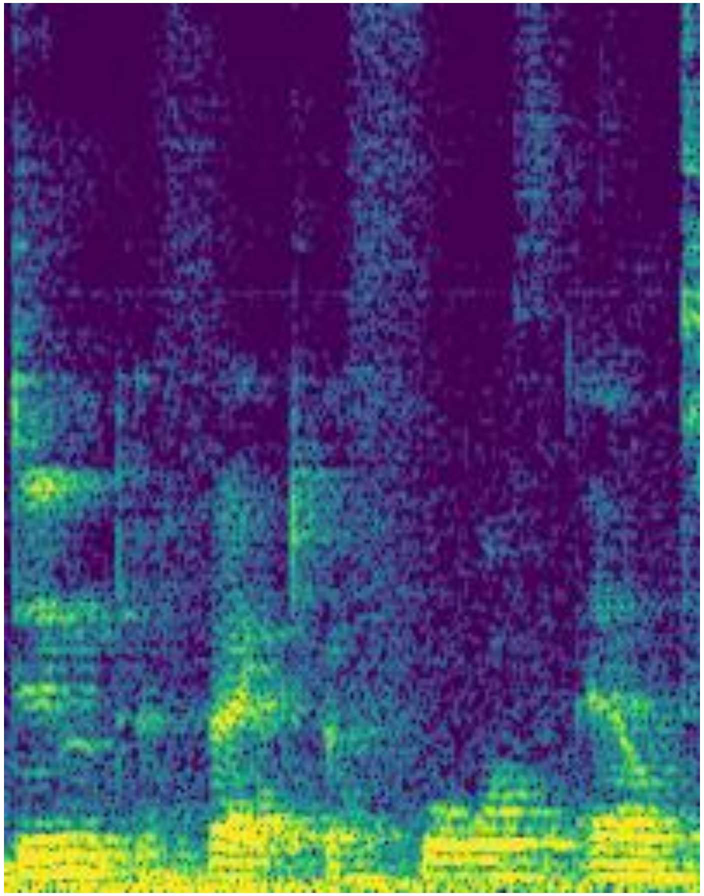
Change Training Targets

- Change NN training target:
  - ➢ $M_{\mathbf{x},tf}$: Mask to predict which tf-bin is dominated by $\mathbf{x}_{tf}^{\text{early}}$
  - ➢ $M_{\mathbf{n},tf}$: Mask to predict which tf-bin is dominated by $\mathbf{x}_{tf}^{\text{late}} + \mathbf{n}_{tf}$
- Everything else remains unchanged

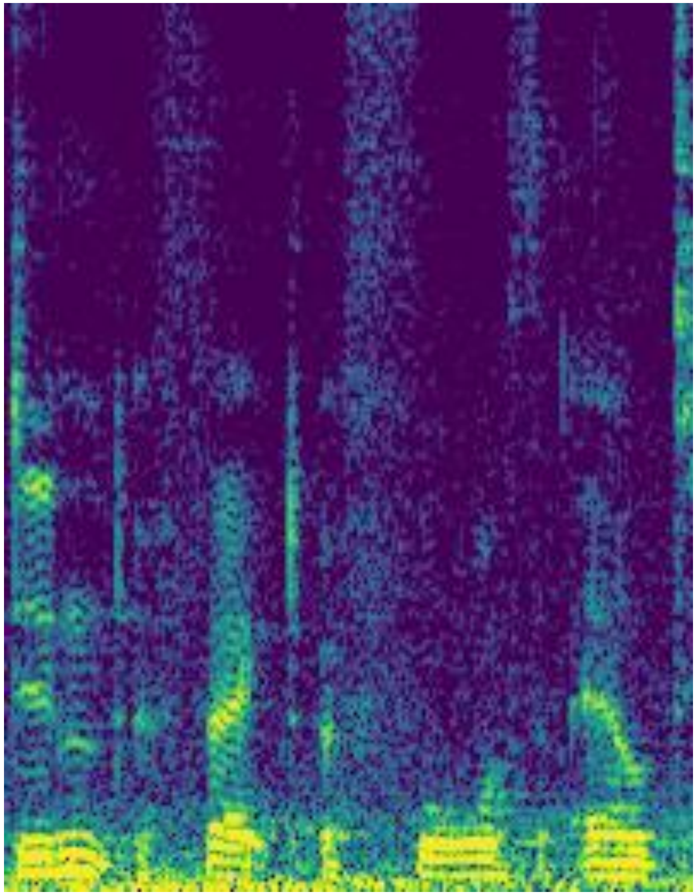


Example Spectrogram

Observed:



Enhanced:



WER Results

[Drude et al. 2018]

- Comparison with Weighted Prediction Error (WPE) dereverberation [Nakatani 2008]
- On REVERB:
  - ➢ Reverberant WSJ, T_{60} = 300 – 700 ms, SNR = 20 dB, real
  - ➢ On par with WPE
- On (WSJ + VoiceHome RIRs + VoiceHome noises)
  - ➢ Noisy reverberant WSJ, T_{60} = 400 – 600 ms, SNR = [illegible] dB, simu

WSJ + VoiceHome	**# c[illegible]**			
	1	[illegible]	4	8
Unprocessed	[illegible]	[illegible]		
WPE	[illegible]7.0	37.1	35.6	34.6
[illegible]	40.0	30.2	19.9	15.3

See Presentation on Thursday

Table of Contents

- Neural network for „desired signal presence“ probability estimation,
  to support
  - ➢ Acoustic beamforming
  - ➢ Dereverberating beamforming
  - ➢ (Noise tracking)
  - ➢ Blind source separation
- Integration with backend ASR

From SPP to DSPP

- Desired Signal Presence Probability (DSPP)

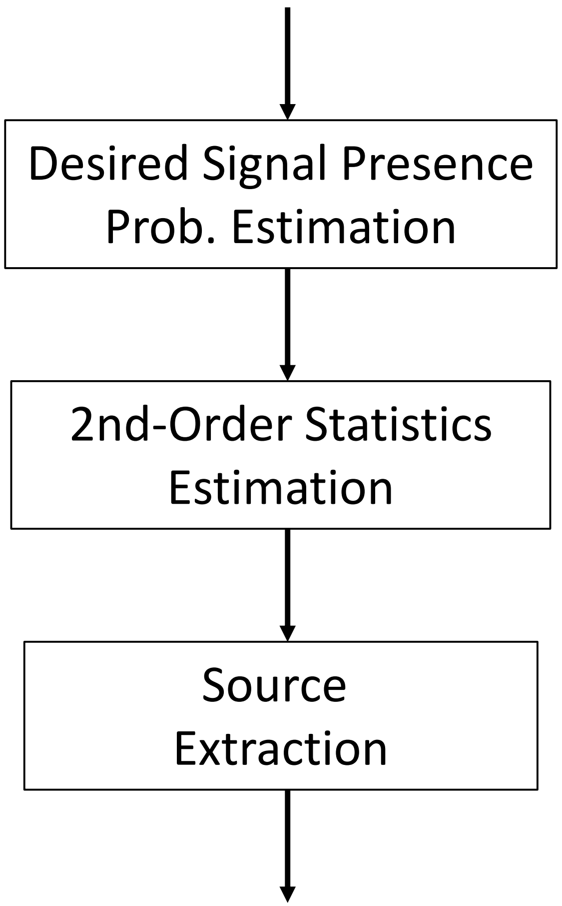
- Sparsity and W-disjoint orthogonality
  - Speech occupies only few tf-bins
  - Those are quite different from speaker to speaker

- Generative model:

$$\mathbf{y}_{tf} = \begin{cases} \mathbf{n}_{tf} & z_{tf} = 0 \\ \mathbf{a}_{tfk} s_{tfk} + \mathbf{n}_{tf} & z_{tf} = k;\ 1 \leq k \leq K \end{cases}$$

- Hidden variable $\mathbf{z}_{tf}$ (source activity indicator) indicates dominant source

Blind Source Separation



Source activity indicator estimation

Spatial mixture model

Beamforming (multiple beams),
Masking

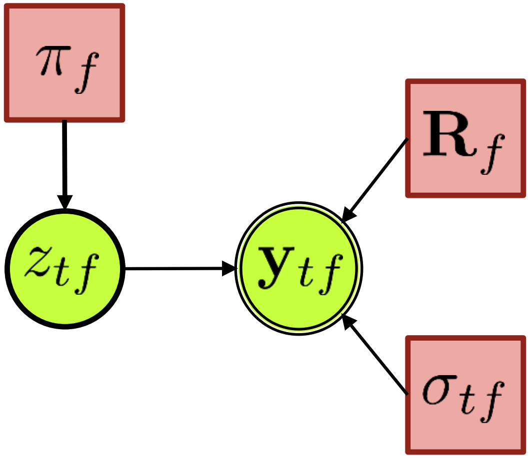
DSPP Estimation

- Formulated as unsupervised learning approach
- Formulated as supervised learning problem

Unsupervised Learning Approach: EM

- EM Algorithm
  - ➢ E-Step
    - Estimate source activity indicator: $\gamma_{tfk} = \Pr(z_{tf} = k | \mathbf{y}_{tf})$
  - ➢ M-Step
    - Estimate params of spatial mixture model

- Example spatial mixture model
  - ➢ Time-variant complex Gaussian mixture model [Ito, 2014]

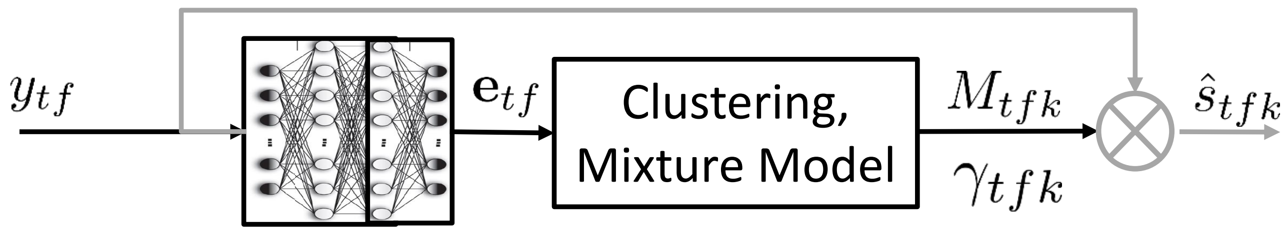


$$p(\mathbf{y}_{tf}) = \sum_k \Pr(z_{tf} = k) p(\mathbf{y}_{tf} | z_{tf} = 1)$$
$$= \sum_k \pi_{fk} \mathcal{N}_{\mathcal{C}}(\mathbf{y}_{tf}; \mathbf{0}, \sigma_{tfk} \cdot \mathbf{R}_{fk})$$

Supervised Learning Approach: NN

- Source activity indicator estimation
  - Deep clustering [Hershey, 2016]
  - Deep attractor networks [Zhou, 2017]
- Estimate embedding space where speakers form clusters
- Cluster using k-means or learn spectral mixture model on $\mathbf{e}_{tf}$

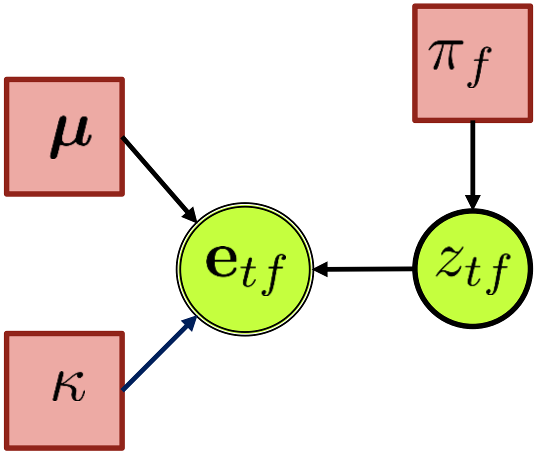
DC:



Mixture Model for Embeddings

- Mixture of von Mises-Fisher Distributions:

$$p(\mathbf{y}_{tf}) = \sum_k \Pr(z_{tfk} = 1) p(\mathbf{e}_{tf} | z_{tfk})$$
$$= \sum_k \pi_{fk} \cdot \mathrm{vMF}(\mathbf{e}_{tfk}; \boldsymbol{\mu}_k, \kappa_k)$$



Integrated Model

[Drude & Haeb-Umbach, 2017]

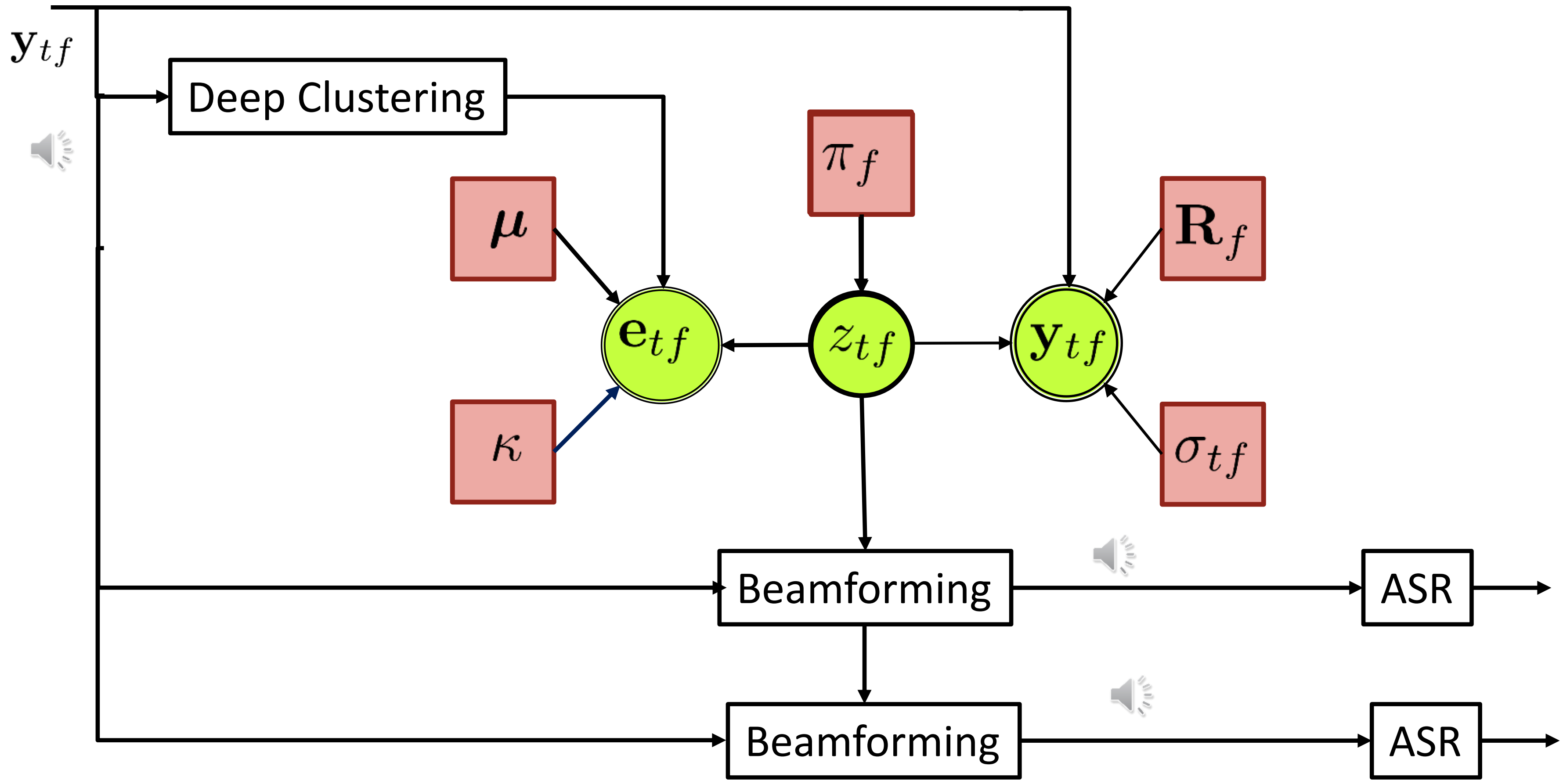
Spectral mixture model

Spatial mixture model

Integrated mixture model

- Coupling via latent class affiliation variable
- Better parameter estimation, when estimated jointly

Overall BSS Model



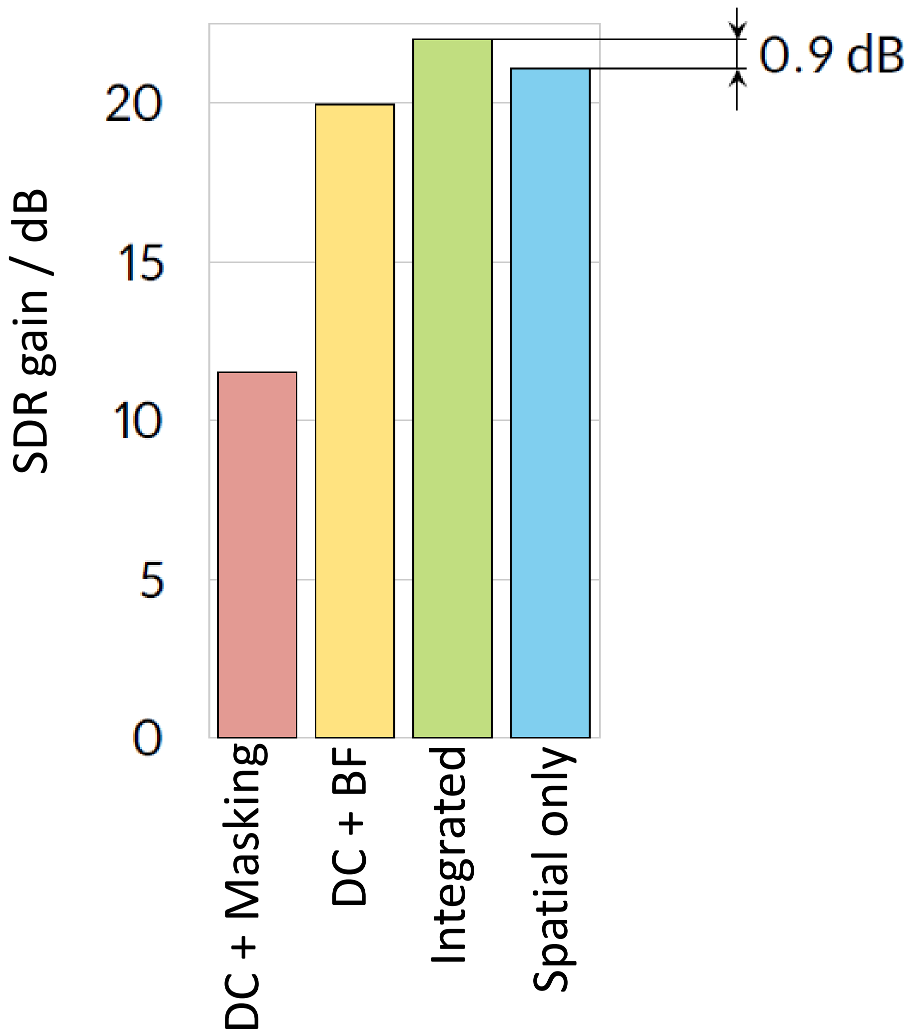
- Source extraction via beamforming or by masking

Evaluation

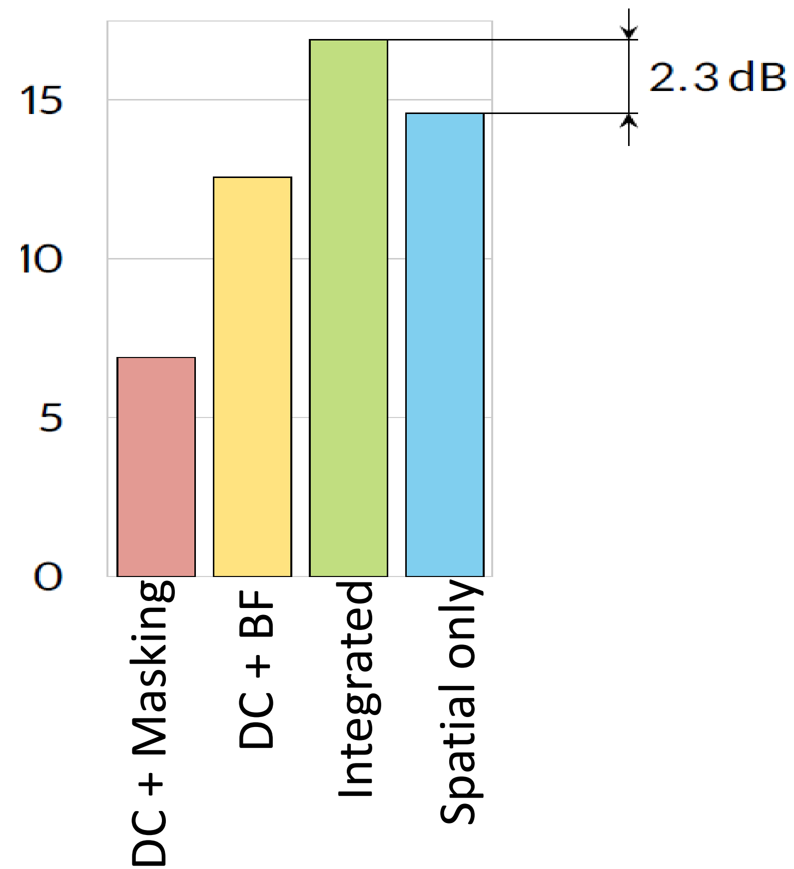
- Train DC on single channel WSJ utterances [Isik, 2016]
  - Randomly mixed, 2- and 3-speaker mixtures
- Simulate multi-channel signals (6 channels)
  - Image method to generate RIRs
  - Random source and array positions

Signal-to-Distortion Ratio (SDR) Gain

Low reverb (T_{60} = 50 … 100 ms)



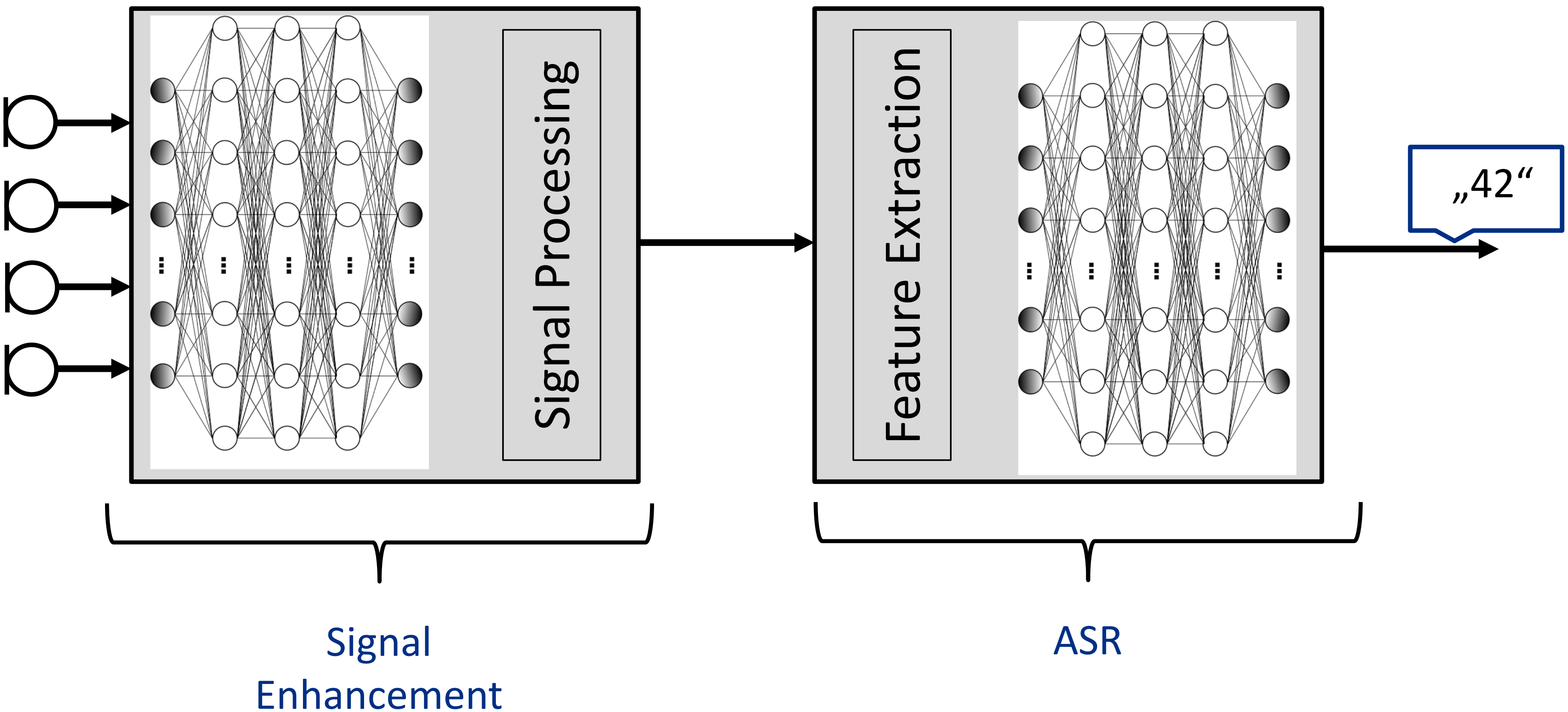
Medium reverb (T_{60} = 200 … 300 ms)



UNIVERSITÄT PADERBORN

Integration with ASR

System Setup

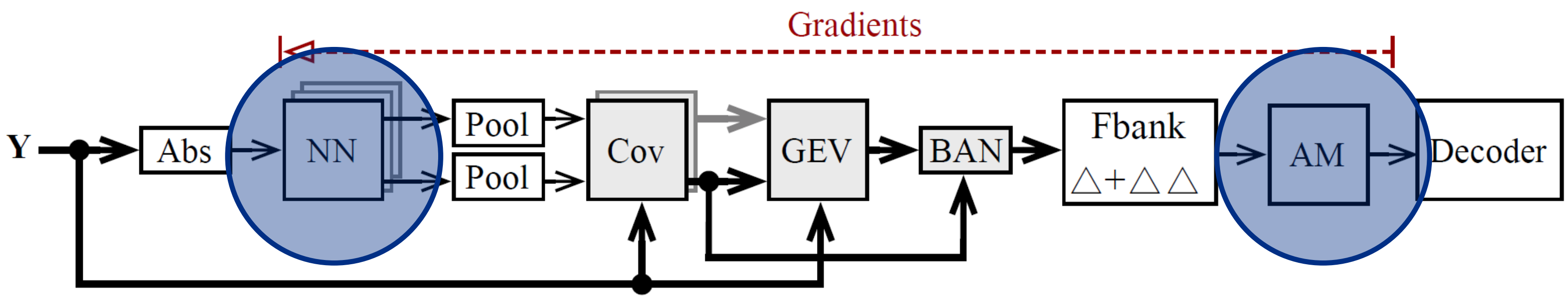


Integration with ASR

- We now have (at least) two neural networks
  - NN in enhancement stage
  - NN as acoustic model of ASR
- With different objective functions
- Advantages of joint training
  - Common objective function
  - No need for parallel data
- Note: Networks are not connected head-to-tail
  - Intermediate processing

Example: NN Supported Beamforming

[Heymann et al. 2017, Boeddeker et al, 2017]



- Gradient through signal processing tasks
  - Feature extraction
  - Beamforming
- Complex-valued gradients

WER Results (1/2)

CHiME-4:

Beamformer – AM trng	Eval Simu	Eval Real
BeamformIt – separat	10.2	9.4
separat – separat	4.6	5.8
Joint: scratch – scratch	5.6	8.8
Joint: scratch – finetune	4.1	5.8
Joint: finetune – finetune	3.9	5.4

Parallel data no longer required!

WER Results (2/2)

[Heymann et al. 2018]

Google Voice Search:

	# channels			
	1	**2**	**4**	**8**
Baseline	30.6			
Beamformer		28.4	27.3	27.4
Joint (scratch – scratch)		42.1	38.6	37.8
Joint + mask[1]		37.3	31.8	30.4

[1]: Joint + mask: Beamformer training with IBM mask as additional trng target

- Joint training worse than baseline
  - ➢ Overfitting to the specific characteristics of the beamformer?
  - ➢ Too much variability removed?

Conclusions

- NN *supported* signal enhancement for multi-channel input is advantageous
  - ➢ Model serves as regularizer
- Unsupervised vs supervised approaches
  - ➢ Supervised approaches tend to be more powerful, but require parallel data
  - ➢ Unsupervised approaches are more versatile, however limited in performance
- Joint training has to be taken with care

References

- [Heymann et al., 2018]: J. Heymann, M. Bacchiani, T. Sainath: Performance of mask-based statistical beamforming in a smart home scenario, in Proc. ICASSP 2018
- [Boeddeker et al., 2017]: C. Boeddeker, P. Hanebrink, L. Drude, J. Heymann, R. Haeb-Umbach: Optimizing neural-network supported acoustic beamforming by algorithmic differentiation, in Proc. ICASSP 2017
- [Heymann et al., 2017]: J. Heymann, L. Drude, C. Boeddeker, P. Hanebrink: Beamnet: end-to-end training of a beamformer-supported multi-channel ASR system, in Proc. ICASSP 2017
- [Drude & Haeb-Umbach, 2017]: L. Drude, R. Haeb-Umbach: Tight integration of spatial and spectral features for BSS with deep clustering embeddings, in Proc. Interspeech 2017
- [Chinaev et al., 2016]: A. Chinaev, J. Heymann, L. Drude, R. Haeb-Umbach: Noise-presence-probability-based noise PSSD estimation by using DNNs, in Proc. ITG Conference on Speech Communication, 2016
- [Heymann et al, 2016]: J. Heymann, L. Drude, R. Haeb-Umbach: Neural network based spectral mask estimation for acoustic beamforming, in Proc. ICASSP 2016
- [Heymann et al., 2015]: J. Heymann, L. Drude, A. Chinaev, R. Haeb-Umbach: BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge, in Proc. ASRU 2015
- [Tran & Haeb-Umbach, 2010]: D.H. Tran Vu, R. Haeb-Umbach: An EM approach to multichannel speech separation and noise suppression, in Proc. IWAENC, 2010
- [Tran & Haeb-Umbach, 2012]: D.H. Tran Vu, R. Haeb-Umbach: Exploiting temporal correlations in joint multichannel speech separation and noise suppression using hidden Markov Models, in Proc. IWAENC, 2012