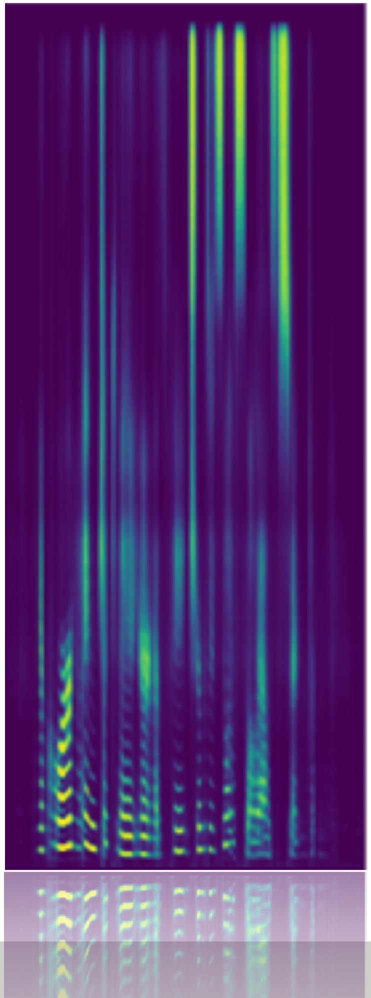
PADERBORN UNIVERSITY
The University for the Information Society

Department of Communications Engineering

Neural Network based Spectral Mask Estimation for Acoustic Beamforming



Jahn Heymann
Lukas Drude
Reinhold Haeb-Umbach

Multi channel processing with neural networks

MOTIVATION

Motivation

- Neural networks are powerful single-channel models
  - Rendered many feature enhancement techniques superfluous
- Few multi-channel approaches
  - Stack channels
  - Work on raw waveforms

- Our approach: Combine neural network with a traditional beamformer

GEV & MVDR

ACOUSTIC BEAMFORMING

Acoustic beamforming

- **MVDR**
  - Minimize noise
  - Source distortionless

- **GEV**
  - Maximize SNR
  - Introduces distortions

$$\underset{\mathbf{F}}{\operatorname{argmin}}\, \mathbf{F}^{\mathrm{H}} \mathbf{\Phi}_{\mathbf{NN}} \mathbf{F} \quad \text{s.t.} \;\; \mathbf{F}^{\mathrm{H}} \mathbf{d} = 1.$$

$$\mathbf{d} = \mathcal{P}\left\{\mathbf{\Phi}_{\mathbf{XX}}\right\}$$

$$\mathbf{F}_{\mathrm{MVDR}} = \frac{\mathbf{\Phi}_{\mathbf{NN}}^{-1} \mathcal{P}\left\{\mathbf{\Phi}_{\mathbf{XX}}\right\}}{\mathbf{d}^{\mathrm{H}} \mathbf{\Phi}_{\mathbf{NN}}^{-1} \mathcal{P}\left\{\mathbf{\Phi}_{\mathbf{XX}}\right\}}$$

$$\underset{\mathbf{F}}{\operatorname{argmax}}\, \frac{\mathbf{F}^{\mathrm{H}} \mathbf{\Phi}_{\mathbf{XX}} \mathbf{F}}{\mathbf{F}^{\mathrm{H}} \mathbf{\Phi}_{\mathbf{NN}} \mathbf{F}}$$

$$\mathbf{\Phi}_{\mathbf{XX}} \mathbf{F} = \lambda \mathbf{\Phi}_{\mathbf{NN}} \mathbf{F}$$

Acoustic beamforming

- Both beamformers depend only on signal statistics
  - Cross-Power Spectral Density of speech and noise
  - Independent of microphone array
  - No assumption on acoustic transfer function
- We estimate PSD matrices using masks

$$\mathbf{\Phi}_{\nu\nu} = \frac{1}{T}\sum_{t=1}^{T} M_\nu(t)\mathbf{Y}(t)\mathbf{Y}(t)^{\mathrm{H}} \quad \text{where} \quad \nu \in \{X, N\}$$

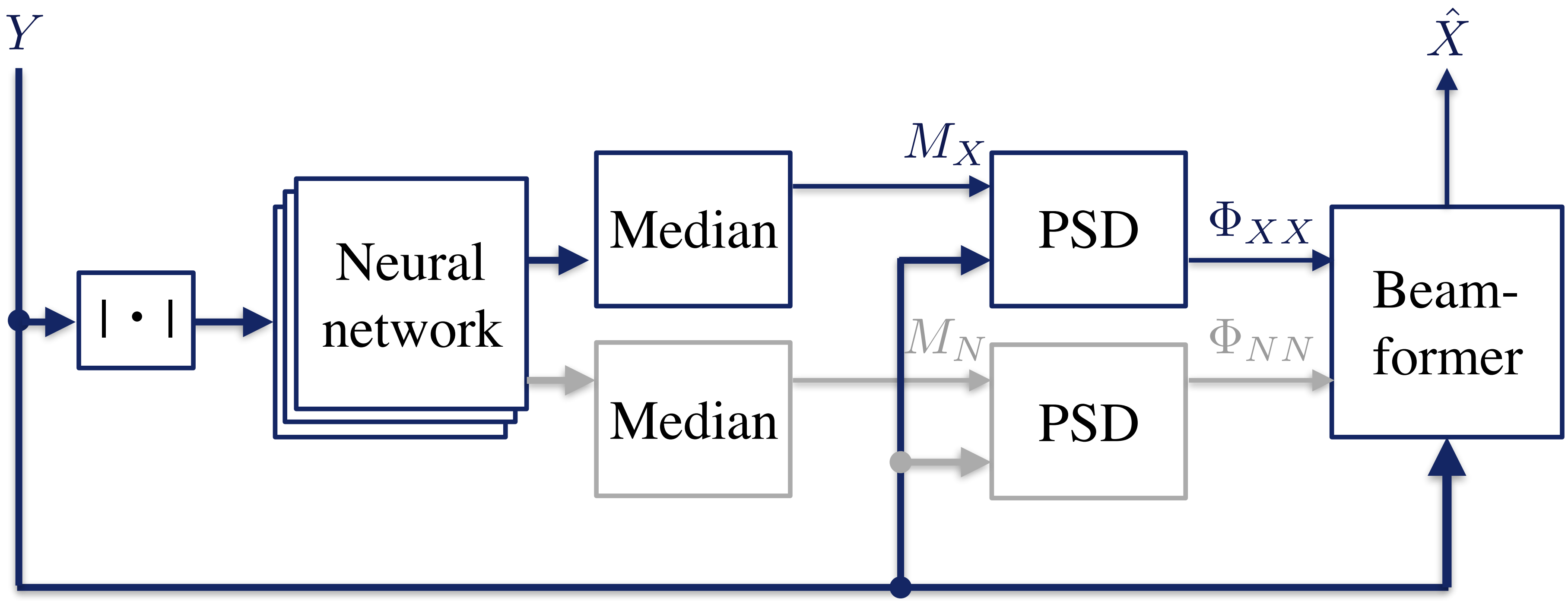
- This allows us to incorporate a neural network

Neural mask estimation

SYSTEM OVERVIEW

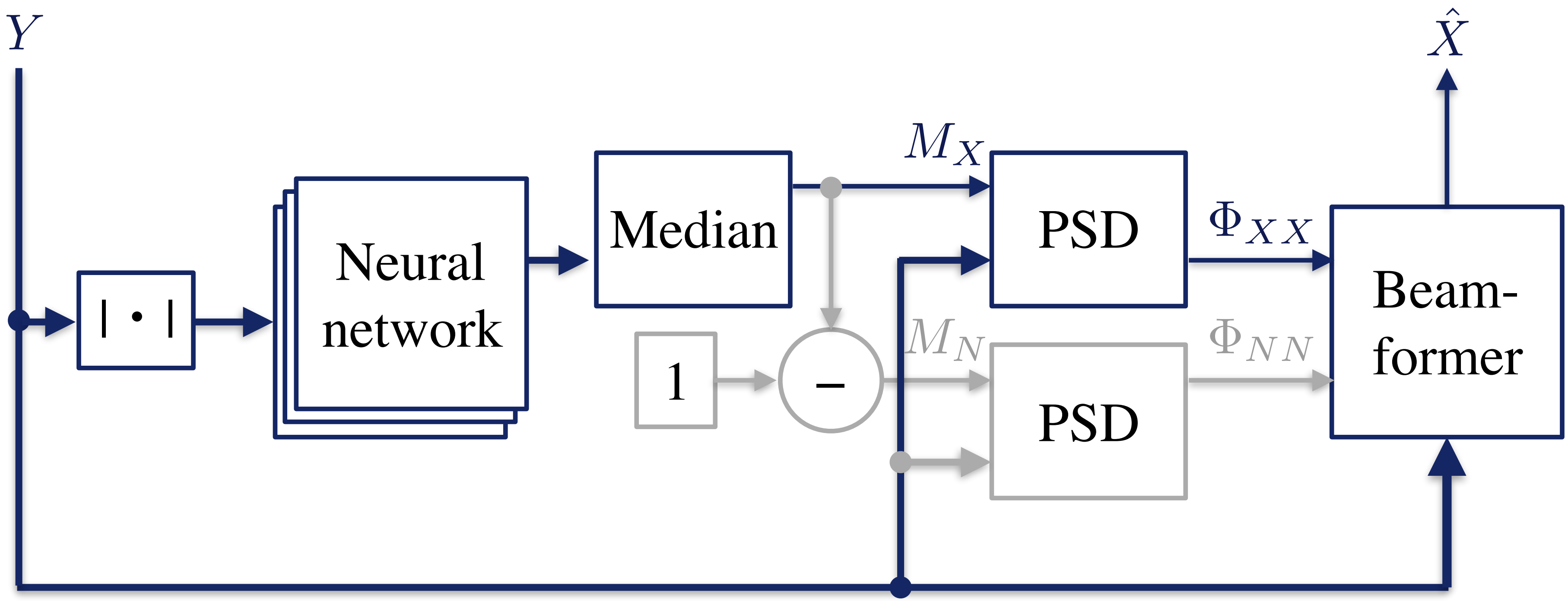
System overview

noise-aware



System overview

clean



System overview

BLSTM

Sophisticated

Layer	Units	Type	Non-linearity	dropout
1	256	BLSTM	Tanh	0.5
2	513	FF	ReLU	0.5
3	513	FF	ReLU	0.5
4	513/1026	FF	Sigmoid	0.0

FF

Simple

Layer	Units	Type	Non-linearity	dropout
1	513	FF	ReLU	0.5
2	513/1026	FF	Sigmoid	0.0

Experimental setup

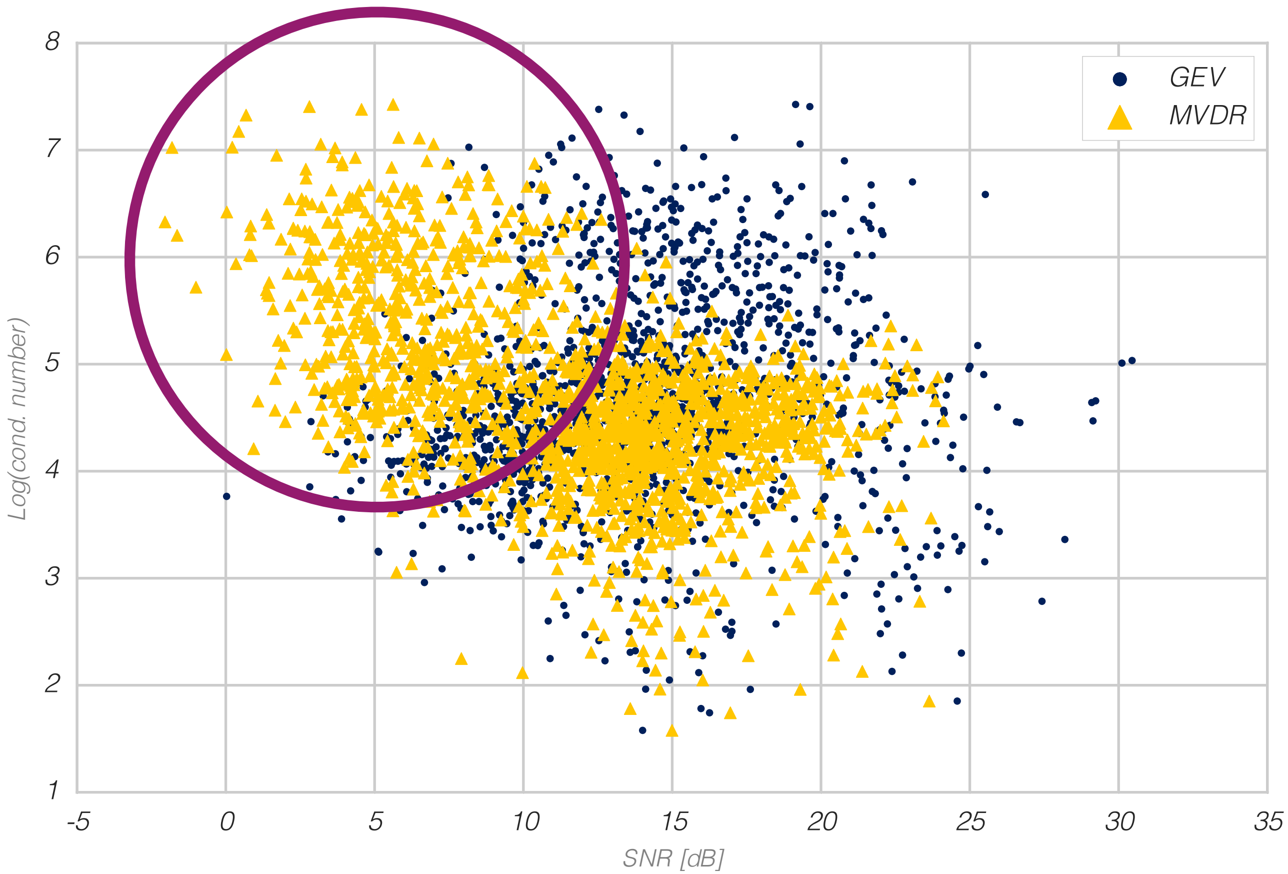
- CHiME III challenge
  - 6 channels
  - 4 different real-world background noise types
- Metrics
  - PESQ / WER
- Compared to
  - Parametric source separation approaches
    - [Tran10]
    - [Ito13]
  - BeamformIt! (only ASR)

MVDR vs. GEV, Speech Enhancement, Speech Recognition

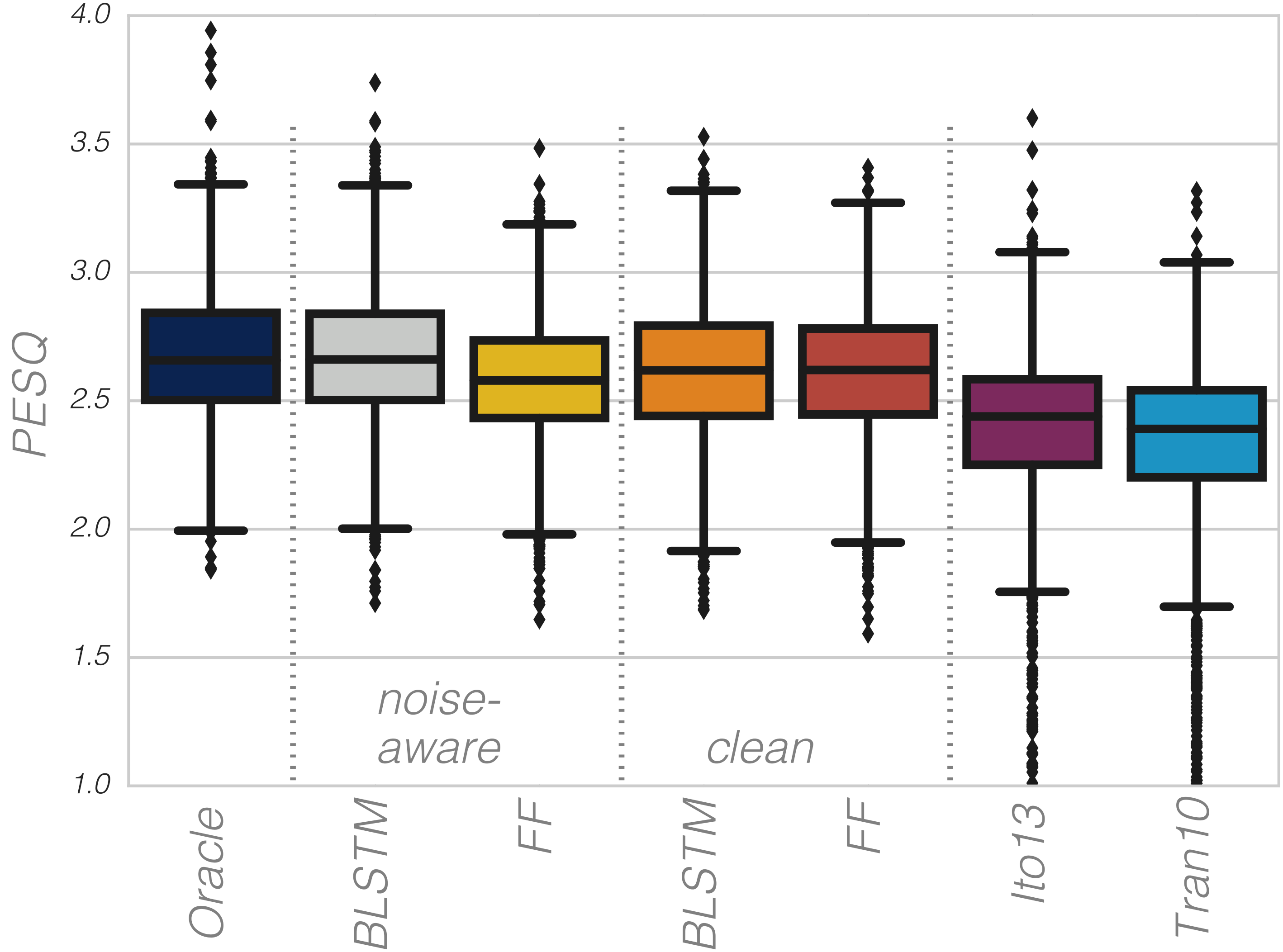
RESULTS

Results

- GEV works better with our masks as it avoids the matrix inversion



Results


4.0
3.5
3.0
2.5
2.0
1.5
1.0
PESQ
noise-
aware
clean
Oracle
BLSTM
FF
BLSTM
FF
Ito13
Tran10

Results

	Evaluation *real*	
	clean	noise-aware
Baseline	40.17	
BLSTM	22.28	**15.42**
FF	21.93	**17.85**
BeamformIt!	22.65	
Ito13	27.32	
Tran10	22.70	

BeamformIt!*	12.79	
BLSTM*		7.45

CONCLUSIONS

Conclusion

- Beamformer supported by Neural Network
- Significant performance gains
- Independent of microphone array configuration
- Small & simple network possible
- Robust against mismatch conditions

Code available:
https://github.com/fgnt/nn-gev