PADERBORN UNIVERSITY

Computer Science, Electrical Engineering and Mathematics

NT Communications Engineering
Prof. Dr.-Ing. Reinhold Häb-Umbach

Wide Residual BLSTM Network with Discriminative Speaker Adaptation for Robust Speech Recognition

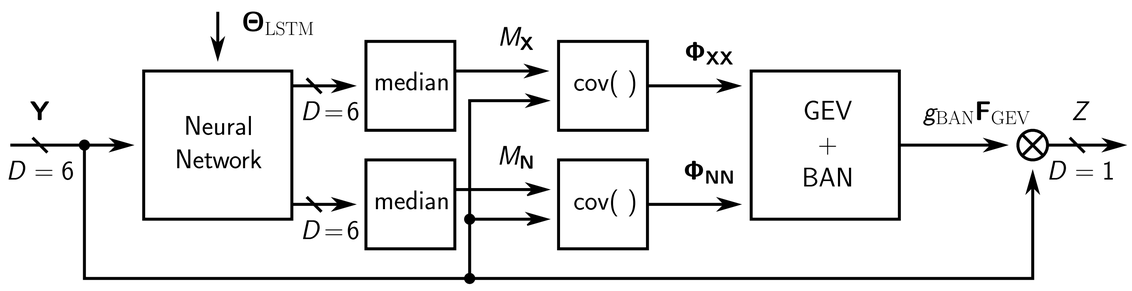
Jahn Heymann, Lukas Drude, Reinhold Haeb-Umbach

Paderborn University
Department of Communications Engineering
Warburger Str. 100, Paderborn, Germany

Introduction

- NN-GEV front-end
- Wide Residual BLSTM back-end
- Custom language model
- Discriminative speaker adaption

Front-end (GEV)



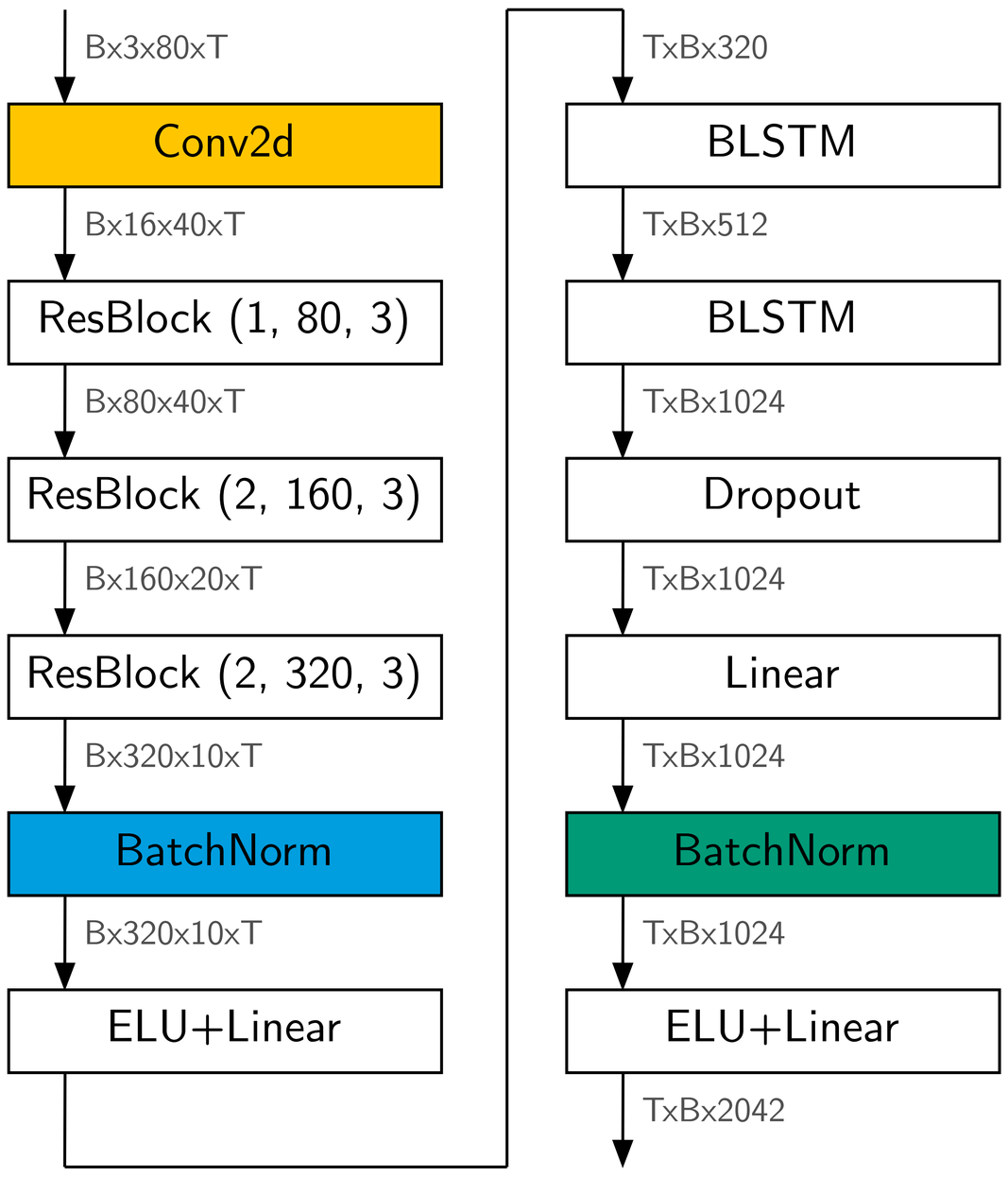
- Neural Network calculates two masks for each channel
- GEV beamformer optimizes SNR for each frequency:

$$\mathbf{F}_{\mathrm{GEV}}(f) = \arg\max_{\mathbf{F}} \frac{\mathbf{F}^{\mathrm{H}}\mathbf{\Phi_{XX}}(f)\mathbf{F}}{\mathbf{F}^{\mathrm{H}}\mathbf{\Phi_{NN}}(f)\mathbf{F}}$$

- Independent of sensor array geometry & speaker location
- Blind analytic normalization to compensate for distortions

Back-end (WRBN)

- Wide residual network part to generate robust features
- Classifier based on two bi-directional LSTM layers
- Works on whole utterance instead of frames
- Trained on all six channels without any pre-processing



Speaker Adaptation (SA)

- Calculate statistics for Batch-Normalization also during decoding
- Add a single transformation matrix (80×80) for each speaker
- Train matrix using backpropagation and alignments from SI model
- Also reduces mismatch between training and processed test data

Language model (LSTM-LM)

- Two-layer LSTM language model
- Sentence wise training without truncated backpropagation
- Number of unknowns in training set unconstrained

Results

- 43 % WER reduction compared to winning CHiME-3 system for 6ch track
- Significant improvements over baseline system

Track	System	Dev		Test	
		real	simu	real	simu
1ch	Baseline	11.57	12.98	23.70	20.84
	WRBN	6.64	9.09	11.8	13.78
	+BN	5.69	7.53	10.4	12.67
	+LSTM-LM	5.69	6.95	9.94	12.35
	+SA	5.19	6.69	9.34	11.11
2ch	Baseline	8.23	9.5	16.58	15.33
	Kaldi+GEV	6.93	8.03	13.76	9.9
	WRBN+GEV	4.67	5.38	7.65	6.53
	+BN	4	4.76	6.96	6.22
	+LSTM-LM	3.76	4.45	6.68	6.19
	+SA	3.54	4.05	5.96	5.16
6ch	Baseline	5.76	6.77	11.51	10.90
	Kaldi+GEV	3.7	3.72	5.66	4.34
	WRBN+BFIT	4.43	5.27	7.33	7.85
	WRBN+GEV	3.16	3.2	4.52	3.41
	+BN	3.06	2.99	4.07	3.51
	+LSTM-LM	2.82	2.61	3.87	3.15
	+SA	2.73	2.34	3.48	2.76

Conclusion

- Even with a strong back-end multi-channel processing (i.e. beamforming) provides a significant gain
- System independent of array configuration
- Simple adaptation leads to notable better performance

Outlook

- Jointly train the AM and the beamformer

This work has been supported by the DFG under contract no. Ha3455/11-1