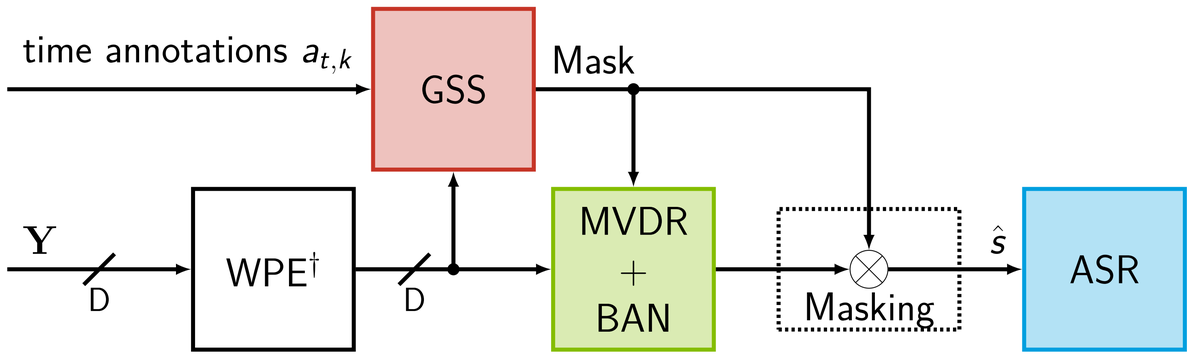
Front-End Processing for the CHiME-5 Dinner Party Scenario

Christoph Boeddeker*, Jens Heitkaemper*, Joerg Schmalenstroeer, Lukas Drude, Jahn Heymann, Reinhold Haeb-Umbach
Department of Communications Engineering, Paderborn University, Germany

Introduction

- WPE, Guided Source Separation (GSS), BF and Masking front-end
- Track A with baseline back end architecture
- No system combination
- Front-end processing based on unsupervised learning techniques without Neural Network (NN)

Overview



† https://github.com/fgnt/nara_wpe

MVDR beamformer

$$\mathbf{w}_f(\mathbf{r}) = \frac{\mathbf{\Phi}_{NN_f}^{-1}\mathbf{\Phi}_{XX_f}\mathbf{r}}{\mathrm{tr}\left\{\mathbf{\Phi}_{NN_f}^{-1}\mathbf{\Phi}_{XX_f}\right\}}, \qquad \mathbf{r} = \underset{\mathbf{e}}{\mathrm{argmax}} \left\{\frac{\mathbf{w}_f^{\mathsf{H}}(\mathbf{e})\mathbf{\Phi}_{XX_f}\mathbf{w}_f(\mathbf{e})}{\mathbf{w}_f^{\mathsf{H}}(\mathbf{e})\mathbf{\Phi}_{NN_f}\mathbf{w}_f(\mathbf{e})}\right\}$$

$\mathbf{e}$ . . . unit vector

Single array to multiple arrays

Options:
- Process each/reference array independently
- Stack all channels to one big array

Conclusion:
- WPE gains from stacking
- GSS gains from stacking
- Beamforming gains from stacking

Guided Source Separation (GSS)

- Complex angular central Gaussian mixture model
- Time annotations indicate when a speaker is active
  - When inactive: Force the posterior to be zero
  - Avoid frequency permutation
  - Avoid global permutation
- Optional noise class
- In-ear: Session-wise blind source separation with utterance-wise finetuning (No time annotations, no WPE, no beamforming)
- Array: Consider left and right context (Not entire session)

Acoustic Model (AM)

- In-Ear TDNN: Trained on enhanced in-ear signals
- Baseline TDNN: Default TDNN trained on left in-ear and 100k utterances from random arrays

Source Activity Detection (SAD)

- Replacing time annotations $a_{t,k}$ with a NN source activity estimate
- NN input: observations, $a_{t,k}$, source activity mask estimated by GSS
- Only used in final results

Conclusion

- Speech enhancement can provide a significant gain already without modified AM architecture
- The only front-end with significant gains from multi array (see RWTH/UPB)
- GSS is suitable for scenarios like CHiME 5
- Main gains without a NN in the front-end
- Achives the 3rd best WER with a stronger back-end (see RWTH/UPB)

Dev WER: Single array track (w/o SAD)

Context in s	WPE	GSS	GSS Noise class	Beamforming	In-Ear TDNN w/o Masking	In-Ear TDNN w Masking	Baseline TDNN w/o Masking	Baseline TDNN w Masking
0	-	-		BeamformIt	96.98	-	81.69	-
				Sum Channels	93.94	-	81.13	-
15	-	✓	-	Channel 3	94.72	86.49	82.28	74.67
			✓			80.84		84.84
0	-	✓	-	MVDR + BAN	96.71	96.87	86.31	86.57
0			✓		94.65	95.09	81.98	82.56
2			-		89.08	86.40	78.16	76.58
2			✓		86.52	80.43	76.16	81.50
15			-		88.20	83.30	77.23	74.42
15			✓		84.86	79.89	75.11	84.76
2	✓	✓	-	MVDR + BAN	85.66	82.71	76.97	75.10
2			✓		82.69	**77.15**	74.82	78.87
15			-		85.11	79.35	74.42	**73.08**
15			✓		82.02	77.21	74.08	83.12

Dev WER: Multiple array track (Context 15 s, noise class and w/o SAD)

Stack all channels WPE	GSS	MVDR + BAN	In-Ear TDNN w/o Masking	In-Ear TDNN w Masking	Baseline TDNN w/o Masking	Baseline TDNN w Masking
-	-	-	82.02	77.21	74.08	83.12
-	-	✓	77.19	77.36	69.98	83.17
-	✓	-	71.95	66.24	65.50	78.78
✓	✓	✓	67.93	**62.51**	**64.41**	76.80

-14.7 -9.67

Final results

Track A	Dev	Eval
single	71.43	69.60
multi	61.73	68.98

-9.02

5th CHiME Workshop 2018, Hyderabad, India

{boeddeker, heitkaemper, schmalen, drude, heymann, haeb}@nt.upb.de