PADERBORN UNIVERSITY

NTT

Part VI.
Other Topics

Reinhold Haeb-Umbach

Table of contents in part VI

- NN supported enhancement: Overcome need for parallel clean and distorted training data
  - Motivation
  - Joint training
  - Teacher-student approach
  - Direct optimization of likelihood
- Should we do speech enhancement also on the ASR training data?

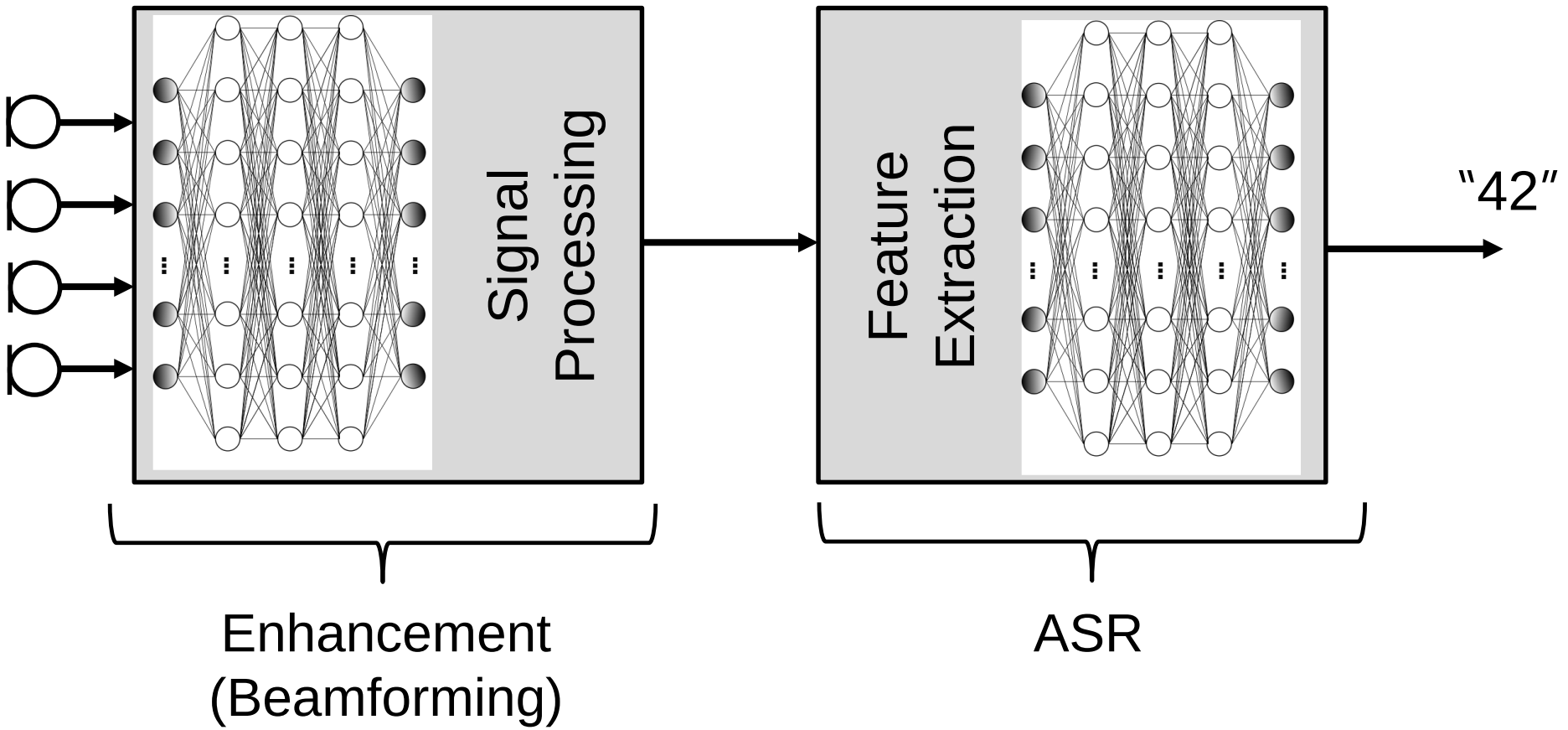
Table of contents in part VI

Motivation

- We have seen different uses of neural networks in enhancement
  - E.g., speech presence probability (mask) estimation
- Those networks were trained by supervised learning
  - Corrupted signal at input
  - Desired/clean signal as target
- This requires parallel (clean and distorted) data
  - Which is unavailable for real recordings of distorted speech
  - Training only on simulated (= artificially distorted) data possible
- Thus
  - No training on real recordings of distorted speech possible
  - Certain effects are hard, if impossible, to realistically simulate
    - e.g., Lombard speech

Goal: Get rid of need for parallel data in NN training!

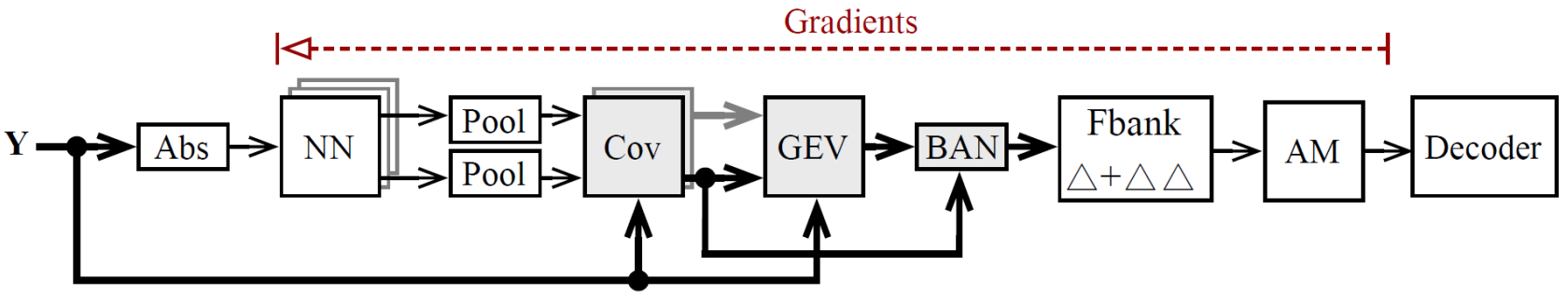
Option 1: Joint training



- Train NNs in front-end and back-end jointly
- Back-propagate gradient of cross entropy loss all the way to enhancement NN

Example NN-supported beamforming

[Heymann et al. 2017a, Ochiai et al., 2017]



- Gradient passed through signal processing tasks
  - ASR feature extraction
  - Beamforming
- Complex-valued gradients
  - See [Boeddeker et al., 2017] for a large collection of complex-valued gradients of various operations

Discussion

- Possible advantages of joint training
  - Parallel clean and noisy data no longer required
  - Training on real recordings of distorted speech
  - Mask estimator trained with criterion closer related to WER

- Possible disadvantages of joint training
  - Weaker acoustic model (AM)
    - Beamforming reduces the number of input channels to one. Thus fewer training data for acoustic model (AM)
    - Beamforming improves SNR, thus AM exposed to less variability
  - Weaker beamformer
    - AM learns to ignore certain distortions, thus beamformer does not need to remove them, meaning that beamforming is less effective in cleaning the data

WER results on CHiME-4

	Beamformer trng	AM traning	Eval Simu	Eval Real
parallel data required				
(a)	i) independent	i) independent on unenh. data	6.8	7.3
(b)	i) independent	ii) indep. on enhanced data	6.6	8.9
no parallel data required				
(c)	i) jointly from scratch	i) jointly from scratch	6.9	9.1
(d)	ii) using gradient from AM	i) separate on unenh. data	7.4	7.6

Training order: first i), then ii)

(a) & (c) Joint training degrades performance, in particular on real data

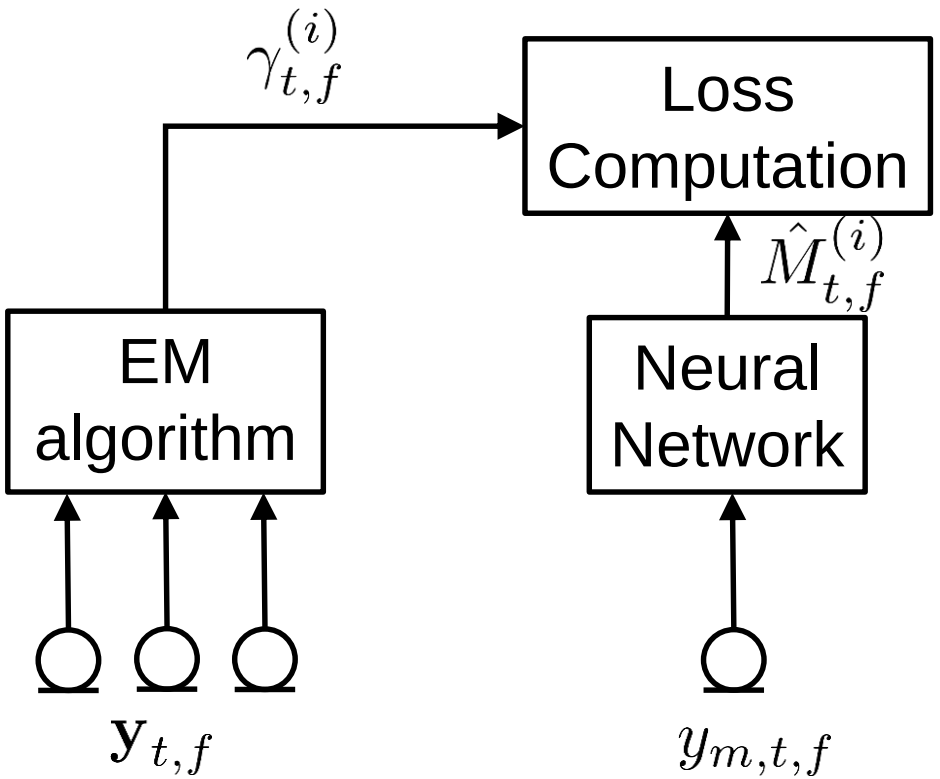
(b) & (d) The cause appears to be the weaker AM;

degradation can be reduced if AM sees enough variability in training

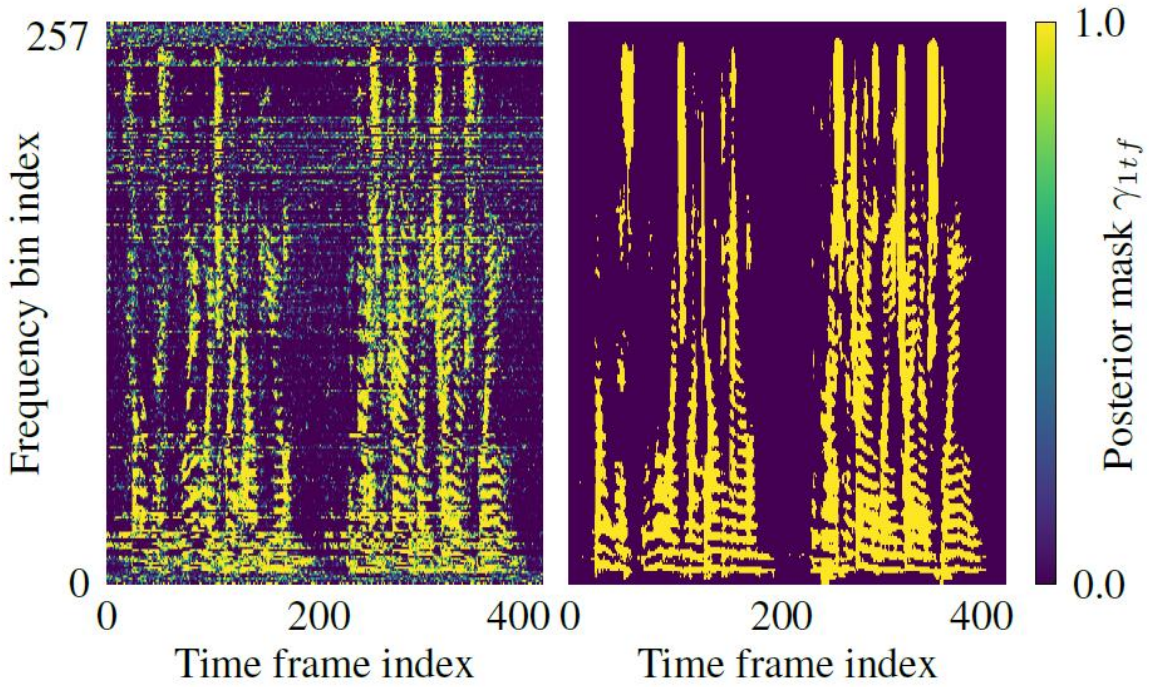
Option 2: Teacher – student approach

[Drude et al., 2019a, Seetharaman et al., 2019, Tzinis et al., 2019]

- Speaker presence probs ($\gamma_{t,f}^{(i)}$) obtained from spatial mixture model used as training targets of NN mask estimator



Example result for BSS [Drude et al., 2019a]



Teacher:
spatial mixture model

Student:
neural network

Results [Drude et al., 2019a]

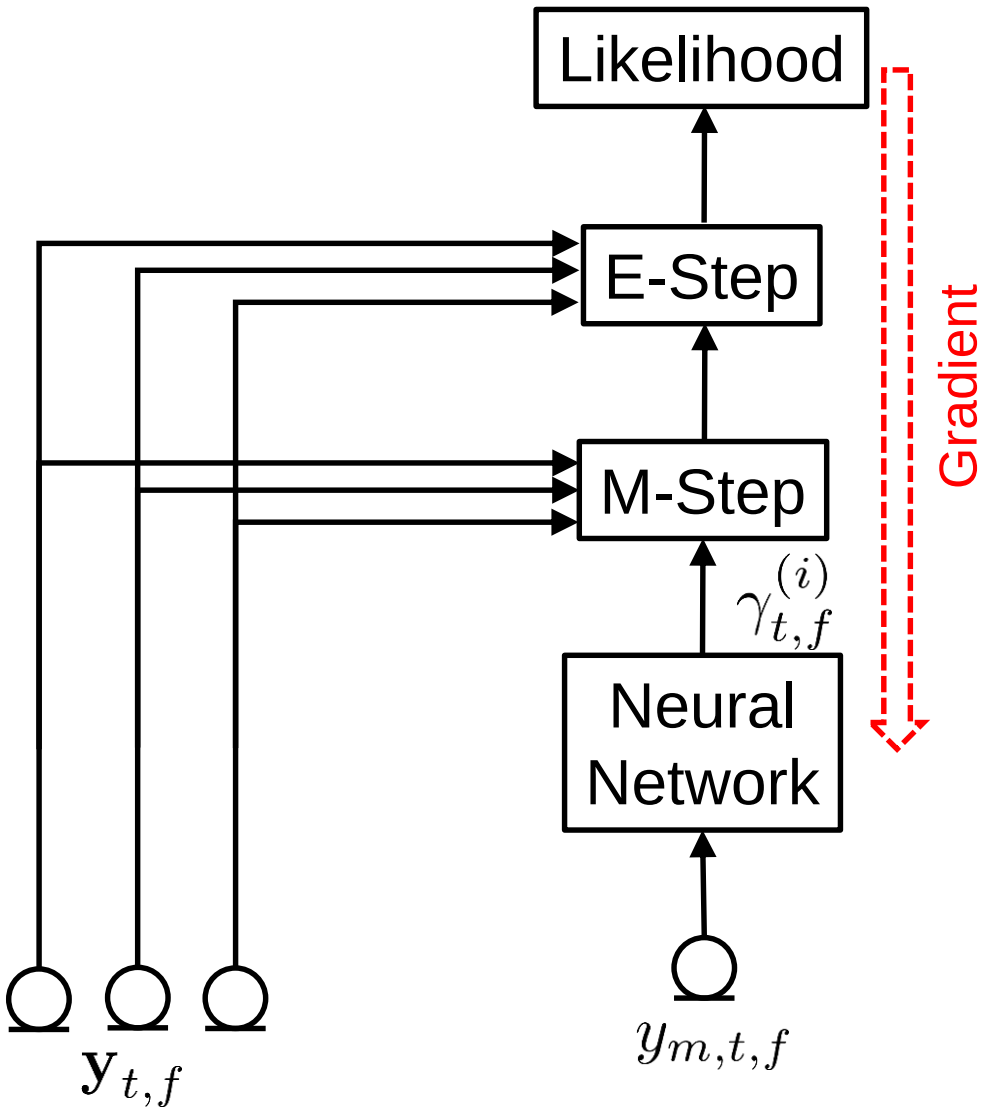
- Database: spatialized multi-channel wsj-2mix
- Source extraction via beamforming

	Model	Training	Initialization on test utt.	WER [%]
(a)	spatial mixt. model	-	random	28.0
(b)	deep clustering	Supervised	-	26.5
(c)	deep clustering	taught by mixt. model	-	29.0
(d)	spatial mixt. model	-	deep clustering from (c)	20.7
(e)	spatial mixture model	-	oracle ideal binary mask	19.9

(d) On test utterance, first apply DC to obtain initial values for $\gamma_{t,f}^{(i)}$. Then run EM to obtain updated $\gamma_{t,f}^{(i)}$.

Option 3: Direct optimization of likelihood

[Drude et al., 2019b, session Tue-O-3-5]



- Optimize likelilhood of spatial mixture model
- Backpropagate gradient of likelihood through E-step and M-step of spatial mixture model to class affiliation posteriors and then to NN parameters
- Optional: additional EM-step at inference time on test utterance

Results [Drude et al., 2019b]

- Beamforming
- CHiME-4 real test set
- Additional EM step on test utterance

Estimator of $\gamma_{t,f}^{(i)}$	**Training**	**WER [%]**
spatial mixture model	-	13.0
neural network	Oracle masks	7.7
neural network	teacher-student	7.9
neural network	likelihood	7.8

Table of contents in part VI

- NN supported enhancement: Overcome need for parallel training data
  - Motivation
  - Joint training
  - Teacher-student approach
  - Direct optimization of likelihood
- Should we do speech enhancement on the ASR training data?

Enhancement on ASR training data?

Pros:

- Acoustic model can learn artifacts of the enhancement
- Cleaner training data → better alignments → better models

Cons:

- Acoustic model is exposed to less variability
- Can reduce the amount of training data (e.g., if only the beamformed signal is used for training instead of all raw channels)

Example results

- Beamforming on CHiME-4

	Training Data	WER [%] Eval Simu	WER [%] Eval Real
(a)	all six channels	6.8	7.3
(b)	all six channels + beamformed	6.4	7.7
(c)	single channel	6.9	7.6
(d)	beamformed only	6.9	9.6
(e)	clean	11.7	16.3

(a) & (d) enhancement in trng hurts performance, in particular on real data

(c) & (d) The reason is not fewer trng data, but removal of variability

But look at these results

- CHiME-5
  - Extremely degraded: lots of overlapped speech, reverberation, …
  - Weak enhancement: (BeamformIt: variant of Delay-Sum-Beamformer)
  - Strong: guided source separation [Kanda et al., 2019, session Tue-O-3-5]

WER [%] on eval	Enhancement in Test		
Enhancemnt in Trng	**none**	**weak**	**strong**
none	59.9	59.7	51.6
weak (BeamformIt)	59.1	58.5	49.9
strong (GSS)	73.1	69.2	45.7

- Matched is best
- Enhancement in trng beneficial, as long as it is weaker than in test
- If data is extremely poor, enhance for alignment extraction, not for NN training itself

Summary of part VI

- There are several options to avoid the need for parallel clean and noisy training data
  - Direct optimization of likelihood is the (arguably) conceptually most appealing one
    - Sofar only developed for beamforming
  - Joint training of front end NN and acoustic model is tricky

- Enhancement of ASR training data
  - Is only advisable as long as the training data contains still at least as much variability as the test data

Table of contents