Technische Universität Braunschweig
Institut für Nachrichtentechnik

Acoustic Model Fusion for Phoneme Recognition According to the Turbo Principle

LISTEN Workshop / Summer School

Tim Fingscheidt
(with contributions from W. Li, T. Lohrenz, S. Receveur, D. Scheler, R. Weiss)

1 A Brief Introduction to the Turbo Principle (Digital Communications)

Turbo principle introduced into forward error correction (FEC) by Berrou et al., 1993

- 2 parallel (weak) convolutional encoders operating on interleaved bitstreams
- 2 parallel convolutional decoders operating iteratively

  (modified Viterbi algorithm or modified BCJR algorithm applied)
- In the decoder: Iterative exchange of (a posteriori) probabilities / likelihood ratios
- Error performance very close to theoretical bounds

Today, turbo codes are known in many variants and are deployed in many communication systems (3G, LTE, ….)

NEAR SHANNON LIMIT ERROR - CORRECTING CODING AND DECODING : TURBO-CODES (1)

Claude Berrou, Alain Glavieux and Punya Thitimajshima

Claude Berrou, Integrated Circuits for Telecommunication Laboratory

Alain Glavieux and Punya Thitimajshima, Digital Communication Laboratory

Ecole Nationale Supérieure des Télécommunications de Bretagne, France

(1) Patents N° 9105279 (France), N° 92460011.7 (Europe), N° 07/870,483 (USA)

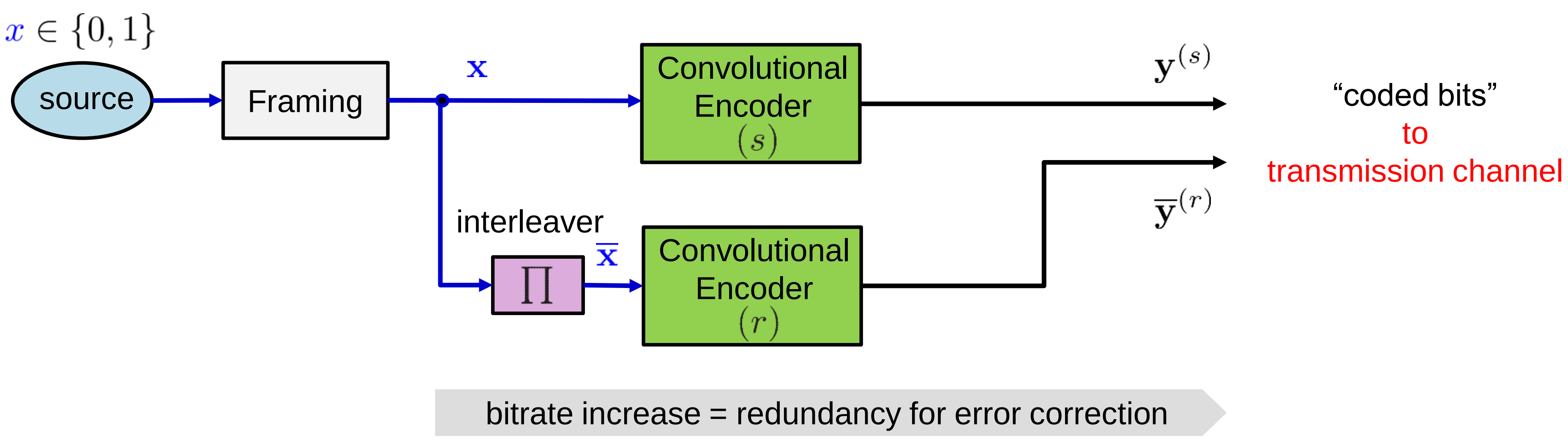
Abstract - This paper deals with a new class of convolutional codes called *Turbo-codes*, whose performances in terms of Bit Error Rate (BER) are close to the SHANNON limit. The Turbo-Code encoder is built using a parallel concatenation of

$$P_r\{a_k = 0 / a_1 = \varepsilon_1, .. a_{k-1} = \varepsilon_{k-1}\} = P_r\{d_k = \varepsilon\} = 1/2 \quad (4)$$

with ε is equal to

The Encoder (simplified)

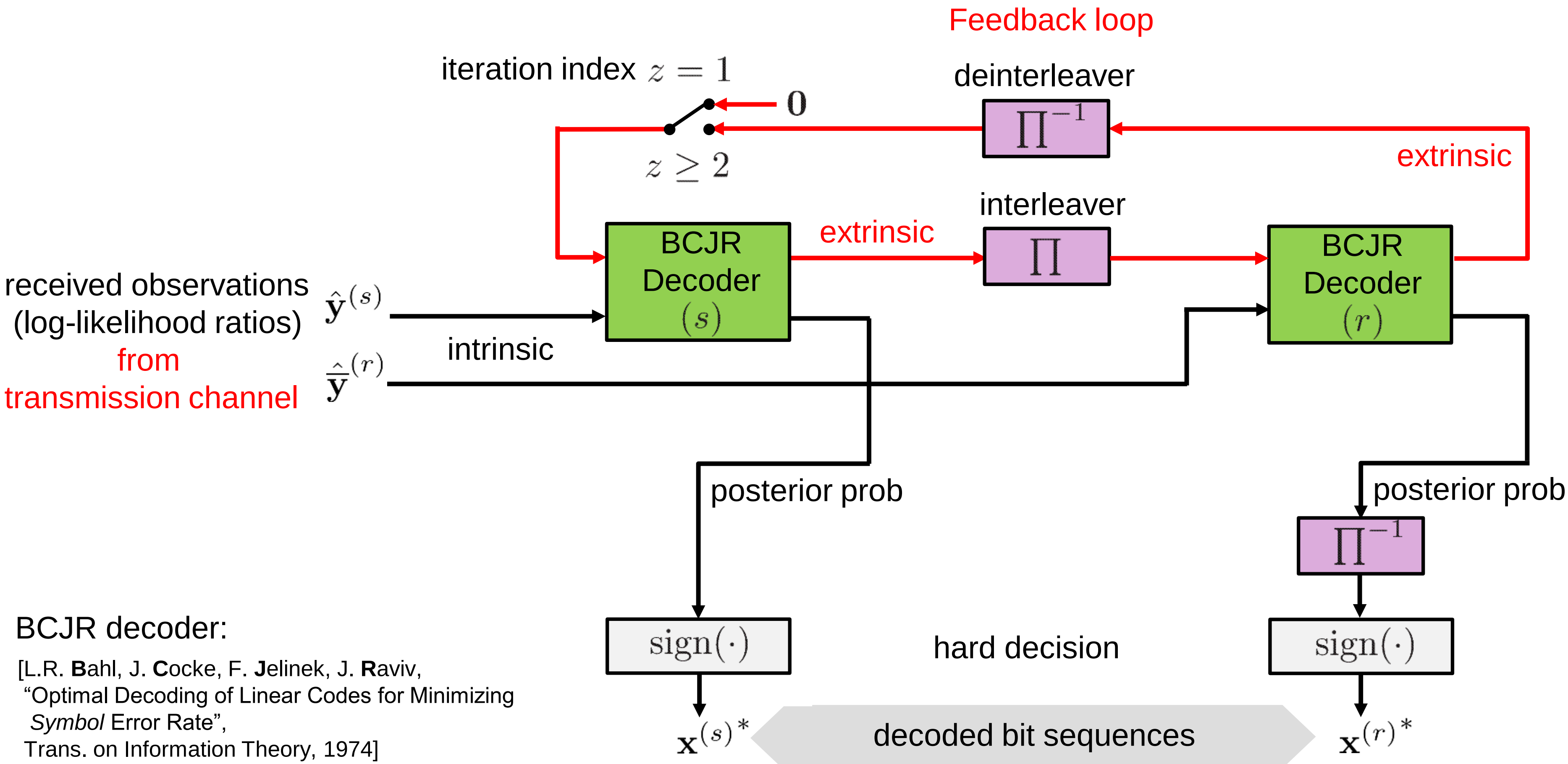
2 parallel (weak) convolutional encoders:



A Brief Introduction to the Turbo Principle

The Decoder

2 parallel convolutional decoders operating iteratively (BCJR or soft-input/soft-output Viterbi):



BCJR decoder:

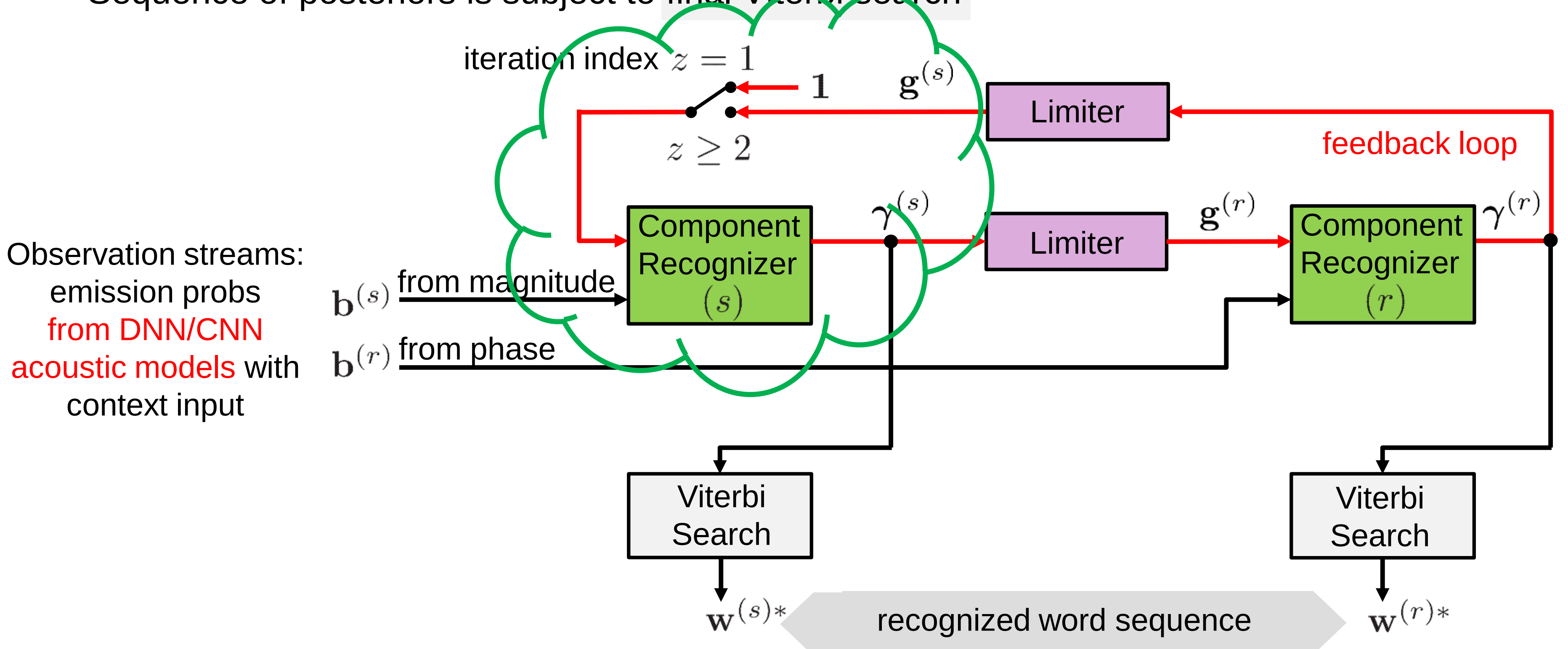
[L.R. **B**ahl, J. **C**ocke, F. **J**elinek, J. **R**aviv,
"Optimal Decoding of Linear Codes for Minimizing *Symbol* Error Rate",
Trans. on Information Theory, 1974]

2 The Turbo Fusion Approach

The Decoder

Turbo fusion of magnitude and phase feature models: Forward-backward algorithms in iteration

- Dynamic range limitation of exchanged information (posteriors) between iterations
- Sequence of posteriors is subject to final Viterbi search

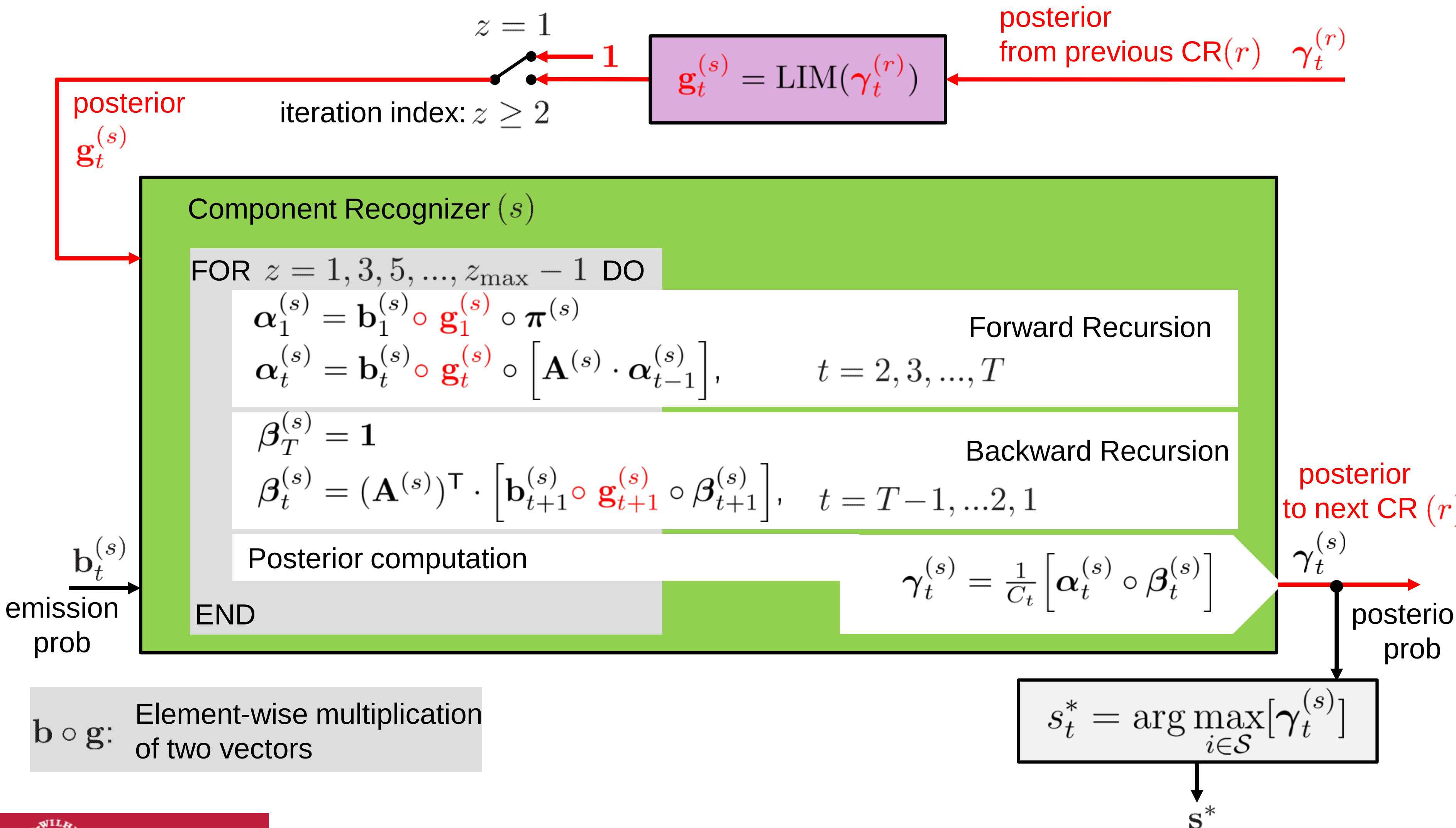


[S. Receveur, R. Weiss, T. Fingscheidt, "Turbo Automatic Speech Recognition", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 5, pp. 846-862, May 2016]

[T. Lohrenz and T. Fingscheidt, "Turbo-Fusion of Magnitude and Phase Information for DNN-Based Phoneme Recognition", in Proc. of ASRU, pp. 118-125, Okinawa, Japan, Dec. 2017]

The Turbo Fusion Approach

The Decoder (Details)



$\mathbf{b} \circ \mathbf{g}$: Element-wise multiplication of two vectors

2 The Turbo Fusion Approach

Limiting the Exchanged Posterior Information (Extrinsic)

In theory, the extrinsic information (posterior) has to be converted to the state space of the subsequent component recognizer by a **state-space transformation matrix**.

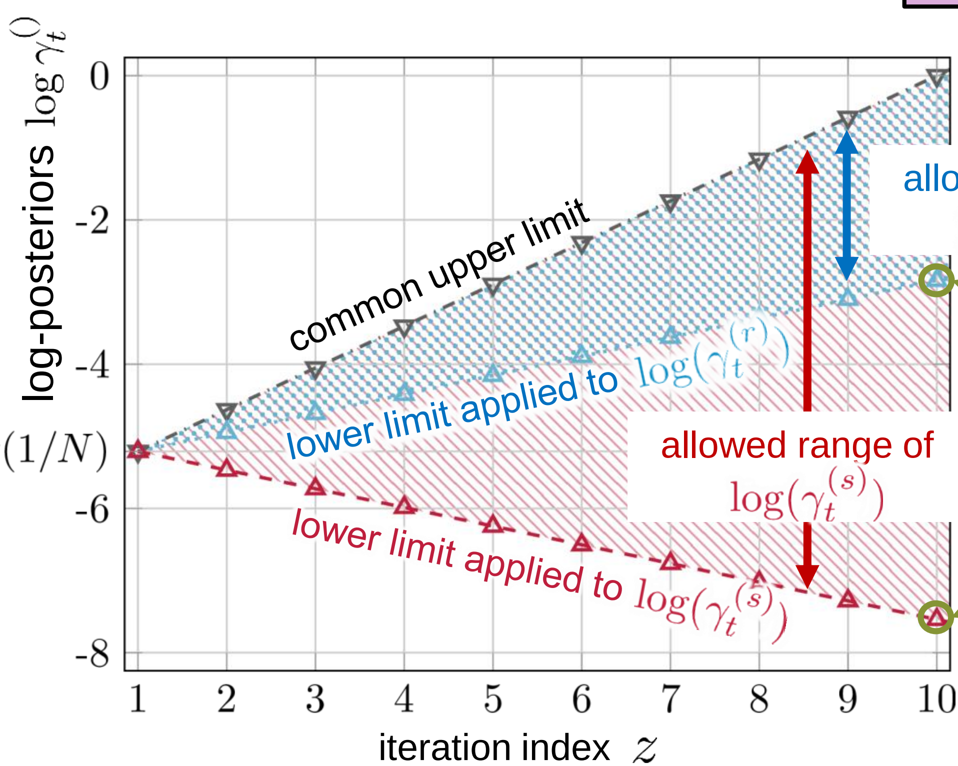
Analogy to the turbo codes: **Interleaver matrix** $\boxed{\Pi}$ also performs a (state) space transform.

For equal state spaces, this can be efficiently done by a simple Limiter function:

$$\mathcal{S} = \mathcal{R}$$

$$|\mathcal{S}| = N$$

Last iteration: No upper limit for extrinsic



2 adjustable hyperparameters = final lower limits

First iteration: Only uniform extrinsic is allowed

3 Simulation Setup
Task Definition, Database and Feature Extraction

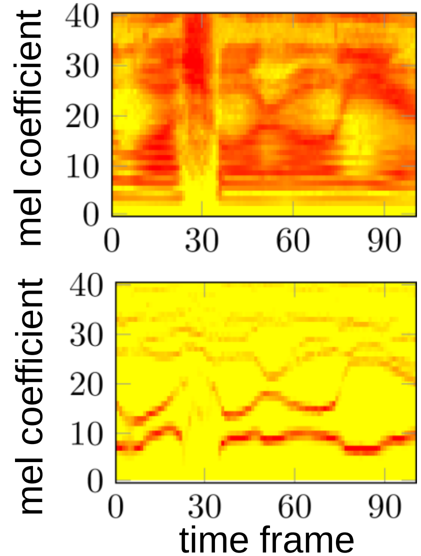
Experiments on the TIMIT dataset of continuous speech (phoneme recognition task)

- 462 speaker training set, 50 speaker development set, and 24 speaker core test set
- 61 phones during decoding, merged to 39 phones for scoring (common practice)
- Hybrid context-independent acoustic monophone models with 3 states per phone (= 183 HMM states)
- Phoneme error rate is evaluated as $\mathrm{PER} = \left(1 - \frac{N-D-I-S}{N}\right) \cdot 100\%$

- Magnitude features: 40 standard mel-filterbank coeffs (FBANK) + $\log E$ with appended $\Delta + \Delta\Delta$
- Proven to be suitable for DNN/CNN processing

- Phase features: 40 mel-filterbank coeffs extracted from all-pole model-based group delay function + $\log E$ + $\Delta + \Delta\Delta$
- Narrow peaks in formant regions yields complementarity

[E. Loweimi, S.M. Ahadi, and T. Drugman, "A new Phase-Based Feature Representation for Robust Speech Recognition", in Proc. of ICASSP, pp. 7155-7159, Sep. 2013]



% phonemes exclusively correct:

6.85%

3.94%

3 Simulation Setup
Acoustic Modelling

Training in clean conditions

Three different model topologies were used:

DNN: *33.5M parameters*

- 8 fully connected layers with 2048 sigmoid units
- RBM initialized weights, using dropout
- 15 frames input context (-7 / +7)

[G.E. Hinton, N Srivastava, A. Krizhevsky, I. Sutskever, and R.R. Salakhutdinov, "Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors", arXiv:1207.0580, pp. 1-18, 2012]

CNN1: *5.4M parameters*

- Limited weight sharing (LWS) for 1-D convolution along spatial dimension
- 15 frames input context (-7 / +7)

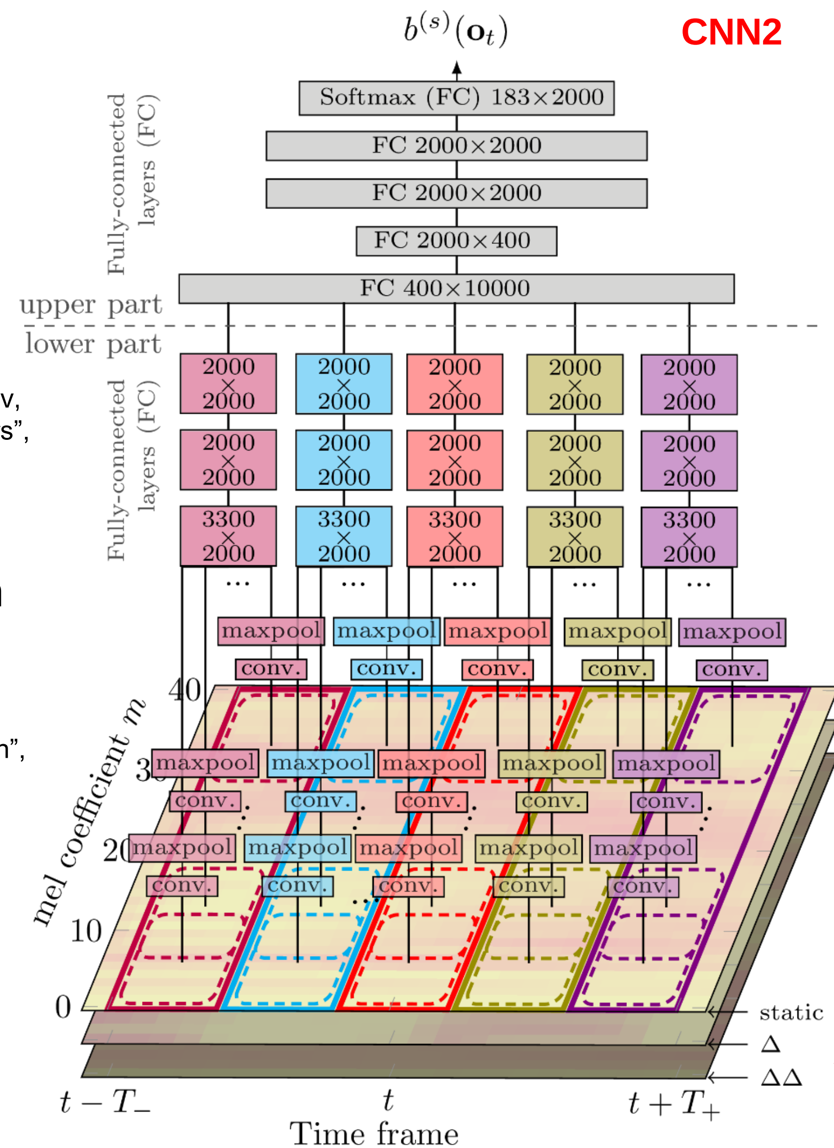
[O. Abdel-Hamid et al., "Convolutional Neural Networks for Speech Recognition", in IEEE/ACM Trans. on ASLP, vol. 22, no. 10, pp.1533-1545, Oct. 2014]

CNN2: *87.1M parameters*

- Hierarchical network with 5 subnetworks
- LWS for 2-D convolution on local sections
- 25 frames input context (-12 / +12)

[O. Abdel-Hamid et al., "Phone Recognition with Hierarchical Deep Maxout Networks", in EURASIP Journ. on ASMP, vol. 2015, no. 1, pp.1-13, 2015]

CNN2



Baselines / Other Fusion Methods

Baselines without information fusion:

- Magnitude features (**-mag**)
- Phase features (**-phase**)

single-model approaches

Baselines with information fusion:

- Feature-level fusion: Feature concatenation (**Fusion-CONCAT**)
- Classifier-level fusion: Synchronuous multi-stream HMMs (**Fusion-MSHMM**):

$$b_i^{(\mathrm{MSHMM})}(\mathbf{o}_t, \mathbf{u}_t) = \left(b_i^{(s)}(\mathbf{o}_t)\right)^{\varphi_s} \cdot \left(b_{k=i}^{(r)}(\mathbf{u}_t)\right)^{\varphi_r} \qquad a_{j,i}^{(\mathrm{MSHMM})} = \xi_s a_{j,i} + \xi_r a_{\ell=j,k=i}$$

(two independent hyperparameters φ_s, ξ_s)

[A.V. Nefian, L. Liang, X. Pi, X. Liu, K. Murphy, “Dynamic Bayesian Networks for Audio-Visual Speech Recognition”, in EURASIP Journal on Applied Signal Processing 11(1), pp. 1274-1288, 2002]

- Classifier-level fusion: Linear combination of emission probs (**Fusion-WA**):

$$b_i^{(\mathrm{WA})}(\mathbf{o}_t, \mathbf{u}_t) = w_s b_i^{(s)}(\mathbf{o}_t) + w_r b_{k=i}^{(r)}(\mathbf{u}_t)$$

(one independent hyperparameter w_s)

[H. Misra, H. Bourlard, V. Tyagi, “New Entropy Based Combination Rules in HMM/ANN Multi-Stream ASR”, Proc. of ICASSP, Hong Kong, China, pp. 741-744, 2003]

- Decision-level fusion: Recognizer output voting error reduction (**Fusion-ROVER**)

(maximum *weighted* confidence scoring, one independent hyperparameter)

[B. Hoffmeister, T. Klein, R. Schlüter, H. Ney, “Frame Based System Combination and a Comparison With Weighted ROVER and CNC”, in Proc. of Interspeech, Pittsburgh, PA, USA, pp. 537-540, Sep. 2006]

Turbo Fusion and Other Fusion Methods

Phoneme Error Rate (% PER)

Benchmarking of DNN-based turbo fusion w.r.t. [Hinton et al.], single-model DNN baselines, and some simulated fusion baselines

	Dev Set	Core Test Set	
DNN	-	19.70	[Hinton et al.]
DNN-mag	18.24	19.85	(our impl. of Hinton et al.)
DNN-phase	21.32	23.29	
Fusion-CONCAT	19.92	21.58	
Fusion-WA	18.04	19.74	
Fusion-MSHMM	17.94	19.62	
Fusion-ROVER	18.34	20.01	
T-Fusion-DNN+DNN (start: mag)	**17.90**	**19.46**	[Lohrenz, Fingscheidt]
T-Fusion-DNN+DNN (start: phase)	18.15	19.78	[Lohrenz, Fingscheidt]

[G.E. Hinton, N Srivastava, A. Krizhevsky, I. Sutskever, and R.R. Salakhutdinov, "Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors", arXiv:1207.0580, pp. 1-18, 2012]

[T. Lohrenz and T. Fingscheidt, "Turbo-Fusion of Magnitude and Phase Information for DNN-Based Phoneme Recognition", in Proc. of ASRU, pp. 118-125, Okinawa, Japan, Dec. 2017]

4 Simulation Results
Turbo Fusion and Other TIMIT Benchmarks

Phoneme Error Rate (% PER)

Benchmarking of turbo fusion w.r.t. [Hinton et al.], single-model DNN or CNN1 or CNN2 baselines, and various other TIMIT baselines

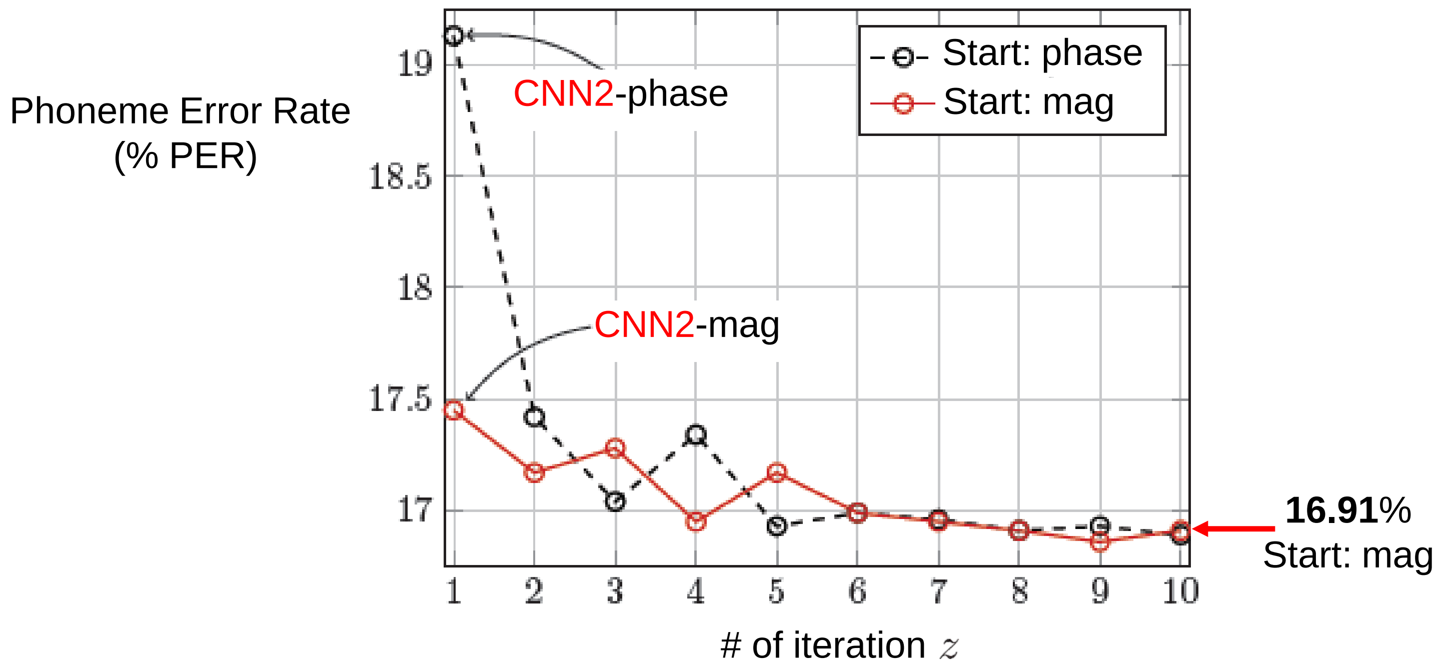
	Core Test Set	
CNN	19.92	[Abdel-Hamid et al., 2014]
DNN	19.70	[Hinton et al., 2012]
T-Fusion-DNN (start: mag)	19.46	[Lohrenz, Fingscheidt, 2017]
CNN2-phase	19.13	Phase baseline of [Lohrenz, Fingscheidt, 2018, subm.]
DNN+RNN	18.80	[Li Deng, Chen, 2014]
T-Fusion-DNN+CNN1	18.80	[Lohrenz, Li, Fingscheidt, accepted for publication 2018]
WaveNet on raw audio	18.80	[v.d. Oord et al., 2016]
Connectionist Temporal Classification	18.40	[Graves, Mohamed, Hinton, 2013]
HMM-BLSTM	17.90	[Graves, Jaitly, Mohamed, 2013]
RNN Transducer	17.70	[Graves, Mohamed, Hinton, 2013]
CNN2-mag	17.45	Magn. baseline of [Lohrenz, Fingscheidt, 2018, subm.]
T-Fusion-CNN2+CNN2 (start: mag)	**16.91** 4.4% rel.	[Lohrenz, Fingscheidt, 2018, submitted]

[T. Lohrenz and T. Fingscheidt, "A New TIMIT Benchmark for Context-Independent Phone Recognition Using Turbo Fusion", submitted to SLT, Athens, Greece, Dec. 2018]

4 Simulation Results

Behavior of Turbo Fusion Over Iterations z

How does T-Fusion-CNN2+CNN2 (magnitude and phase turbo fusion) perform over the iterations $z = 1, 2, ..., 10$?



- Both component recognizers (CRs) reach good consensus after a few iterations
- Even the weaker phase CNN2 model improves the CNN2-mag model after 10 iterations by 17.45% - 16.91% = 0.54% absolute (3.1% relative)

5 Conclusions

Applying the “turbo principle” from Communications to model fusion in ASR:

- Minor modification to the forward-backward algorithm (FBA)
- Feedback of posteriors and simple dynamic range limitation
- HMMs trained separately for each input stream (flexible and scalable towards multiple CRs!)

Performance:

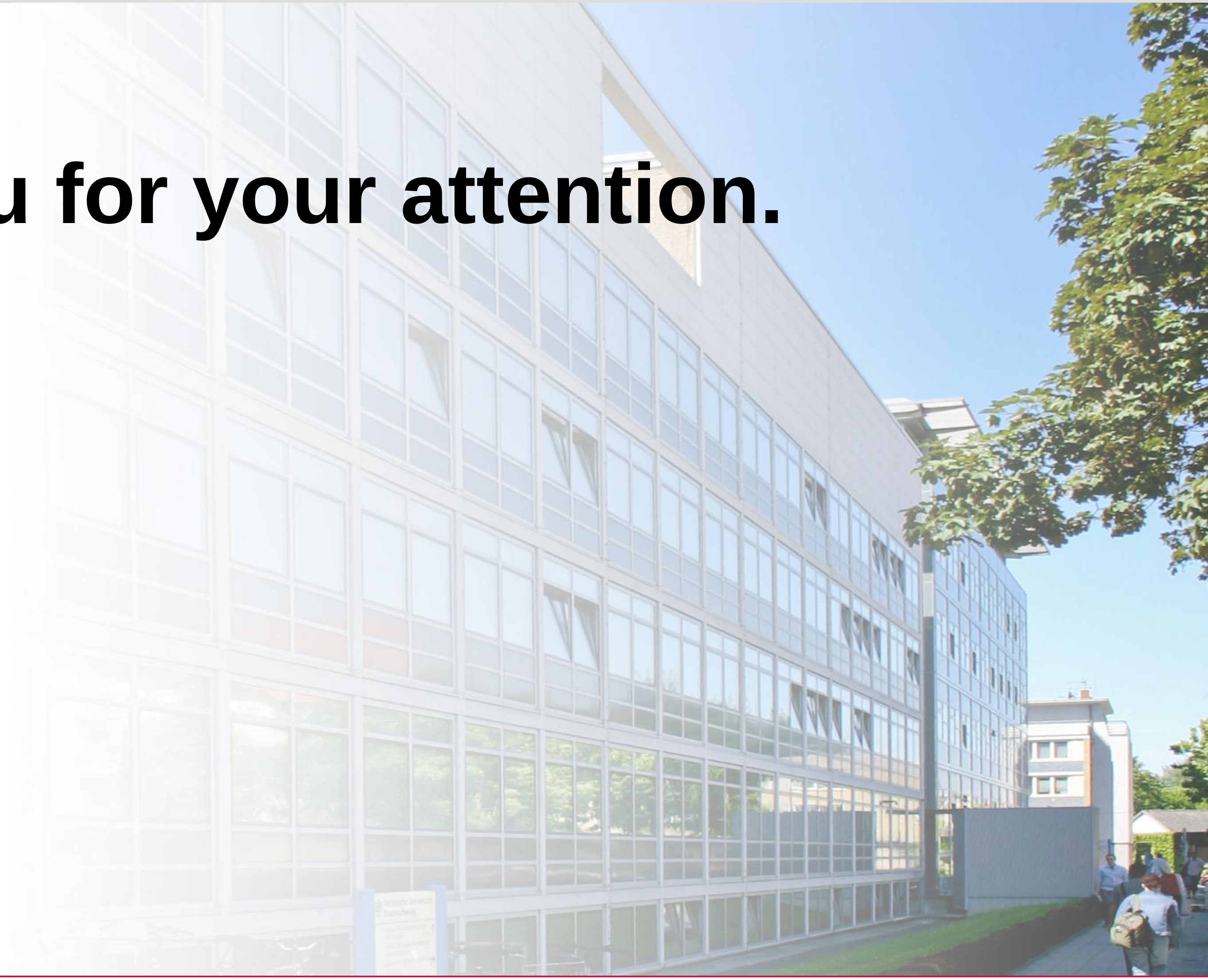
- Control of limiter range over iterations makes the component recognizers “listen and talk” to each other
- Any recognizer output can be used as final result after some iterations
- Turbo fusion of magnitude/phase models outperforms all investigated reference methods on TIMIT:

  16.91% PER for context-independent models

Outlook: Turbo model fusion for ASR …

- … could replace multicondition training in the future
- … could be realized by using BLSTMs instead of the FBA
- … could be used in really distributed intelligence and recognition
- … could be used in acoustic sensor networks

Thank you for your attention.



Tim Fingscheidt

t.fingscheidt@tu-bs.de

References on Iterative ASR and Turbo Fusion for ASR

[S.T. Shivappa, B.D. Rao, M.M. Trivedi, “Multimodal Information Fusion Using the Iterative Decoding Algorithm and its Application to Audio-Visual Speech Recognition”, in Proc. of ICASSP 2008, pp. 2241-2244, Las Vegas, NV, USA, Mar. 2008]

[S. Receveur, T. Fingscheidt, “A Turbo-Decoding Weighted Forward-Backward Algorithm for Multimodal Speech Recognition”, in Proc. of IWSDS 2014, pp. 4--15, Napa Valley, CA, USA, Jan. 2014]

[S. Receveur, T. Fingscheidt, “A Compact Formulation of Turbo Audio-Visual Speech Recognition”, in Proc. of ICASSP 2014, pp. 5554-5558, Florence, Italy, May 2014]

[S. Receveur, R. Weiss, T. Fingscheidt, “Multimodal ASR by Turbo Decoding vs. Feature Concatenation: Where to Perform Information Integration?”, in Proc. of 11. ITG Symposium on Speech Communication, pp. 1-4, Erlangen, Germany, Sep. 2014]

[S. Zeiler, R. Nickel, N. Ma, G.J.. Brown, D. Kolossa, “Robust audiovisual speech recognition using noise-adaptive linear discriminant analysis”, in Proc. of ICASSP 2016, pp. 2797-2801, Shanghai, China, Mar. 2016]

[S. Receveur, R. Weiss, T. Fingscheidt, “Turbo Automatic Speech Recognition”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 5, pp. 846-862, May 2016]

[S. Gergen, S. Zeiler, A. Hussen Abdelaziz, R. Nickel, D. Kolossa, “Dynamic Stream Weighting for Turbo-Decoding-Based Audiovisual ASR”, in Proc. of INTERSPEECH 2016, pp. 2135-2139, San Francisco, CA, USA, Sep. 2016]

[S. Zeiler, H. Meutzner, A. Hussen Abdelaziz, D. Kolossa, “Introducing the Turbo-Twin-HMM for Audio-Visual Speech Enhancement”, in Proc. of INTERSPEECH 2016, pp. 1750-1754, San Francisco, CA, USA, Sep. 2016]

[T. Lohrenz, S. Receveur, T. Fingscheidt, “EXIT Charts for Turbo Automatic Speech Recognition: A Case Study”, in Proc. of 12. ITG Symposium on Speech Communication, pp. 1-5, Paderborn, Germany, Oct. 2016]

[S. Receveur, T. Lohrenz, T. Fingscheidt, “Introducing Block-Wise Processing into Turbo Viterbi ASR”, in Proc. of 12. ITG Symposium on Speech Communication, pp. 1-5, Paderborn, Germany, Oct. 2016]

[T. Lohrenz, T. Fingscheidt, “Turbo Fusion of Magnitude and Phase Information for DNN-Based Phoneme Recognition”, ASRU, pp. 118-125, Okinawa, Japan, Dec. 2017]

[T. Lohrenz, W. Li, T. Fingscheidt, “DNN/CNN Acoustoc Model Turbo Fusion for Phoneme Recognition”, accepted for 13. ITG Symposium on Speech Communication, pp. 1-5, Oldenburg, Germany, October 2018]