Channel-selection for distant-speech recognition on CHiME-5 dataset

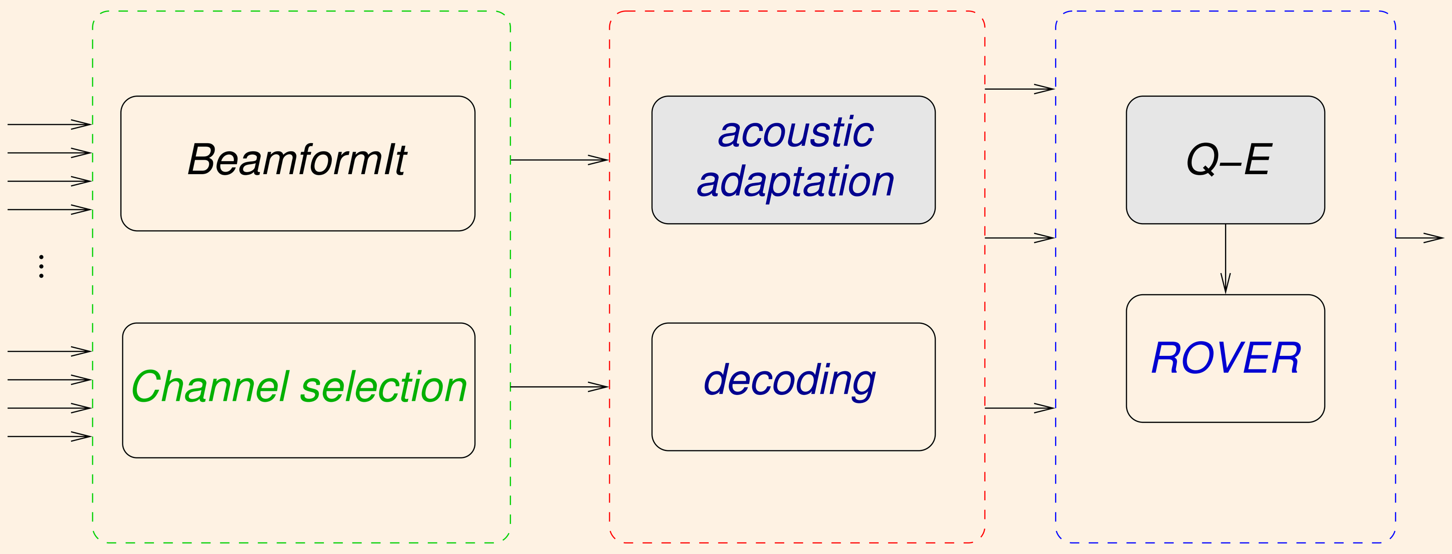
H. Unterholzner[1], L. Pfeifenberger[1], F. Pernkopf[1],
M. Matassoni[2], A. Brutti[2], D. Falavigna[2]

unterholzner@student.tugraz.at, lpfeifen@gmail.com pernkopf@tugraz.at,
matasso@fbk.eu, brutti@fbk.eu, falavi@fbk.eu

[1]Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria
[2]SpeechTek, Fondazione Bruno Kessler, Italy

System Overview

Proposed System:



- **Channel selection using a DNN multi-label classifier: Predicts best channels according to oracle channels**
- **Acoustic model adaptation: based on transfer learning, using a selected subset of the utterances**
- **Automatic quality estimation (Q-E): sentence confidence score**
- **Hypothesis fusion at utterance level with ROVER via majority voting**

Oracle Results

- Theoretical performance gain expected from hypothesis combination.
- Oracle: Upper performance bound by selecting the best hypothesis among a set of decoded channels on utterance-level.
- Using all decoded channels leads to an absolute word error rate reduction of 18.9% compared to the baseline.

Channels	Dev		
	S02	S09	Overall
Baseline (U_ref + BFIt) (1)	83.4	81.1	82.5
U_ref (4)	76.1	72.8	74.8
U + BFIt (5)	70.8	68.2	69.3
U (20)	66.3	63.3	65.1
U + BFIt, U (25)	65.5	62.3	64.3
U + BFIt, U, U_ref (29)	64.6	62.2	63.6

U is a single array channel, U_ref is a channel from the reference array.

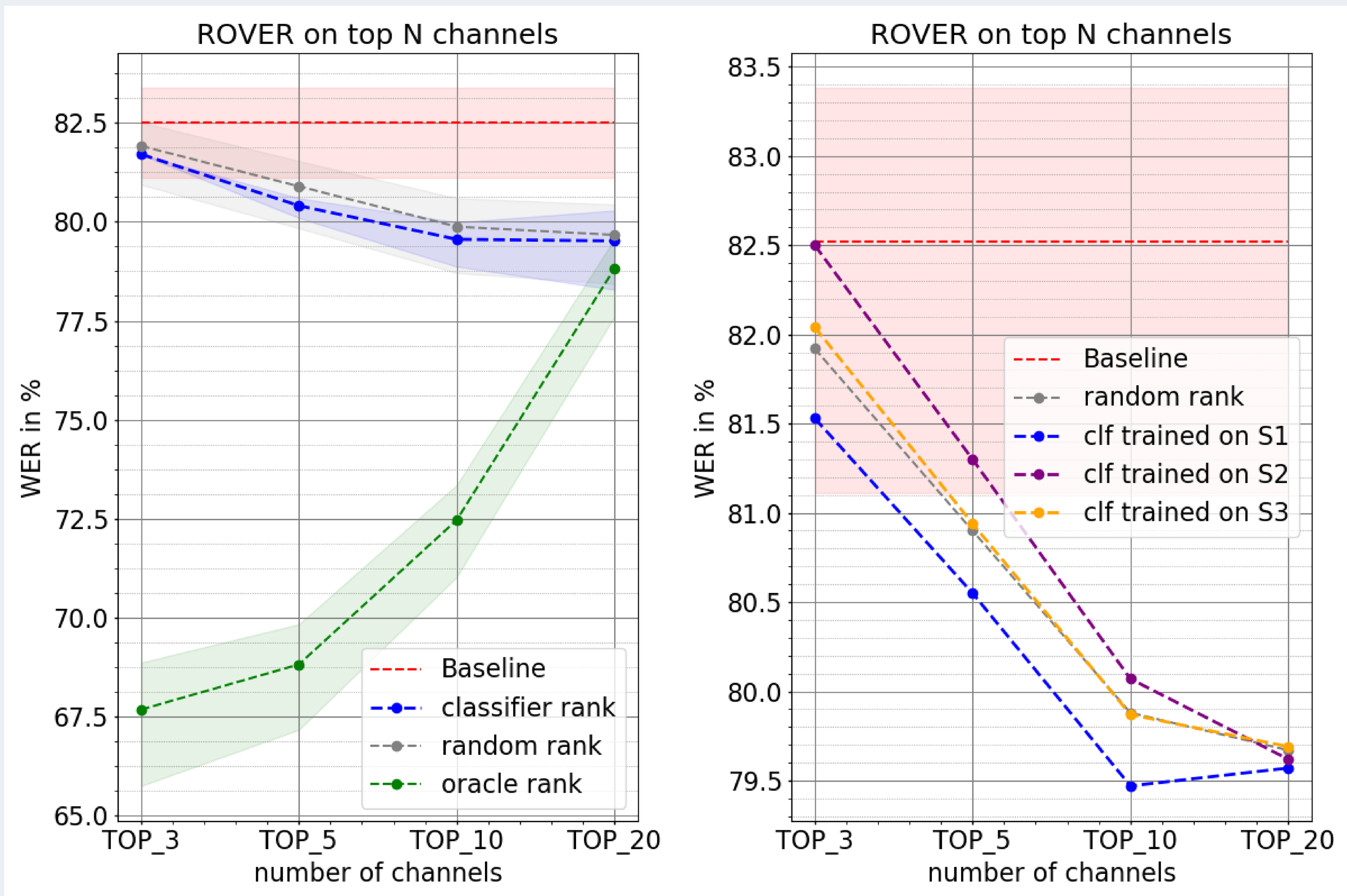
Acoustic model adaptation

- Oracle-selected utterances are used to adapt the baseline DNN-based acoustic model
- Transfer learning: single epoch, very low learning rate for all layers, last layer with higher learning-rate

Adaptation set	Dev		
	S02	S09	Overall
S02 (supervised)	62.6	84.5	70.9
S09 (supervised)	86.9	56.5	75.3
S02 (oracle WER$\leq$60)	83.1	84.7	83.7
S09 (oracle WER$\leq$60)	86.8	80.8	84.5

Channel Selection

- A multi-label DNN using filter bank features is employed to predict if channel is oracle or not.
- Network architecture: 3 hidden layers (LSTM layer + two fully connected layers), sigmoid activation in the output layer used for channel ranking.
- Training: Using different subsets of the CHiME-5 training data with binary cross-entropy as the loss function.
- ROVER results using the N best classified channels.



Transparency colored regions states performance deviation among the two development sessions. Classifier trained on 4 sessions (S1), 6 sessions (S2) and 10 sessions (S3).

Results on Development Set

- Results for the best system. WER (%) per session and location together with the overall WER on the development set.

Track	Session	Kitchen	Dining	Living	Overall
Single	S02	88.7	80.8	78.4	81.5
	S09	81.1	81.1	77.4	
Multiple	S02	83.6	79.5	77.3	79.6
	S09	78.4	78.8	79.5	

Conclusion

- According to the oracle results channel selection seems promising.
- Results using energy or spatial information for channel selection are not convincing.
- Ongoing investigation on model adaptation [1] and an enhancement stage based on Beamforming and other denoising techniques [2] [3].

References

[1] M. Matassoni, M. Ravanelli, S. Jalalvand, A. Brutti, and D. Falavigna, "The FBK system for the CHiME-4 challenge," in 4th International Workshop on Speech Processing in Everyday Environments, San Francisco, US, September 2016.

[2] L. Pfeifenberger, M. Zöhrer, and F. Pernkopf, "Dnn-based speech mask estimation for eigenvector beamforming," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 2017, pp. 66–70.

[3] T. Schrank, L. Pfeifenberger, and M. Z Deep Beamforming and Data Augmentation for Robust Speech Recognition: Results of the 4th CHiME Challenge.