On DNN Posterior Probability Combination in Multi-Stream Speech Recognition for Reverberant Environments

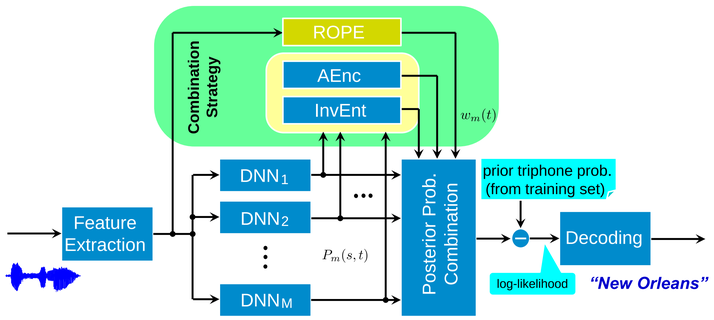
Feifei Xiong[1,2,3], Stefan Goetze[2,3], Bernd T. Meyer[4]

[1]Medizinische Physik, Carl von Ossietzky Universität Oldenburg, Germany [2]Fraunhofer Institute for Digital Media Technology, Project Group Hearing, Speech and Audio Technology, Oldenburg, Germany
[3]Cluster of Excellence Hearing4all, Oldenburg, Germany [4]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA

ABSTRACT

- A multi-stream framework with DNN classifiers → to improve ASR performance in various reverberant environments
- Combination strategy is the crucial issue
  - neural network posterior probability combination
  - frame-wise stream merging
  - higher weights to more reliable streams
- To determine the stream-specific weights
  - Inverse entropy (InvEnt) [1]
  - Autoencoders (AEnc) [2]
  - ROom Parameter Estimator (ROPE) model

SYSTEM



$$P(s,t) = \sum_{m=1}^{M} w_m(t)\, P_m(s,t)$$

- $P_m(s,t)$ → the *m*th DNN posterior probability matrix at the HMM state *s* and time frame *t*
- $w_m(t)$ → the combination weight for *m*th DNN stream at time frame *t*
  - stream weighting: $\sum_{m=1}^{M} w_m(t) = 1$
  - winner-takes-all: $w_m(t) = \begin{cases} 1 & m = \arg\max w_m(t), \forall m \\ 0 & \text{else} \end{cases}$
  - utterance-based mode: temporal averaging

- ASR setup based on the REVERB Challenge DNN/HMM framework in Kaldi [4]
  - 7861 utterances for training, 1088 for each test set
  - No additive noises
  - FBANK for ASR DNN
  - Context-dependent triphone states → posteriors

METHODS

Inverse Entropy (InvEnt)

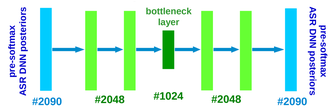
- Distribution analysis of the posteriors (*unsupervised*)
- Weight ←→ inverse entropy value

$$w_m(t) = \frac{1/e_{\mathrm{InvEnt},m}(t)}{\sum_{m=1}^{M} 1/e_{\mathrm{InvEnt},m}(t)} \qquad e_{\mathrm{InvEnt}}(t) = -\sum_{s=1}^{S} P(s,t)\log_2(P(s,t))$$

Autoencoders (AEnc)

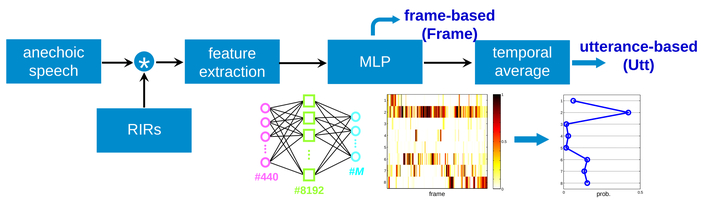
- Train an autoencoder to learn the posterior distribution (*supervised*)
- Weight ←→ inverse reconstruction error square value

$$w_m(t) = \frac{1/\|e_{\mathrm{AEnc},m}(s,t)\|^2}{\sum_{m=1}^{M} 1/\|e_{\mathrm{AEnc},m}(s,t)\|^2}$$



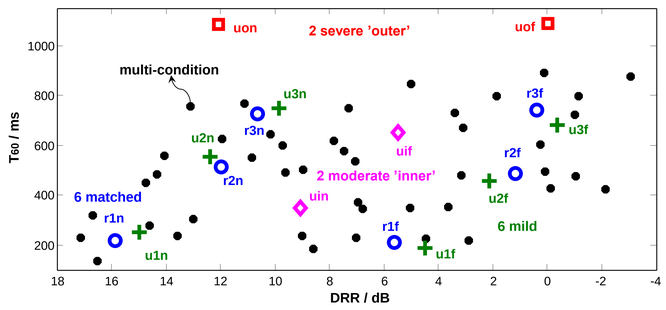
ROom Parameter Estimator (ROPE)

- ROPE output posteriors correlate with the relative performances between all DNN streams (*supervised*) [3]
- Weight ←→ MLP posterior probabilities



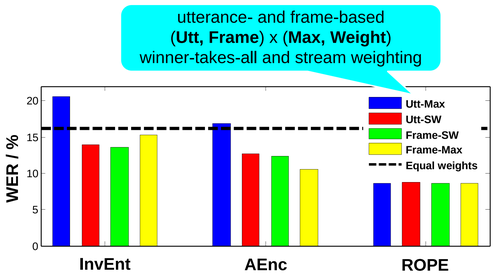
EXPERIMENTAL SETUP

- ***M* = 8** expert streams: clean- and multi-condition, 6 specific conditions (r1n, r1f, r2n, r2f, r3n, r3f)
- Set A: clean test and the chosen 6 matched conditions
- Set B: 6 mild, 2 moderate, 2 severe mismatched conditions



RESULTS & DISCUSSION

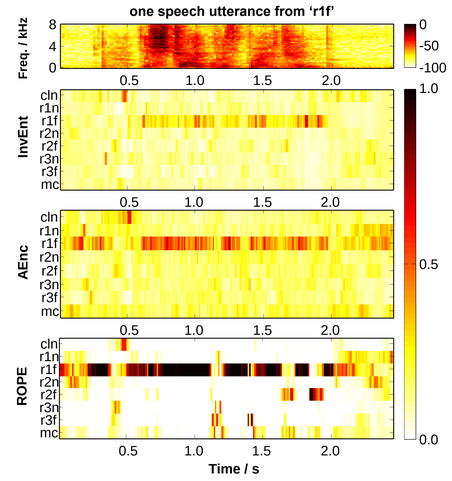
- Equal weights → mediocre results
- In general, AEnc better than InvEnt and 'Frame' better than 'Utt'
- ROPE → lowest and consistent WERs
- InvEnt & AEnc: independent frame processing → isolated noisy frames, severely affecting Utt-Max
- ROPE: some temporal smoothing due to the spliced input features (11 frames)



	Train \ Test	Set A	Set B: (mild)	(moderate)	(severe)	Avg.
Single-stream	Clean-cond.	31.43	32.29	29.62	57.23	34.55
	Multi-cond.	8.40	**8.60**	**7.92**	**13.62**	9.03
Multi-stream	Equal weights	13.61	14.76	12.43	33.18	16.18
	InvEnt Frame-Weight	11.52	12.31	10.06	28.15	13.58
	AEnc Frame-Max	9.27	9.79	8.97	18.46	10.50
	ROPE Frame-Max	**7.04**	8.72	8.06	14.41	**8.62**

- ROPE > AEnc > InvEnt > Equal weights
- Multi-cond. here is a very strong baseline
  - generalization with 44 RIRs
- Multi-stream system with ROPE still provides comparable results to multi-cond.
  - outperforms multi-cond. in matched test Set A
  - multi-cond. advantageous for unseen highly reverberant conditions
- More investigation into multi-stream system is required!!!

- One example to inspect the obtained frame-wise combination weights
- ROPE provides consistently higher and far less noisy estimates than InvEnt and AEnc



CONCLUSIONS

- **ROPE**: new method to determine stream weights for combination of DNN posterior probabilities in a multi-stream DNN/HMM framework
- Outperforming **InvEnt** (46% relative) and **AEnc** (29% relative) in known and unknown reverberant scenarios for stream weighting or selection
- Stable results independently of (weighting or winner-takes-all) & (frame-wise vs. utterance-level) → real-time ASR

[1] H. Misra, H. Bourlard, and V. Tyagi, "New Entropy based Combination Rules in HMM/ANN Multi-Stream ASR," in IEEE ICASSP, Apr. 2003, pp. 741–744.

[2] S. H. Mallidi, T. Ogawa, K. Veselý, P. S. Nidadavolu, and H. Hermansky, "Autoencoder based Multi-Stream Combination for Noise Robust Speech Recognition," in Proc. Interspeech, Dresden, Germany, Sep. 2015, pp. 3551–3555.

[3] F. Xiong, S. Goetze, and B. T. Meyer, "Estimating Room Acoustic Parameters for Speech Recognizer Adaptation and Combination in Reverberant Environments," in IEEE ICASSP, Florence, Italy, May 2014, pp. 5559–5563.

[4] F. Xiong, B. T. Meyer, N. Moritz, R. Rehr, J. Anemüller, T. Gerkmann, S. Doclo, and S. Goetze, "Front-End Technologies for Robust ASR in Reverberant Environments - Spectral Enhancement-based Dereverberation and Auditory Modulation Filterbank Features," EURASIP Journal on Advances in Signal Processing, vol. 2015:70, pp. 1–18, 2015.

This work was funded by the DFG (Cluster of Excellence 1077/1 "Hearing4All" and the SFB/TRR 31 "The Active Auditory System") as well as the BMBF and the EC (project KNOTS, grant no. AAL-2013-6-144.