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

Feifei Xiong1,2,3, Stefan Goetz2,3, Bernd T. Meyer4
1Medizinische Physik, Carl von Ossietzky Universität Oldenburg, Germany 2Fraunhofer Institute for Digital Media Technology, Project Group Hearing, Speech and Audio Technology, Oldenburg, Germany 3Cluster of Excellence Hearing4all, Oldenburg, Germany 4Center 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] ROPE Parameter Estimator (ROPE) model

METHODS

Inverse Entropy (InvEnt)

Distribution analysis of the posteriors (unsupervised)
• Weight \( \propto \text{inverse entropy value} \)
• \( W_{\text{InvEnt}}(t) = \frac{1}{T} \sum_{t=1}^{T} P_{\text{train}}(s,t) \log(P_{\text{train}}(s,t)) \)

Autoencoders (AEnc)

• Train an autoencoder to learn the posterior distribution (supervised)
• Weight \( \propto \text{inverse reconstruction error square value} \)
• \( W_{\text{AEnc}}(t) = \frac{1}{T} \sum_{t=1}^{T} \left( \frac{1}{2} \text{AEnc}(c(s,t)) - \frac{1}{2} \text{AEnc}(\hat{c}(s,t)) \right)^2 \)

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)

EXPERIMENTAL SETUP

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

CONCLUSIONS

• 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


This work was funded by the DFG (Cluster of Excellence 1077 “Hearing4all” and the SFB/TRR 13 “The Active Auditory System”). As well as the BMBF and the EC project KNOTS, grant no. AAL-2013-6-144.