Situation Informed End-to-End ASR for Noisy Environments

Suyoun Kim*, Siddharth Dalmia*, Florian Metze
Carnegie Mellon University, School of Computer Science

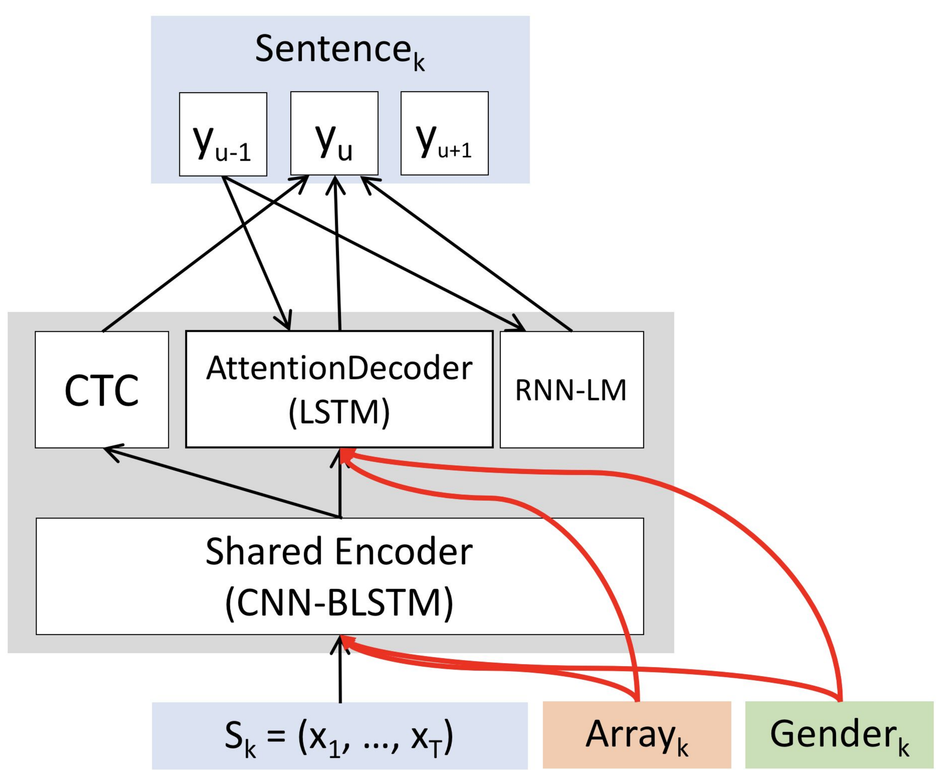
Introduction

- Problem and Challenge:
  - Distant microphone conversational speech recognition in everyday home environments[1]
  - Insufficient amount of training data
- Our goal:
  - Improve the performance of the **end-to-end ASR** modeling with using **context information**, without using any speech enhancement technique or data augmentation or data cleanup or lexicon information.

Our proposed methods

1. **Acoustic Environment Modeling**
   - Different arrays have different acoustic conditions both in terms of type of noise, and also the topic that is generally discussed.
   - Males and Females often carry-on different conversations and differ significantly in acoustic properties.

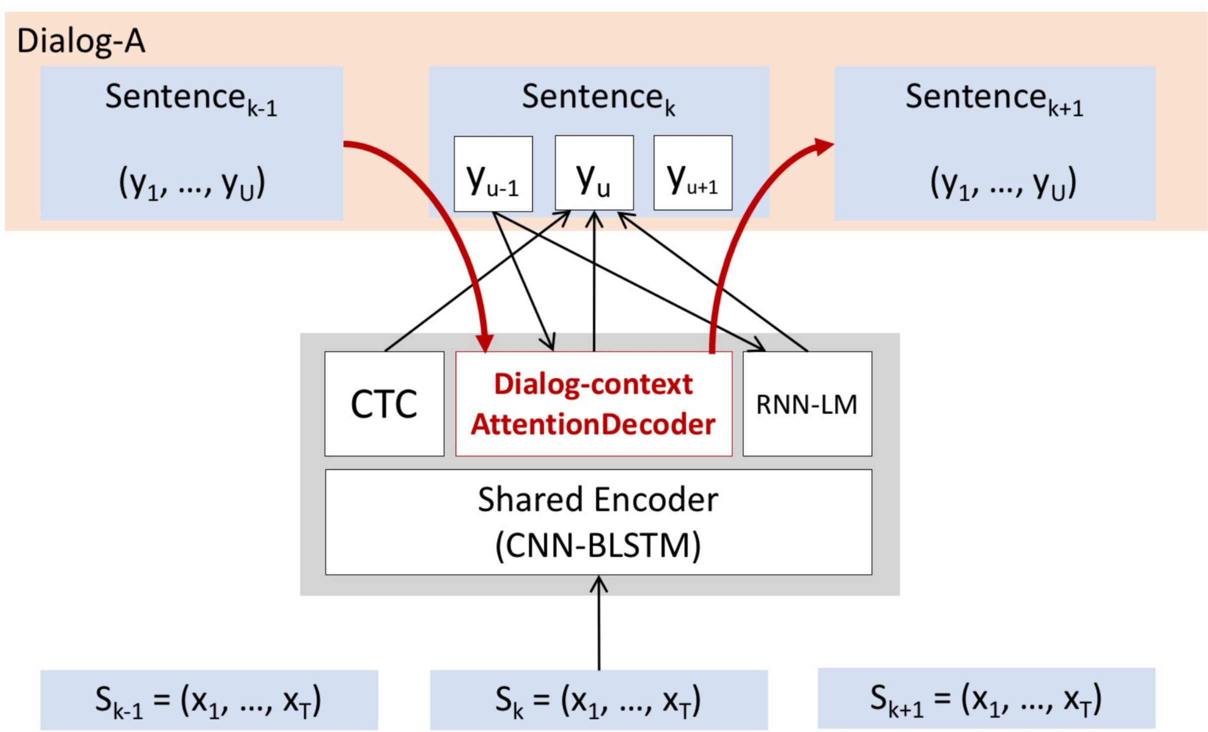
Figure 1: Acoustic Environment Modeling



- To modulate these variations in the network's internal representation we extend the end-to-end speech recognition models[2,3] to explicitly use location of microphone array, and gender of speaker.
- We generate one-hot gender, array, location ID and add them to decoder network as well as encoder network

2. **Dialog Context Modeling[4]**
   - We extend the end-to-end speech recognition models[2,3] to explicitly use dialog-context information, e.g. higher-level knowledge that spans across sentences.

Figure 2: Dialog Context Modeling

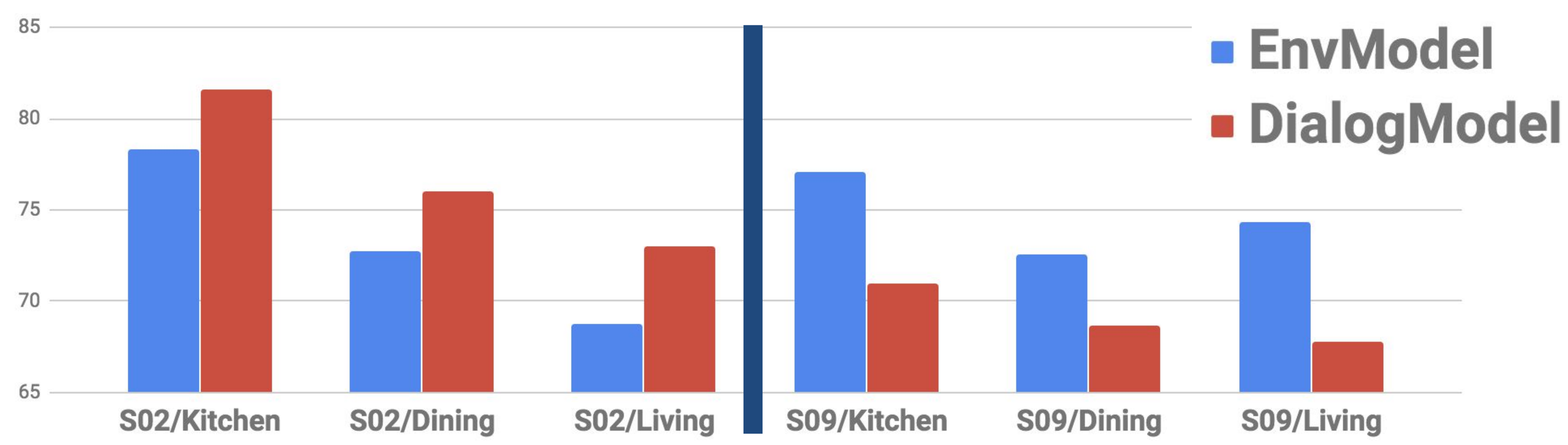


Our proposed methods

- We serialize dataset based on their onset times and sessions
- Our decoder is additionally conditioning on the dialog-context vector which represents the preceding sentence.
- One way to represent the dialog-context vector: the final hidden decoder states of preceding sentence.

Results and Analysis

Figure 3: WER Comparison of EnvModel and DialogModel



- Interestingly the improvement of DialogModel models seemed to be dependent on the session/dialog. DialogModel worked better for S09, however, EnvModel worked better for S02. Shows that the 2 models are contrastive and model combination can help!
- We got considerable improvements over the end-to-end baseline and almost matched the LF-MMI TDNN model numbers without doing any new feature engineering, or data cleaning, or data augmentation.

Table 1: WER Comparison

Models	Dev	Eval
End-to-End baseline	94.7	N/A
LF-MMI TDNN	**81.14**	73.27
Our End-to-End model	82.08	**71.82**

- Experiment setup
  - Dataset: provided beamformed data (40hrs x 6 arrays), without any extra feature enhancement
  - 83d input features, 45 char-level distinct outputs

Conclusions and Future work

- Our end-to-end ASR model obtained 12.6% absolute WER improvement and outperformed LF-MMI TDNN.
- Our model can be easily combined with other speech enhancement techniques, such as multi-array processing enhancement, or single-array enhancement via close-talk data, and expect further improvement.
- With better enhancement techniques and model combinations of our 2 contrastive systems we hope to close the gap between the best model and our e2e model.

References:

1. J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth CHiME Speech Separation and Recognition Challenge: Dataset, task and baselines," in Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018), Hyderabad, India, Sep. 2018.
2. S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," IEEE Journal of Selected Topics in Signal Processing, vol. 11,no. 8, pp. 1240–1253, 2017.
3. S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N.E.Y. Soplin, J. Heymann, M. Wiesner, N. Chen et al.,"Espnet: End-to-end speech processing toolkit," arXiv preprint arXiv:1804.00015, 2018.
4. S. Kim, and F. Metze. "Dialog-context aware end-to-end speech recognition." arXiv preprint arXiv:1808.02171 (2018).