Situation Informed End-to-End ASR for CHiME-5 Challenge

Suyoun Kim[1*], *Siddharth Dalmia*[2*], *Florian Metze*[2]

[1]Electrical & Computer Engineering
[2]Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213; U.S.A.
{suyoung1|sdalmia|fmetze} @andrew.cmu.edu

Abstract

This paper describes an end-to-end speech recognition system for the 5th CHiME challenge that addresses continuous conversation in everyday environments using distributed microphone arrays. The main contribution of our system is the investigation of an effective adaptation method within the end-to-end system based on speaker gender information, microphone array information, and conversational history information for better generalization. Without using any speech enhancement technique, or data augmentation, or data cleaning up, or lexicon information, our proposed system produces better ASR performance than the baseline system (LF-MMI TDNN) which requires the lexicon information and a complicated conventional modeling process (i.e. HMM/GMM, triphone-based acoustic modeling, fMLLR, SAT, i-vector, Data cleaning up, etc). Our final ASR system achieves an absolute word error rate reduction of 12.6% on development set in comparison to the end-to-end baseline system, and an absolute word error rate reduction of 1.5% on evaluation set in comparison to conventional baseline system (LF-MMI TDNN) in a single-array track.

1. Setup for CHiME-5 Challenge

In this section, we describe the dataset of the single-array track for the 5th CHiME challenge [1] and our end-to-end baseline system based on the ESPnet toolkit [2, 3, 4].

1.1. Dataset

We investigated the performance of the proposed models on the CHiME-5 task which has a 40 hours training set from the close-talk microphone and 6 distant microphone arrays. For the 6 distant microphone arrays, we used the beamformed data by using BeamformIt toolkit [5] which are provided from the challenge organizers. In total, we used 40x7 hours of the beamformed data for training (560k utterances), and 4.5 hours of the beamformed data from a reference distant microphone array for hyper-parameter tuning. Evaluation was carried out on the evaluation set, which has 5 hours of the beamformed data from a reference distant microphone array.

We sampled all audio data at 16kHz, and extracted 80-dimensional log-mel filterbank coefficients with 3-dimensional pitch features, from 25 ms frames with a 10ms frame shift. We used 83-dimensional feature vectors to input to the network in total. We used 45 distinct labels including 26 characters, several special characters, and start-o,f-speech/end-of-speech, and blank tokens. Note that no pronunciation lexicon was used in any of the experiments.

[*] Equal Contributions

1.2. End-to-End ASR

We use joint CTC/Attention end-to-end speech recognition architecture with the Chainer [6] deep learning library and ESPnet toolkit [2, 3, 4]. We used a CNN-BLSTMP encoder followed the baseline system, except we down-sampled by 3 along with the time frequency axis within CNN. The CNN layers are followed by a 6-layer BiLSTM with 320 cells. We used a location-based attention mechanism, where 10 centered convolution filters of width 100 were used to extract the convolutional features. The decoder network was a one-layer LSTM with 300 cells. We also built a character-level RNN-LM on the CHiME-5 training transcription which was optimized using the AdaDelta algorithm. We used $\lambda = 0.1$ for joint CTC/Attention training objective:

$$\mathcal{L} = \lambda \mathcal{L}_{\text{CTC}} + (1 - \lambda)\mathcal{L}_{\text{Attention}}. \tag{1}$$

For decoding of the models, we used joint decoder which combines the output label scores from the AttentionDecoder, CTC, and character-level RNN-LM by using shallow fusion [7]:

$$\begin{aligned} \boldsymbol{y}* = \text{argmax}\{ & \log p_{att}(\boldsymbol{y}|\boldsymbol{x}) \\ & + \alpha \log p_{ctc}(\boldsymbol{y}|\boldsymbol{x}) \\ & + \beta \log p_{rnnlm}(\boldsymbol{y})\} \end{aligned} \tag{2}$$

The scaling factor of CTC, and character-level RNN-LM scores were $\alpha = 0.1$, and $\beta = 0.1$, respectively. We used a beam search algorithm similar to [8] with the beam size 20 to reduce the computation cost. We adjusted the score by adding a length penalty, since the model has a small bias for shorter utterances. The final score is normalized with a length penalty 0.1.

2. Contributions

In this section, we describe our proposed models which is built on top of the end-to-end baseline (in Section 1.2).

2.1. Stabilized End-to-End Baseline

When we build the end-to-end baseline system that the challenge organizer is provided, we found that there exist many numerical stability issues during the calculation of CTC loss since there are many cases that the length of the subsampled input frames is shorter than the output character sequence. The original baseline dropped the whole minibatch when any example in that batch led to a NaN while calculating the CTC loss. This led to many good examples being skipped. In order to minimize to drop the minibatches, we calculate the CTC loss for each example in the minibatch and mask out the NaN loss rather than

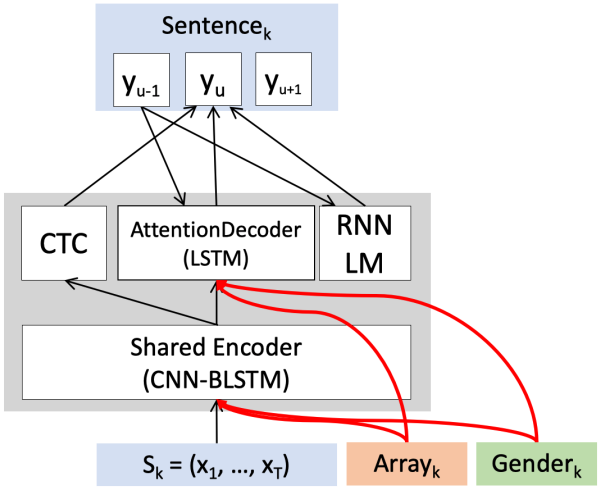

Figure 1: *The architecture of our end-to-end speech recognition model with speaker gender and microphone array information. The red curved line represents the speaker gender and microphone array information flow to the encoder network and the decoder network.*

dropping the entire minibatch. In addition, we changed the rate of subsampling rate from 4 to 3, to minimize the examples that the input length is shorter than the output length. From this modification, we improved the end-to-end baseline from 94.7% to 90.2%.

2.2. Acoustic Environment Modeling

Different arrays are recorded in 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. Several meta-information, such as speaker gender identity, microphone array identity, location of microphone array are available in the evaluation set as well as the training/development sets. Such additional information can be exploited to help adapt our model to different acoustic conditions in terms of type of speaker, and the environment where the conversation occurs. This use of external knowledge has been shown helpful in a lot of speech tasks, like speaker adaptation [9] using i-vectors [10], visual adaptation [11], environmental noise adaptation [12] etc.

To modulate these variations in the networks internal representation we extend the end-to-end speech recognition models to explicitly use the information of the speaker gender identity (`gender`), the microphone array identity (`array`), the location of microphone array (`location`). We first create a one-hot indicator vector for each three identities, `gender`, `array`, and `location`. We then append these vectors to the original Mel-filterbank features as an auxiliary input to the model. Additionally, this auxiliary input is also forwarded to the decoder network to adapt our model further. A visualization of our model can be seen in Figure 1.

2.3. Conversational-Context Modeling

The 5th CHiME Challenge considers the problem of conversational speech recognition in a dinner party scenario with four speakers. The training dataset has 16 conversations and each conversation is around 2-hour long. The conversational-context, dynamic contextual flow across multiple sentences, provides important information that can improve speech recognition, especially in the case of such long conversations. However, existing speech recognition systems are typically built at the sentence level and does not employ conversational-context information. In this work, we propose to use a conversational-context aware speech recognition model [13], which explicitly uses context information beyond sentence-level information, in an end-to-end fashion. Our conversational-context model captures a history of sentence-level contexts, so that the whole system can be trained with conversational-context information in an end-to-end manner.

The core idea of our approach is the integration of conversational-context into an attention-based decoder network. Figure 2 shows the fundamental principle of the *DialogAttentionDecoder* subnetwork, an extension of the *AttentionDecoder*, which can be found in standard end-to-end models. Let we have a dataset consists of N-number of conversations, $D = \{d_1, \cdots, d_N\}$ and each conversation $d_i = (s_1, \cdots, s_K)$ has K utterances. k-th utterance s_k is represented as a sequence of U-length output characters ($\boldsymbol{y}$) and T-length input acoustic features ($\boldsymbol{x}$). Given the high-level representation ($\boldsymbol{h}$) of input acoustic features ($\boldsymbol{x}$) generated from *Encoder* subnetwork, both the standard *AttentionDecoder* and our proposed *DialogAttentionDecoder* generates the probability distribution over characters (y_u), conditioned on ($\boldsymbol{h}$), and all the characters seen previously ($y_{1:u-1}$). Our proposed *DialogAttentionDecoder* additionally conditioning on conversational-context vector (c_k), which represents the information of the preceding utterance in the same conversation as:

$$\boldsymbol{h} = \mathit{Encoder}(\boldsymbol{x}) \tag{3}$$

$$y_u \sim \begin{cases} \texttt{standard decoder network:} \\ \quad \mathit{AttentionDecoder}(\boldsymbol{h}, y_{1:u-1}) \\ \texttt{proposed decoder network:} \\ \quad \mathit{DialogAttentionDecoder}(\boldsymbol{h}, y_{1:u-1}, c_k) \end{cases} \tag{4}$$

In order to learn conversational-context during training, we serialize the utterances based on the session identity that they are part of, and then their onset times, as is normally done during decoding. We shuffle data at the session level and create the minibatches across multiple sessions. We bootstrap the training with the model trained on the shuffled data. We explore two

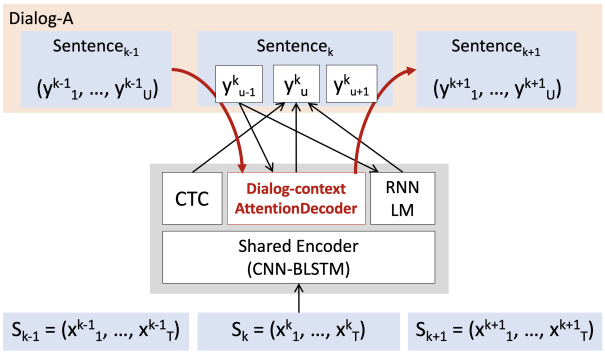

Figure 2: *The architecture of our end-to-end speech recognition model with conversational history information. The red curved line represents the conversational context information flow within the same dialog.*

methods to generate the context vector (c_k) to represent the preceding sentence: method (a) *last-hidden-state* and method (b) *all-outputs* method. In method (a), the last decoder state of the previous sentence represents the context vector, c_k. The context vector c_k is propagated to the initial decoder state of current sentence. In method (b), every output information of preceding sentence is integrated with additional attention mechanism and represents the context vector, c_k. This context vector is then propagated to every decoder state for each output time step of current sentence.

3. Experimental evaluation

Figure 3 shows the WER (%) results of our improved end-to-end baseline and two different proposed models tested on the development set. The `Impr.Baseline` is our improved end-to-end baseline by fixing the CTC loss calculation described in 2.1. The `EnvModel` is our speaker gender identity and microphone array identity informed model described in 2.2, and the `DialogModel` is our conversational context informed model described in 2.3. In overall, we obtained WER improvements using both the proposed models, `EnvModel` and `DialogModel`, over the `Impr.Baseline`. The interesting observation is that the result of `DialogModel` seems to be dependent on the session identity. The `EnvModel` shows similar performance across the session, however, the `DialogModel` seems to be more effective on session 09 (`S09`). One possible reason is that there exists more coherent conversational flow in session 09. From this result, we expect the model combination can improve further, although we chose the `EnvModel` as our final system for the challenge submission.

Table 1 summarizes the WER (%) results for our final system (`EnvModel`) tested on the development set for each session and room [1]. Note that the numbers reported in this table uses the kaldi scoring script mentioned in `http://spandh.dcs.shef.ac.uk/chime_challenge/submission.html`. Similar to the baseline results [1], the performance in the kitchen condition is the poorest probably due to the kitchen background noises and greater degree of speaker movement that occurs in this location.

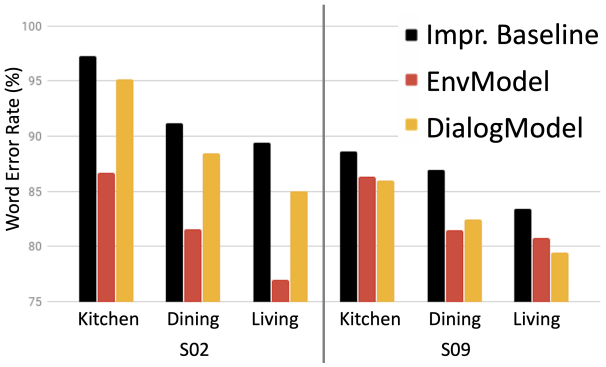

Figure 3: *The WER (%) comparison of our EnvModel with the speaker gender and microphone array information and DialogModel with the conversational-context information.*

[1]Due to inconsistencies in the scoring script, the poster presented at the CHiME workshop in Hyderabad on September 9, 2018, showed a different, incorrect, word error rate. This has been fixed in this version, and the conclusions have been updated accordingly.

Table 1: *Results for our systems. WER (%) per session and location together with the overall WER.*

Track	Session		Kitchen	Dining	Living	Total
Single	dev	S02	86.7	81.6	77.0	82.1
		S09	86.3	81.5	80.8	
	eval	S01	76.9	65.3	83.8	71.8
		S21	74.7	66.1	69.1	

Table 2 shows the comparison of WER (%) results of our final end-to-end system, the end-to-end baseline system, and the conventional baseline system (LF-MMI TDNN) in the single-array track. Our proposed end-to-end system achieved 12.6% absolute WER improvement on development set over the baseline end-to-end system. Note that we could not compare with the result on the evaluation set of baseline end-to-end system since the number was not available. The most noticeable result is that our model outperformed the conventional baseline system (LF-MMI TDNN [14]) on the evaluation set, without using any speech enhancement technique, or data augmentation, or data cleaning, or lexicon information.

Table 2: *The WER (%) comparison of our final end-to-end system, the baseline end-to-end system, and the conventional baseline system (LF-MMI TDNN).*

Models	Dev	Eval
End-to-End baseline [1]	94.7	N/A
LF-MMI TDNN [1]	**81.1**	73.3
Our End-to-End model	82.1	**71.8**

4. Conclusions

In this paper, we introduced an end-to-end speech recogntition system for CHiME-5 challenge. By explicitly using the speaker gender information, the microphone array information, and the conversational history, our proposed model achieved an absolute word error rate reduction of 12.6% on development set in comparison to the end-to-end baseline system. The most noticeable result is that our end-to-end ASR system outperformed the baseline system (LF-MMI TDNN) which requires the lexicon information, the complicated conventional modeling process (i.e. HMM/GMM, triphone-acoustic modeling, fMLLR, SAT, i-vector, Data cleaning up, etc), without using any speech enhancement technique, or data augmentation, or data cleaning up, or lexicon information. In addition, our proposed method can be easily combined with other speech enhancement techniques, such as multi-array processing or single-source enhancement via close-talk microphone data, and we expect further improvement.

5. Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research. The authors would like to thank other members of the CMU speech team for sharing their insights.

6. 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. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4835–4839.

[3] 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.

[4] 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.

[5] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[6] P. Networks, "Chainer," in *"http://chainer.org/"*.

[7] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, "Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm," *arXiv preprint arXiv:1706.02737*, 2017.

[8] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[9] A. Senior and I. Lopez-Moreno, "Improving dnn speaker independence with i-vector inputs," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 225–229.

[10] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[11] A. Gupta, Y. Miao, L. Neves, and F. Metze, "Visual features for context-aware speech recognition," *arXiv preprint arXiv:1712.00489*, 2017.

[12] S. Kim, B. Raj, and I. Lane, "Environmental noise embeddings for robust speech recognition," *arXiv preprint arXiv:1601.02553*, 2016.

[13] S. Kim and F. Metze, "Dialog-context aware end-to-end speech recognition," *arXiv preprint arXiv:1808.02171*, 2018.

[14] D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence trained neural networks for asr based on lattice free mmi (author's manuscript)," The Johns Hopkins University Baltimore United States, Tech. Rep., 2016.