Neural Incremental Speech Recognition Through Attention Transfer

Sashi Novitasari[1] Andros Tjandra[1] Sakriani Sakti[1,2] Satoshi Nakamura[1,2]
[1]Nara Institute of Science and Technology, Japan
[2]RIKEN AIP, Japan
{sashi.novitasari.si3, andros.tjandra.ai6, ssakti, s-nakamura}@.is.naist.jp

1 Introduction

One of the challenges that have to be confronted to achieve a simultaneous speech translation system is the incremental ASR (ISR) development. Hidden Markov model (HMM) ASR [10, 5] performs a low-delay recognition, but it cannot do an end-to-end modeling. The *state-of-the-art* end-to-end ASR [8, 4] with an attention-based sequence-to-sequence (seq2seq) deep learning framework has a long delay due the necessity to attend a complete input. Neural transducer (NT) framework [7] previously proposed for the seq2seq incremental recognition by segment-based recognition. It is done by learning the segment-level alignments from an external system such as forced-alignment by HMM ASR or from the alignment approximation based on the training states.

The existing ISR framework has a more complicated mechanism than the standard ASR because of the incremental step learning and its preparation. In this work, we constructed incremental ASR (ISR) for a low-latency recognition by exploiting an attention-based non-incremental ASR framework. Here we refer our approach as attention-transfer ISR (AT-ISR). The non-incremental ASR is treated as a teacher to teach the ISR through attention transfer. Our experiment shows that our ISR CER with a delay 0.54 sec is only 2% behind the non-incremental model with a delay more than 6 sec.

2 Seq2Seq ASR Overview

The non-incremental character-level seq2seq ASR [3, 2] consists of encoder, decoder, and attention modules to directly models the conditional probability $P(\mathbf{Y}|\mathbf{X})$, where $\mathbf{X}$ is a speech feature frames sequence (length S) and $\mathbf{Y}$ is a characters sequence (length T). The encoder transforms $\mathbf{X}$ into hidden representative information $\mathbf{h}^e$. The decoder predicts the target sequence probability p_{y_t}, given the previous output $\mathbf{Y}_{<t}$, the current context information c_t, and the current decoder hidden state h_t^d. The context c_t is produced by attention module [1] at time t based on encoder and decoder hidden states. The context c_t calculation produces alignment scores between the encoder and decoder states, which can be represented as a matrix. The AT-ISR is constructed by utilizing these alignment scores.

3 Proposed Seq2seq ISR

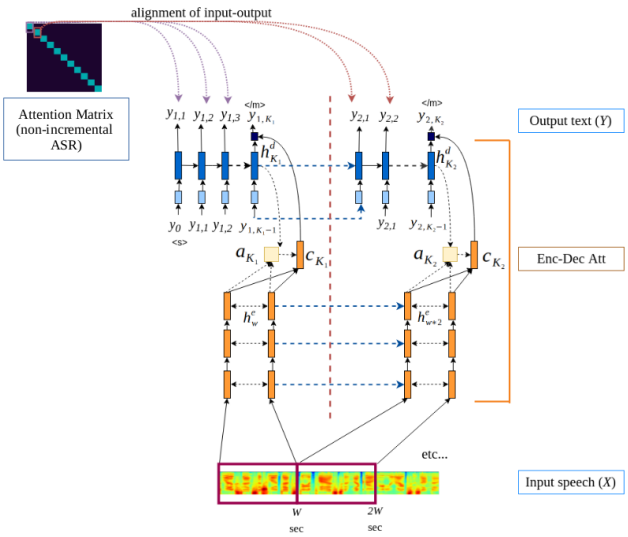

Figure 1: Training method of proposed AT-ISR.

AT-ISR, with the same structure as the non-

incremental ASR, performs several recognition steps incrementally to recognize an utterance, as shown in Figure 1. For each recognition step n, the mechanisms in AT-ISR are the following:

- **Encoder**: Encode **Xn**, a segment from **X** with the length W where $W < S$.
- **Decoder**: Attend the encoder states based on **Xn** and decode for **Yn**, a segment of **Y** with the length Kn where $K_n \leq K < T$. Decoding in step n stop if an end-of-segment $</m>$ token is predicted or the length reaches K tokens. This token is learned during the training phase by attaching $</m>$ at the end of **Yn**.

The recognition in step $n+1$ will be done by keeping the model states and shifting the input window W frames. The AT-ISR delay equals to the input segment size.

AT-ISR is trained via attention knowledge transfer from the non-incremental ASR as the teacher. The segment-level alignments are generated once through teacher-forcing text generation by the teacher. Each output token is aligned to an encoder state, which corresponds to input frames, with the highest attention score. By using the segment-level alignments, AT-ISR is taught to encode a speech segment to decode the target output segment.

4 Experiment

4.1 Setting

We utilized LJ Speech [6] and Wall Street Journal (WSJ) *si284* [9] datasets for our experiments. Our features consist of the 80-Mel spectrogram with window size 50 ms and shift length 12.5 ms.

Our non-incremental ASR and ISR have an identical structure. The encoder, which consists of an FFN layer and three bidirectional LSTM layers, applies state downsampling of eight input frames into one state. In our basic delay, an output token is aligned to eight frames or one frame-block (0.14 sec). The decoder consists of an embedding layer and an LSTM layer.

4.2 Result

The experiment result is shown in Table 1. In this experiment, we utilized the basic delay as the ISR delay to see the performance in the delay as short as possible. The topline is the teacher ASR and the baseline is incremental recognition by using the teacher, which was not trained through attention-transfer for an incremental task.

Table 1: Speech recognition performance. (m=main block; la=look-ahead block; 1 block=8 frames)

Model	**Delay (sec)**	**CER**
LJ Speech		
Teacher ASR	6.54 (avg)	2.78
Baseline ISR	0.14	80.34
AT-ISR (input/step: 1 m)	0.14	23.04
AT-ISR (input/step: 1 m + 4 la)	0.54	4.45
WSJ-si284		
Teacher ASR	7.88 (avg)	6.80
AT-ISR (input/step: 1 m+ 4 la)	0.54	9.06

AT-ISR resulted in a close performance to the teacher, especially when the input includes look-ahead blocks. The experiment with LJ Speech shows that a longer look-ahead block resulted in a better performance. The look-ahead block, which is the frame-block next to the main input block, provides contextual information to complete the information in the main input. AT-ISR cannot perform well without it because of the information limitation in the main input block. The experiment with WSJ or multi-speaker dataset also shows a similar result. For both datasets, the CER difference between the teacher, with a delay 6 sec, and AT-ISR, with a delay 0.54 sec by including look-ahead input blocks, is around 2%. This implies that AT-ISR is able to perform well with a short delay by learning the knowledge from a non-incremental ASR.

5 Conclusion

We constructed ISR that performs a low-delay recognition by transferring the attention knowledge from a non-incremental attention-based ASR that has an identical structure. By delaying the recognition 0.54 sec, the character-level AT-ISR able to

produce the output with the overall quality 2% behind the non-incremental ASR with delay more than 6 sec.

6 Acknowledgement

Part of this work was supported by JSPS KAKENHI Grant Numbers JP17H06101 and JP17K00237.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. Attention-based models for speech recognition. In *Proceedings of the Advances in neural information processing systems (NIPS)*, pp. 577–585, Montreal, Canada, 2015.

[3] William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4960–4964, Shanghai, China, 2016.

[4] Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Ekaterina Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, and Michiel Bacchiani. State-of-the-art speech recognition with sequence-to-sequence models. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018.

[5] David Huggins-Daines, Mohit Kumar, Arthur Chan, Alan W Black, Mosur Ravishankar, and Alex I. Rudnicky. POCKETSPHINX: A free, real-time continuous speech recognition system. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 185–188, Toulouse, France, 2006.

[6] Keith Ito. The LJ speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.

[7] Navdeep Jaitly, Quoc V. Le, Oriol Vinyals, Ilya Sutskever, David Sussillo, and Samy Bengio. An online sequence-to-sequence model using partial conditioning. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pp. 5067–5075, Barcelona, Spain, 2016.

[8] Suyoun Kim, Takaaki Hori, and Shinji Watanabe. Joint CTC-attention based end-to-end speech recognition using multitask learning. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, 2017.

[9] Douglas B. Paul and Janet M. Baker. The design for the Wall Street Journal-based CSR corpus. In *Proceedings of the Workshop on Speech and Natural Language (HLT)*, pp. 357–362, 1992.

[10] Laurence Gillick Robert Roth Paul Bamberg, Yen-lu Chow and Dean Sturtevant. The Dragon continuous speech recognition system: A real-time implementation. In *Proceedings of the Workshop on Speech and Natural Language*, Pennsylvania, USA, 1990.