MULTI-ENCODER SEQUENTIAL ATTENTION NETWORK FOR CONTEXT-AWARE SPEECH RECOGNITION IN JAPANESE DIALOG CONVERSATION

Nobuya Tachimori[1], *Sakriani Sakti*[2,3], *Satoshi Nakamura*[2,3]

[1]Onkyo Corporation, Japan
[2]Nara Institute of Science and Technology, Japan
[3]RIKEN, Center for Advanced Intelligence Project (AIP), Japan

ABSTRACT

Japanese automatic speech recognition (ASR) in conversation is considered challenging and highly ambiguous because the subjects are usually omitted and the number of homonyms is quite large. Since the speaker's utterance alone may not provide enough information, the context of the previous utterance from the dialog partner might help. A limited number of studies have addressed the idea of incorporating dialog-context information in end-to-end ASR tasks. However, these have mainly focused on the English language. Furthermore, although such models exploit dialog context history, the decoder handles the current input and the context simultaneously, ignoring the sequence of a conversation in the process. In this study, we propose a Japanese ASR that considers the dialog partner's conversation context in a sequential manner. Specifically, we use a multi-encoder sequential attention network and investigate several possible architectures. The experimental results reveal that context information helps to improve recognition performance over standard ASR and systems that process the context at the same time.

Index Terms: end-to-end speech recognition, hierarchical attention, context-aware, multi-encoder

1. INTRODUCTION

With the significant improvement of automatic speech recognition (ASR) technology, voice-based services and devices have become mainstream in recent years. For example, using our voice to directly control our TV or smart speakers is becoming more prevalent. Leveraging ASR in contact centers for customer service is also gaining attention. Some services use ASR with the intention of replacing human operators with autonomous contact centers run by machine. Others maintain traditional call centers and use ASR to transcribe the conversation between customers and operators. Those transcriptions are then used to improve operational efficiency in various ways, such as checking call content for NG words and simplifying the creation of FAQs by sharing text. In Japan, multiple companies have also started to provide ASR services to contact centers, and this is becoming accepted as a solution that can improve business efficiency.

However, the performance of ASR in Japanese contact centers tends to result in lower recognition rates compared to general speech recognition. One reason for this is that conversation in Japanese often omits grammatical subjects, and the number of homonyms is quite large, making the meaning highly ambiguous. Since the speaker's utterance alone may not provide enough information, it is essential to have the context information of the previous utterances from the dialog partner in addition to the current utterance. Since there may be many ways to incorporate additional information in ASR, an optimum mechanism to incorporate the additional context of the dialog partners' previous utterances needs to be investigated.

Incorporating context has been studied since work began on statistical speech recognition based on the Hidden Markov Model (HMM) framework [1,2]. Most HMM-based or hybrid DNN-HMM-based acoustic models were constructed with context-dependent long acoustic units, such as triphones, pentaphones, or longer [3–6]. Furthermore, in addition to acoustic modeling, most language models also used n-grams as the biasing context [7, 8]. However, the context used here is still within a word or within a single utterance.

Recently, after the resurgence of deep learning, interest has also surfaced in the possibility of applying a long conversational context within the neural-based ASR [9–13]. Masumura et al. [11] attempted to incorporate large context into end-to-end ASR using a hierarchical encoder-decoder model. Han et al. [13], on the other hand, proposed ContextNet, a fully convolutional encoder that incorporates global context information in convolution layers by adding squeeze-and-excitation modules. Moreover, a study by Hori et al. [12] constructed a Transformer-based architecture that accepts multiple consecutive utterances at the same time and predicts an output sequence for the last utterance. Nevertheless, these studies have focused on utilizing the context information of previous long conversations of the same speaker. Only a limited number of research works have addressed incorporating dialog-context information in end-to-end ASR tasks. Furthermore, the existing systems mainly focus on the English language.

This research focuses on transcribing Japanese dialogs and investigating multiple neural-based ASR architectures that recognize target speech and information previously spoken by the dialog partner as additional context to achieve higher recognition accuracy than standard speech recognition without context information. Specifically, our proposed network adopts attention-based end-to-end ASR [14], which uses a multi-encoder mechanism to simultaneously encode speech and context information and convert them into their respective latent features. However, to maintain the sequence of conversation, we propose using a sequential attention network and processing those input features in a sequential manner. The experimental results reveal that context information helps to improve recognition performance over standard ASR as well as versions that process the context at the same time.

2. RELATED WORKS

Among the limited research works that focused on incorporating dialog context, a study by Kim et al. [15, 16] proposed to embedded conversational context information within an end-to-end encoder-decoder ASR framework. Their recent work [17] is likely the only one that has used conversational-context information for processing a long two-speaker dialog conversation. However, although that model exploited dialog context history, the decoder handled the current input and the context simultaneously, ignoring the sequence of conversation in the process. In contrast, our proposed framework processes the dialog partner's conversation context sequentially using a sequential attention network.

Various architectures that incorporate additional information have been explored not only in the ASR research field but also other domains. Especially in machine translation research, many approaches and architectures have been proposed recently to exploit additional context information in machine translation using a multi-encoder and multiple attention mechanisms. Broadly speaking, those methods can be classified into two main types (following a previous definition [18]): the first one is "flat attention combination," where the decoder learns the distribution jointly over all hidden encoder states simultaneously or produces the final output by simply combining two decoder outputs, and the second one is "hierarchical combination," in which the decoder factorizes the distribution over individual encoders one by one. Some studies focused only on a flat attention combination [19, 20] or only on a hierarchical combination [21, 22], while other approaches attempted to explore both methods [18, 23].

However, exploring various architectures to incorporate dialog-context with the ASR framework has not been attempted. In this study, we explore various architectures to consider the dialog partner's conversation context within ASR. In contrast with HAN, which incorporates different levels of inputs (word-level layer, document-level layer, etc.), we focus on sequentially incorporating conversation context within the same level and propose using sequential attention networks. For comparison, we also consider a structure that uses a flat attention combination similar to the one used in a previous work [17].

3. MULTI-ENCODER SEQUENTIAL ATTENTION NETWORK

In a conversational speech with K number of utterances $\mathbf{u} = [u_1, u_2, ..., u_k, ..., u_K]$, given acoustic speech features of u_k $\mathbf{x} = [x_1, x_2, ..., x_t, ..., x_T]$ with length T, attention-based encoder decoder ASR directly models conditional probability $p(\mathbf{y}|\mathbf{x})$ and an output that corresponds character/word sequences $\mathbf{y} = [y_1, y_2, ..., y_n, ..., y_N]$ with length N. The overall structure consists of encoder, decoder, and attention modules. First, the speech encoder transforms the acoustic speech features into hidden vector representation

$$H_{SpEnc} = SpEnc(\mathbf{x}). \tag{1}$$

Then the attention decoder predicts $\mathbf{y}$ based on the hidden representations of the current speech feature and the entire sequence of the previous output

$$\mathbf{y_n} = p(\mathbf{y_n}|\mathbf{y_{<n}}, \mathbf{x}) = AttDec(\mathbf{y_{<n}}, H_{SpEnc}). \tag{2}$$

Now, assuming that our conversational speech is based on a dialog between two-party speakers A and B, within a dialog we have K number of utterances $\mathbf{u} = [{u_1}^A, {u_2}^B, ... , {u_{k-1}}^A, {u_k}^B, ..., {u_{K-1}}^B, {u_K}^B]$. Our aim is to recognize acoustic speech features of $\mathbf{x}$ of utterance ${u_k}^B$ by incorporating the context from the dialog partner of previous utterance ${u_{k-1}}^A$. Here, we investigate several architectures, including the multi-encoder mechanism illustrated in Figure 1.

3.1. Flat Attention Combination

This model (Figure 1(a)), which is similar to that applied previously [17], has two independent encoder-decoder networks: one acts as an ASR, while the other performs text-to-text response generation. Given acoustic speech features of $\mathbf{x}$ of utterance ${u_k}^B$ from speaker B, the encoder-decoder network outputs the corresponding character sequences $\mathbf{z_n^B}$

$$H_{SpEnc} = SpEnc(\mathbf{x^B}). \tag{3}$$

$$\mathbf{z_n^B} = AttDec(\mathbf{y_{<n}^B}, H_{SpEnc}). \tag{4}$$

On the other hand, having the context from speaker A of the previous utterance $\mathbf{u_{k-1}^A}$, the second encoder-decoder network outputs the response of the current utterance, which considers as the dialog context $\mathbf{c_n^A}$

$$H_{CtxEnc} = ContextEnc(\mathbf{u_{k-1}^A}). \tag{5}$$

$$\mathbf{c_n^A} = ContextAttDec(\mathbf{y_{<n}^B}, H_{CtxEnc}). \tag{6}$$

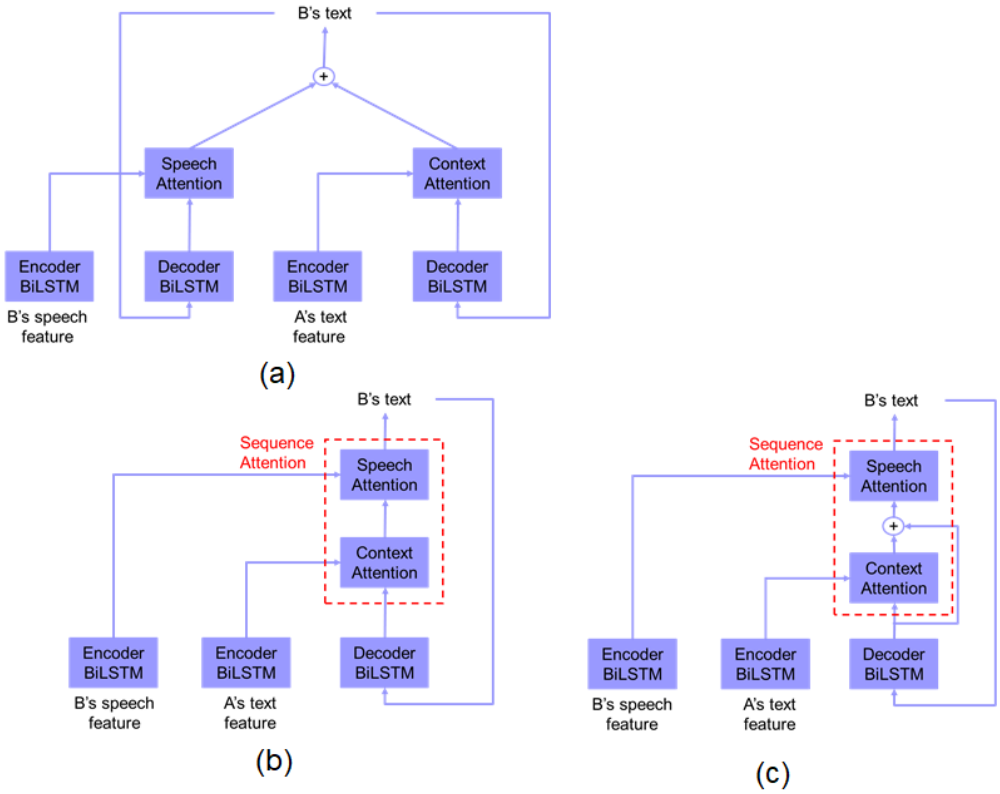

Fig. 1. A model of related works: (a) Flat attention combination ("FlatAttComb"); Proposed Models: (b) Sequential attention combination ("SeqAttComb"); (c) Sequential attention combination and bypass connection ("SeqAttComb+Bypass").

Finally, $\mathbf{z_n^B}$ and $\mathbf{c_n^A}$ are combined to obtain the final text $\mathbf{y_n^B}$.

$$\mathbf{y_n^B} = \mathbf{c_n^A} + \mathbf{z_n^B}. \tag{7}$$

Since the final output is calculated by simply combining two decoder outputs, we call this method the "Flat attention combination" (denoted as "FlatAttComb").

3.2. Proposed Sequential Attention Combination

The first proposed model (Figure 1(b)) uses a multi-encoder sequential attention network. Both speech encoder and contextual encoder transform their respective input, the acoustic speech features of speaker B $\mathbf{x^B}$ and the context from speaker A of the previous utterance $\mathbf{u_{k-1}^A}$, into hidden vector representation

$$H_{SpEnc} = SpEnc(\mathbf{x^B}). \tag{8}$$

$$H_{CtxEnc} = CtxEnc(\mathbf{u_{k-1}^A}). \tag{9}$$

After that, the contextual attention decoder predicts the context vector $\mathbf{c_n^A}$ based on the hidden context representations H_{CtxEnc} and the entire sequence of the previous output

$$\mathbf{c_n^A} = ContextAttDec(\mathbf{y_{<n}^B}, H_{CtxEnc}). \tag{10}$$

Finally, given the context vector $\mathbf{c_n^A}$ and the hidden speech representations H_{SpEnc}, the text transcription is estimated as $\mathbf{y_n^B}$

$$\mathbf{y_n^B} = AttDec(\mathbf{c_n^A}, H_{SpEnc}). \tag{11}$$

We call this method the "Sequential attention combination" (denoted as "SeqAttComb").

The second proposed model (Figure 1(c)) is similar to the second one but with a bypass connection, and the calculation becomes as follows:

$$H_{SpEnc} = SpEnc(\mathbf{x^B}). \tag{12}$$

$$H_{CtxEnc} = CtxEnc(\mathbf{u_{k-1}^A}). \tag{13}$$

$$H_{Dec} = Dec(\hat{\mathbf{y}}_{<\mathbf{n}}^{\mathbf{B}}). \tag{14}$$

$$\mathbf{c_n^A} = ContextAttDec(\mathbf{y_{<n}^B}, H_{CtxEnc}). \tag{15}$$

$$\mathbf{y_n^B} = AttDec(\mathbf{c_n^A}, H_{Dec}, H_{SpEnc}). \tag{16}$$

We call this method the "Sequential attention combination and bypass connection" ("SeqAttComb+Bypass").

Furthermore, in this study, we also investigate two different decoding flows: (1) a non-continuous flow (Figure 2), in which we provide the context information from a correct transcription, and (2) a continuous flow (Figure 3), in which we provide the context information from the model-predicted text.

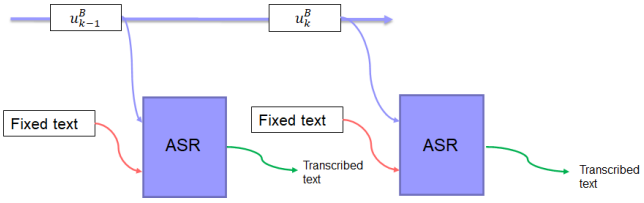

Fig. 2. Non-continuous decoding flow.

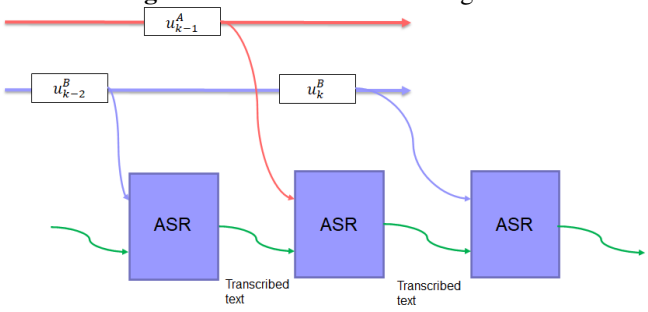

Fig. 3. Continuous decoding flow.

4. EXPERIMENTAL SET-UP AND RESULTS

4.1. Dataset

In this experiment, we utilized two corpora: ATR-APP[1] and ATR-SDB[2], which were used to train and test all ASR models defined in the previous section. From each dataset, 90% of the data are assigned as the training set, 5% as the evaluation set, and 5% as the test set.

Table 1. Specifications of ATR-APP corpus.

Title	Unique speakers	Sentences	Utterance duration
APP3	711	6696	10.1h
APP4	1030	8542	14.2h
APP5	1024	9367	15.7h
APP6	1009	8309	13.4h

Table 2. Specifications of ATR-SDB corpus.

Title	Unique speakers	Sentences	Utterance duration
TRA1	192	4630	6.47h
TRA2	21	6092	10.3h
TRA3	160	5947	8.0h
TRA4	137	6205	10.6h

ATR-APP is a large-scale speech database containing simulated conversations between Japanese people, covering a wide range of regions and ages, making it ideal for speech recognition research on unspecified speakers (detailed specifications in Table 1). The database contains more than 53 hours of simulated conversations between approximately 3,700 speakers in various parts of Japan. The database covers a wide variety of speakers' hometowns and supports a wide range of ages, from 14 to 65 years.

ATR-SDB is a dialog speech database between two speakers of Japanese, where they interact with each other using free speech expressions (detailed specifications in Table 2). Most of the conversations involve a conversation between hotel receptionists and customers over the phone, such as hotel reservations and service inquiries. It contains over 35 hours of simulated conversations with 510 speakers.

4.2. Model Configuration

The sampling rate of all speech utterances was 16 kHz, and we extracted the log Mel-spectrogram (80 dimensions, 50-ms window size, 12.5-ms time steps).

The standard ASR and the proposed ASR models used the same speech encoder-decoder network. The speech encoder consisted of a feed-forward neural network layer (512 units), followed by five bidirectional LSTM layers (512 units each). In each bidirectional LSTM layer, hierarchical subsampling is performed. As a result, the encoder state of the last layer of the encoder represented eight speech frames. Here, we defined eight speech frames as a single speech block.

The context encoder is the character-embedding layer (512 units), which converts 1293 one-shot vectors to 512-dimensional vectors. This is followed by five bidirectional LSTM layers (512 units each).

The composition of attention is the same for both attention to context and attention to audio, and we use the standard MLP-type attention. The decoder consists of a character-embedding layer (512 units), an LSTM layer with an attention mechanism (512 units), and a softmax layer. The loss function of training these models is integrated based on the cross-entropy loss between the correct and predicted symbolic sequences. The batch size was 24 and the number of training steps was 160,000, with the same settings for all models.

5. EXPERIMENTAL RESULTS

The results with a non-continuous decoding flow for the "FlatAttComb" model are 4.1%, 13.2%, and 8.8% for the ATR-APP, ATR-SDB, and both test sets, respectively. Overall, it is not superior to the standard ASR, suggesting that the context information was not being used very effectively. Since the context network and the ASR network are independent of each other in this model, the output seems to tune to the ASR network's output and to ignore the context information.

[1] http://shachi.org/resources/3448
[2] https://www.atr-p.com/products/sdb.html

Table 3. Character Error Rate (CER) of ASR baseline and proposed models on both ATR-APP and ATR-SDB corpora.

Model	CER (%) APP	SDB	APP + SDB
Standard ASR(Baseline)	4.1%	13.0%	8.7%
FlatAttComb	4.1%	13.2%	8.8%
Proposed models			
SeqAttComb	3.8%	12.6%	8.3%
SeqAttComb+Bypass	3.7%	12.4%	8.2%
SeqAttComb+Bypass +Continuous decoding	3.7%	12.3%	8.2%

With a Sequential Attention Network, the proposed model "SeqAttComb" could provide improvements in both corpora (from 4.1% to 3.8% in ATR-APP data and from 13.2% to 12.6% in ATR-SDB data). The results reveal that context information can help speech recognition and improve the recognition rate. The proposed "SeqAttComb+Bypass" model could further improve performance, achieving 3.7%, 12.4%, and 8.2% for the ATR-APP, ATR-SDB, and both tests sets, respectively. This also indicates that the bypass connection could be beneficial to the recognition rate.

For the results with a continuous decoding flow, the system needs to take the ASR error into account. "SeqAttComb+Bypass" model are almost the same as in the non-continuous case. It achieved 3.8%, 12.3%, and 8.2% for the ATR-APP, ATR-SDB, and both test sets, respectively. This indicates that the combination approaches of a sequential attention network is robust to the effects of ASR recognition errors.

Figure 4 shows a typical example when the proposed model performs well in speech recognition results. A's question is asking for B's name and date, and B is answering the question. Standard ASR and "FlatAttComb" recognize it as a word with a similar pronunciation to B's name, Kobayashi. In addition, the sentence after the name is also wrong. On the other hand, "SeqAttComb" correctly recognizes B's name, but misrecognizes some characters in the following sentence. "SeqAttComb+Bypass+Continuous decoding" correctly recognizes both the name and the following sentence.

6. CONCLUSION

This study aimed to improve the speech recognition accuracy of Japanese dialog conversations. Specifically, we utilized a multi-encoder sequential attention network and investigated several possible architectures. The experimental results show that the model using Multi-Encoder Sequence Attention outperforms the standard ASR and that the contextual information from a dialog partner helps speech recognition.

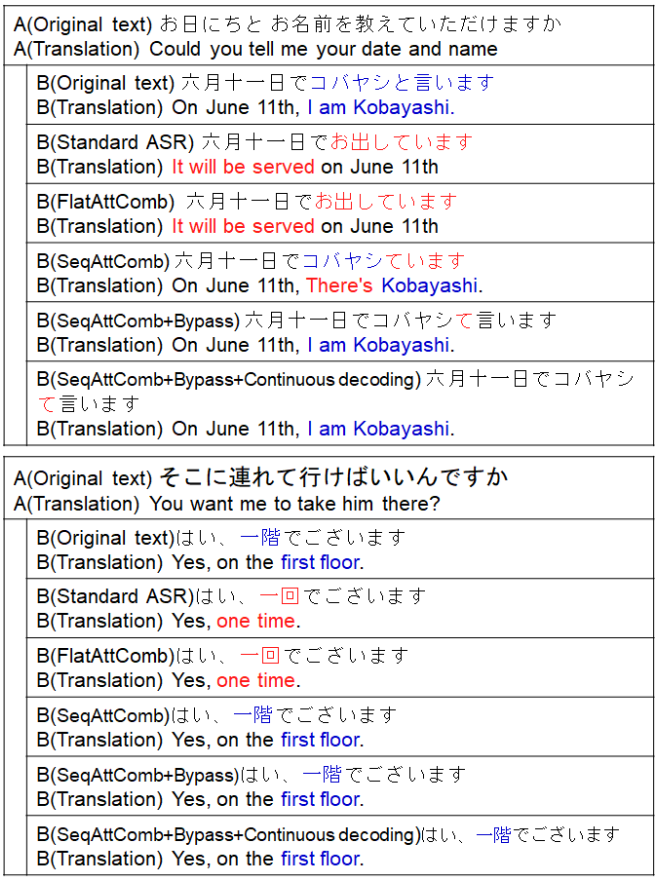

Fig. 4. Comparision of ASR results between ASR baseline with proposed models

7. REFERENCES

[1] F. Jelinek, "Continuous speech recognition by statistical methods," *IEEE*, vol. 64, pp. 532–536, 1976.

[2] J. G. Wilpon, L. R. Rabiner, C. H. Lee, and E. R. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden Markov models," *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. 38, no. 11, pp. 1870–1878, 1990.

[3] K.-F. Lee, "Context-dependent phonetic hidden markov models for speaker independent continuous speech recognition," *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. 38, no. 4, pp. 599–609, 1990.

[4] R. Messina and D. Jouvet, "Context dependent "long units" for speech recognition," in *Proc. INTERSPEECH-ICSLP*, 2004.

[5] S. Sakti, K. Markov, and S. Nakamura, "A hybrid HMM/BN acoustic model utilizing pentaphone-context

dependency," *IEICE Transaction on Information and Systems*, vol. E89-D, no. 3, pp. 954–961, 2006.

[6] G. Dahl, D. Yu, and L. Deng, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. 10, no. 1, pp. 30–42, 2012.

[7] S.-M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. 35, no. 3, pp. 400–401, 1987.

[8] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[9] U. Alon, G. Pundak, and T.-N. Sainath, "Contextual speech recognition with difficult negative training examples," *arXiv preprint arXiv:1810.12170*, 2018.

[10] G. Pundak, T.-N. Sainath, R. Prabhavalkar, A. Kannan, and D. Zhao, "Deep context: end-to-end contextual speech recognition," *arXiv preprint arXiv:1808.02480*, 2018.

[11] R. Masumura, T. Tanaka, T. Moriya, Y. Shinohara, T. Oba, and Y. Aono, "Large context end-to-end automatic speech recognition via extension of hierarchical recurrent encoder-decoder models," in *Proc. ICASSP*, 2019.

[12] T. Hori, N. Moritz, C. Hori, and J. Le Roux, "Transformer-based long-context end-to-end speech recognition," in *Proc. INTERSPEECH*, 2020.

[13] W. Han, Z. Zhang, Y. Zhang, J. Yu, C.-C. Chiu, J. Qin, A. Gulati, R. Pang, and Y. Wu, "Contextnet: Improving convolutional neural networks for automatic speech recognition with global context," in *Proc. INTERSPEECH*, 2020, pp. 3610–3614.

[14] J. Chorowski, D. Bahdanau, D. Serdyuk, K.Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. NIPS*, 2015.

[15] S. Kim and F. Metze, "Dialog-context aware end-to-end speech recognition," in *Proc. SLT*, 2018.

[16] S. Kim, S. Dalmia, and F. Metze, "Gated embeddings in end-toend speech recognition for conversational-context fusion," in *Proc. ACL*, 2019.

[17] ——, "Cross-attention end-to-end asr for two-party conversations," in *Proc. INTERSPEECH*, 2019.

[18] J. Libovický and J. Helcl, "Attention strategies for multi-source sequence-to-sequence learning," in *Proc. ACL (Volume 2: Short Papers)*, 2017, pp. 196–202.

[19] E. Voita, P. Serdyukov, R. Sennrich, and I. Titov, "Context-aware neural machine translation learns anaphora resolution," in *Proc. ACL*, 2018, pp. 1264—-1274.

[20] S. Maruf, A. F. T. Martins, and G. Haffari, "Selective attention for context-aware neural machine translation," in *Proc. NAACL-HLT*, 2019.

[21] J. Zhangy, H. Luany, M. Suny, F. Zhai, J. Xu, M. Zhangx, and Y. Liuyz, "Improving the transformer translation model with document-level context," in *Proc. EMNLP*, 2018, pp. 533–542.

[22] L. Miculicich, D. Ram, N. Pappas, and J. Henderson, "Document-level neural machine translation with hierarchical attention networks," in *Proc. EMNLP*, 2018, pp. 2947—-2954.

[23] B. Li, H. Liu, Z. Wang, Y. Jiang, T. Xiao, J. Zhu, T. Liu, and C. Li, "Does multi-encoder help? a case study on context-aware neural machine translation," in *Proc. ACL*, 2020.