Interspeech 2022
18-22 September 2022, Incheon, Korea

End-to-End Joint Modeling of Conversation History-Dependent and Independent ASR Systems with Multi-History Training

Ryo Masumura, Yoshihiro Yamazaki, Saki Mizuno, Naoki Makishima, Mana Ihori, Mihiro Uchida, Hiroshi Sato, Tomohiro Tanaka, Akihiko Takashima, Satoshi Suzuki, Shota Orihashi, Takafumi Moriya, Nobukatsu Hojo, Atsushi Ando

NTT Corporation, Japan

ryou.masumura.ba@hco.ntt.co.jp

Abstract

This paper proposes end-to-end joint modeling of conversation history-dependent and independent automatic speech recognition (ASR) systems. Conversation histories are available in ASR systems such as meeting transcription applications but not available in those such as voice search applications. So far, these two ASR systems have been individually constructed using different models, but this is inefficient for each application. In fact, conventional conversation history-dependent ASR systems can perform both history-dependent and independent processing. However, their performance is inferior to history-independent ASR systems. This is because the model architecture and its training criterion in the conventional conversation history-dependent ASR systems are specialized in the case where conversational histories are available. To address this problem, our proposed end-to-end joint modeling method uses a crossmodal transformer-based architecture that can flexibly switch to use the conversation histories or not. In addition, we propose multi-history training that simultaneously utilizes a dataset without histories and datasets with various histories to effectively improve both types of ASR processing by introducing unified architecture. Experiments on Japanese ASR tasks demonstrate the effectiveness of the proposed method.

Index Terms: end-to-end speech recognition, conversation-history, crossmodal transformer, multi-history training

1. Introduction

For automatic speech recognition (ASR) based applications, two main ASR systems have been developed. One is a conversation history-independent system that handles isolated speech inputs. The other is a conversation history-dependent system that considers long-range conversation histories as contexts to transcribe individual speech inputs. The former is often used in applications such as voice search systems, and the latter in those such as meeting transcription systems. Conventionally, these two ASR systems have been individually constructed using different models for each application.

In the last few years, both conversation history-dependent and independent ASR systems have been individually implemented with end-to-end ASR (E2E-ASR) modeling as it well performs the overall optimization of the ASR processing. The conversation history-independent E2E-ASR is the most common form of the E2E-ASR [1–6]. The initial studies mainly adopted connectionist temporal classification [1, 2] and recurrent neural network (RNN) encoder-decoders [3, 4]. Recent studies have used the transformer encoder-decoder, which provides much stronger ASR performance [5, 6]. In addition, the conversation history-dependent E2E-ASR, which is also called large-context E2E-ASR, has received increasing attention [7–11]. For the modeling, conventional studies used hierarchical RNN [7,8], hierarchical transformer [9] and context extended transformer [10,11]. These studies reported that the conversation history-dependent E2E-ASR systems outperform the history-independent ones in discourse and meeting ASR tasks. Unfortunately, in previous studies, an appropriate model must be prepared, i.e., either the conversation history-dependent or independent processing, for each application even though E2E modeling is used.

In fact, conventional conversation history-dependent E2E-ASR systems can perform not only history-dependent processing but also the history-independent processing by feeding the model zero padded conversation histories. However, we found that their performance is inferior to history-independent ASR systems (see Section 5). This is because the conventional model architecture and its training criterion are specialized in the cases where conversation histories are given as input. In other words, the conventional model architecture cannot eliminate the inputs from the conversation histories even though isolated speech inputs are given. In addition, the conventional training criterion for the history-dependent systems is not applicable to isolated speech-based training dataset. By addressing these difficulties, systems using a unified model are expected to successfully perform in both history-dependent and independent processes.

In this paper, we propose E2E joint modeling of conversation history-dependent and independent ASR systems. Our key idea is to adopt an architecture that can flexibly switch to use conversation histories or not, without feeding the model zero-padded conversation histories. To this end, the proposed method uses a crossmodal transformer-based architecture. This is motivated by studies of crossmodal representation learning [12, 13] where an encoder network can be utilized for both single-modal and multi-modal processing. The strength of the crossmodal transformer-based architecture is that not only can support both conversation history-dependent and independent ASR processing but can also utilize sharable knowledge of them. In addition, this paper proposes multi-history training that simultaneously utilizes a dataset without histories and datasets with various histories. The training can be regarded as a data augmentation for conversation contexts. This is expected to produce a robust ASR model against both a variety of conversational contexts and none. In experiments on Japanese ASR tasks, we demonstrate that the proposed E2E joint model provides better performance in both history-dependent and independent ASR processing compared with conventional hierarchical transformer [9]. We also demonstrate the effectiveness of the multi-history training.

10.21437/Interspeech.2022-11357

2. Related Work

Recognizing long-form speech: This paper is related to ASR studies that aim to well recognize long-form speech. To this end, there are two main streams: segmentation-driven ASR [7–11] and segmentation-free ASR [14–17]. Our conversation history-dependent ASR is considered to be segmentation-driven ASR, which assumes conversation-level long-form speech, that can be split into utterance-level speech using voice activity detection. While conventional segmentation-driven conversation history-dependent ASR methods are weak against short-form speech, i.e., isolated utterances, our proposed method can robustly recognize both short-form and long-form speech using an unified ASR model.

Speech-text crossmodal encoder architecture: This paper is also related to studies that use a speech-text crossmodal encoder architecture. In ASR tasks, a crossmodal encoder architecture is used for error correction of ASR results [18]. In addition, a speech-text crossmodal architecture is used for the pre-training of spoken language understanding models [19–22]. To the best of our knowledge, this paper is the first study that leverages the speech-text crossmodal encoder architecture to jointly model conversation context-dependent and independent ASR.

3. Preliminaries

This section briefly describes conversation history-independent and dependent E2E-ASR systems. This paper uses an autoregressive model to construct both E2E-ASR systems.

3.1. Conversation history-independent E2E-ASR

Conversation history-independent E2E-ASR is the most common form of E2E-ASR [1–6]. It predicts a generation probability of a transcription $\boldsymbol{W} = \{w_1, \cdots, w_N\}$ from an utterance-level speech $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_M\}$, where w_n is the n-th token in the transcription and $\boldsymbol{x}_m$ is the m-th acoustic feature in the speech. N is the number of tokens in the transcription and M is the number of acoustic features in the speech. In autoregressive modeling, the generation probability of $\boldsymbol{W}$ is defined as

$$P(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{\Theta}^{\mathrm{chi}}) = \prod_{n=1}^{N} P(w_n|w_{1:n-1}, \boldsymbol{X}, \boldsymbol{\Theta}^{\mathrm{chi}}), \quad (1)$$

where $\boldsymbol{\Theta}^{\mathrm{chi}}$ represents the trainable model parameter set and $w_{1:n-1} = \{w_1, \cdots, w_{n-1}\}$. The model parameter set can be trained from a training dataset $\mathcal{D}^{\mathrm{chi}} = \{(\boldsymbol{X}^t, \boldsymbol{W}^t) \mid t \in \{1, \cdots, T\}\}$ where T is the number of utterances in the training dataset. The model parameter set can be optimized by

$$\hat{\boldsymbol{\Theta}}^{\mathrm{chi}} = \underset{\boldsymbol{\Theta}^{\mathrm{chi}}}{\mathrm{argmin}} - \sum_{t=1}^{T} \log P(\boldsymbol{W}^t|\boldsymbol{X}^t, \boldsymbol{\Theta}^{\mathrm{chi}}). \quad (2)$$

For this training, the utterance-level dataset which does not has any conversation histories must be composed.

3.2. Conversation history-dependent E2E-ASR

Conversation history-dependent E2E-ASR is also called large-context E2E-ASR [7–11]. In this E2E-ASR system, we suppose a conversation can be split into utterances. Conversation history-dependent E2E-ASR predicts a generation probability of the t-th utterance-level transcription $\boldsymbol{W}^t = \{w_1^t, \cdots, w_{N^t}^t\}$ from its conversation histories $\boldsymbol{W}^{1:t-1} = \{\boldsymbol{W}^1, \cdots, \boldsymbol{W}^{t-1}\}$ and the t-th utterance-level speech $\boldsymbol{X}^t = \{\boldsymbol{x}_1^t, \cdots, \boldsymbol{x}_{M^t}^t\}$. N^t is the number of tokens in the t-th transcription, and M^t is the number of acoustic features in the t-th speech. In autoregressive modeling, the generation probability of $\boldsymbol{W}^t$ is defined as

$$P(\boldsymbol{W}^t|\boldsymbol{X}^t, \boldsymbol{W}^{1:t-1}, \boldsymbol{\Theta}^{\mathrm{chd}}) = \prod_{n=1}^{N^t} P(w_n^t|w_{1:n-1}^t, \boldsymbol{W}^{1:t-1}, \boldsymbol{X}^t, \boldsymbol{\Theta}^{\mathrm{chd}}), \quad (3)$$

where $\boldsymbol{\Theta}^{\mathrm{chd}}$ is the model parameter set and $w_{1:n-1}^t = \{w_1^t, \cdots, w_{n-1}^t\}$. The model parameter set can be trained from a training dataset $\mathcal{D}^{\mathrm{chd}} = \{(\boldsymbol{X}^{t,c}, \boldsymbol{W}^{t,c}, \boldsymbol{W}^{1:t-1,c}) \mid t \in \{1, \cdots, T^c\}, c \in \{1, \cdots, C\}\}$ where C is the number of conversations in the training dataset and T^c is the number of utterances in the c-th conversation. The model parameter set can be optimized by

$$\hat{\boldsymbol{\Theta}}^{\mathrm{chd}} = \underset{\boldsymbol{\Theta}^{\mathrm{chd}}}{\mathrm{argmin}} - \sum_{c=1}^{C} \sum_{t=1}^{T^c} \log P(\boldsymbol{W}^{t,c}|\boldsymbol{W}^{1:t-1,c}, \boldsymbol{X}^{t,c}, \boldsymbol{\Theta}^{\mathrm{chd}}). \quad (4)$$

For the training, a conversation-level training dataset is needed, and utterance-level training data is not often utilized.

4. Proposed Method

In this paper, we propose joint E2E modeling of conversation history-dependent and independent ASR systems. While previous studies independently model these two systems, the proposed method uses a unified architecture to jointly model both systems. In addition, we propose multi-context training that can produce a robust ASR model against both a variety of conversational contexts and none.

4.1. Joint end-to-end modeling

We define joint end-to-end modeling that enables us to compute both $P(\boldsymbol{W}^t|\boldsymbol{X}^t; \boldsymbol{\Theta})$ and $P(\boldsymbol{W}^t|\boldsymbol{W}^{1:t-1}, \boldsymbol{X}^t; \boldsymbol{\Theta})$ where $\boldsymbol{\Theta}$ is the unified trainable model parameter set. Figure 1 shows how to perform history-dependent and independent E2E-ASR processing using the proposed joint model. Our joint modeling uses a speech encoder, history text encoder, crossmodal encoder, and text decoder. In our method, each type of ASR can be achieved by switching whether to use the history text encoder.

Speech encoder: The speech encoder converts input acoustic features into speech hidden representations. To transcribe the t-th utterance's input speech, the speech encoder converts the acoustic features $\boldsymbol{X}^t$ into the speech hidden representations $\boldsymbol{V}^t$ as

$$\boldsymbol{V}_{\mathrm{co}}^t = \mathrm{ConvolutionPooling}(\boldsymbol{X}^t; \boldsymbol{\theta}_{\mathrm{speech}}^{\mathrm{co}}), \quad (5)$$

$$\boldsymbol{V}_{\mathrm{po}}^t = \mathrm{AddPosition}(\boldsymbol{V}_{\mathrm{co}}^t), \quad (6)$$

$$\boldsymbol{V}_{\mathrm{tr}}^t = \mathrm{TransformerEnc}(\boldsymbol{V}_{\mathrm{po}}^t; \boldsymbol{\theta}_{\mathrm{speech}}^{\mathrm{tr}}), \quad (7)$$

$$\boldsymbol{V}^t = \mathrm{AddSpeechSegment}(\boldsymbol{V}_{\mathrm{tr}}^t; \boldsymbol{\theta}_{\mathrm{speech}}^{\mathrm{se}}), \quad (8)$$

where $\{\boldsymbol{\theta}_{\mathrm{speech}}^{\mathrm{co}}, \boldsymbol{\theta}_{\mathrm{speech}}^{\mathrm{tr}}, \boldsymbol{\theta}_{\mathrm{speech}}^{\mathrm{se}}\} \in \boldsymbol{\Theta}$ is the trainable parameters of the speech encoder. $\mathrm{ConvolutionPooling}()$ is a function composed of convolution layers and pooling layers, $\mathrm{AddPosition}()$ is a function that adds a continuous vector in which position information is embedded, $\mathrm{TransformerEnc}()$ is a function of the transformer encoder blocks that consist of

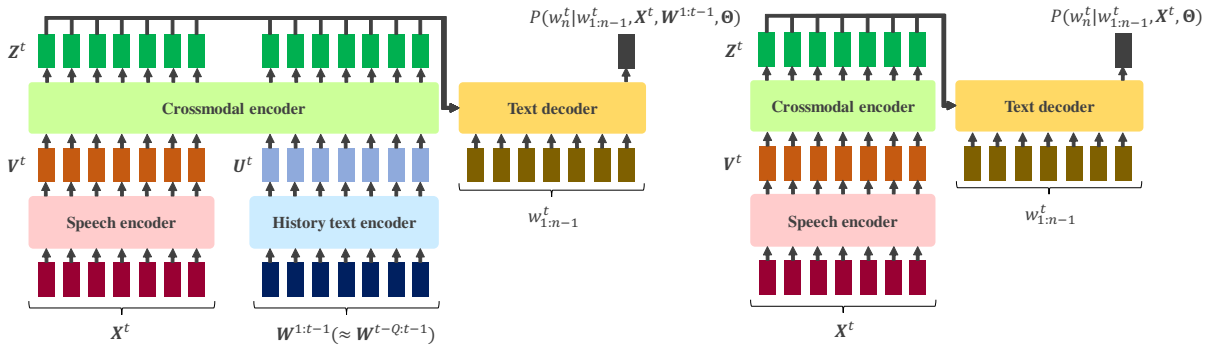

Figure 1: *How to perform history-dependent and independent E2E-ASR processing using our joint E2E modeling.*

multi-head self-attention layers and position-wise feed-forward networks [5], and AddSpeechSegment() is a function that adds a continuous vector in which speech segment information is embedded [23].

History text encoder: The history text encoder embeds conversation contexts into text hidden representations. In general, the conversation contexts are automatic transcriptions of preceding utterances. Since it is impractical to take all conversation contexts into consideration, we limit conversation histories for the t-th utterance as $\boldsymbol{W}^{t-Q:t-1}$, where Q is the number of utterances used for the limited histories. The history text encoder converts $\boldsymbol{W}^{t-Q:t-1}$ into text hidden representations $\boldsymbol{U}^t$ as

$$\begin{aligned}
\boldsymbol{U}_{\mathrm{em}}^t &= \mathrm{Emedding}(\boldsymbol{W}^{t-Q:t-1}; \boldsymbol{\theta}_{\mathrm{history}}^{\mathrm{em}}), && (9)\\
\boldsymbol{U}_{\mathrm{po}}^t &= \mathrm{AddPosition}(\boldsymbol{U}_{\mathrm{em}}^t), && (10)\\
\boldsymbol{U}_{\mathrm{tr}}^t &= \mathrm{TransformerEnc}(\boldsymbol{U}_{\mathrm{po}}^t; \boldsymbol{\theta}_{\mathrm{history}}^{\mathrm{tr}}), && (11)\\
\boldsymbol{U}^t &= \mathrm{AddTextSegment}(\boldsymbol{U}_{\mathrm{tr}}^t; \boldsymbol{\theta}_{\mathrm{history}}^{\mathrm{se}}), && (12)
\end{aligned}$$

where $\{\boldsymbol{\theta}_{\mathrm{history}}^{\mathrm{em}}, \boldsymbol{\theta}_{\mathrm{history}}^{\mathrm{tr}}, \boldsymbol{\theta}_{\mathrm{history}}^{\mathrm{se}}\} \in \boldsymbol{\Theta}$ are the trainable parameters of the history text encoder. Embedding() is a linear transformation function to convert tokens into continuous representations and AddTextSegment() is a function that adds a continuous vector in which text segment information is embedded [23]. Note that we do not consider utterance boundaries of the conversation histories.

Crossmodal encoder: The crossmodal encoder is fed outputs of the speech encoder and the history text encoder. In our method, each type of ASR can be achieved by switching whether to use the history text encoder. Thus, we define hidden representations $\boldsymbol{Z}_0^t$ as

$$\boldsymbol{Z}_0^t = \begin{cases} \mathrm{Concat}(\boldsymbol{V}^t, \boldsymbol{U}^t) & \text{if histories are available,} \\ \boldsymbol{V}^t & \text{otherwise,} \end{cases} \quad (13)$$

where Concat() is a function that concatenates inputs on the axis of time series. Note that the dimensions of $\boldsymbol{V}^t$ and $\boldsymbol{U}^t$ must be same although their sequence length is different. We then convert the hidden representations into $\boldsymbol{Z}^t$ as

$$\boldsymbol{Z}^t = \mathrm{TransformerEnc}(\boldsymbol{Z}_0^t; \boldsymbol{\theta}_{\mathrm{cross}}), \quad (14)$$

where $\boldsymbol{\theta}_{\mathrm{cross}} \in \boldsymbol{\Theta}$ are the trainable parameters of the crossmodal encoder. In this transformer function, crossmodal attention can be performed when histories exist.

Text decoder: The text decoder computes the generation probability of token w_n^t from its preceding tokens within an utterance $w_{1:n-1}^t$ and $\boldsymbol{Z}^t$. It is computed from

$$\begin{aligned}
\boldsymbol{O}_{\mathrm{em}_n}^t &= \mathrm{Embedding}(w_{1:n-1}^t; \boldsymbol{\theta}_{\mathrm{decoder}}^{\mathrm{em}}), && (15)\\
\boldsymbol{O}_{\mathrm{po}_n}^t &= \mathrm{AddPosition}(\boldsymbol{O}_{\mathrm{em}_n}^t), && (16)\\
\bar{\boldsymbol{o}}_n^t &= \mathrm{TransformerDec}(\boldsymbol{O}_{\mathrm{po}_n}^t, \boldsymbol{Z}^t; \boldsymbol{\theta}_{\mathrm{decoder}}^{\mathrm{tr}}), && (17)\\
\boldsymbol{o}_n^t &= \mathrm{Softmax}(\bar{\boldsymbol{o}}_n^t; \boldsymbol{\theta}_{\mathrm{decoder}}^{\mathrm{so}}), && (18)
\end{aligned}$$

where $\{\boldsymbol{\theta}_{\mathrm{decoder}}^{\mathrm{em}}, \boldsymbol{\theta}_{\mathrm{decoder}}^{\mathrm{tr}}, \boldsymbol{\theta}_{\mathrm{decoder}}^{\mathrm{so}}\} \in \boldsymbol{\Theta}$ are the trainable parameters of the text decoder. TransformerDec() is a set of transformer decoder blocks that consist of masked multi-head self-attention layers, multi-head source-target attention layers, and position-wise feed-forward networks. Softmax() is a softmax layer with a linear transformation. The output $\boldsymbol{o}_n^t$ corresponds to $P(w_n^t|w_{1:n-1}^t, \boldsymbol{X}^t, \boldsymbol{\Theta})$ for the histroy-independent ASR model and $P(w_n^t|w_{1:n-1}^t, \boldsymbol{W}^{1:t-1}, \boldsymbol{X}^t, \boldsymbol{\Theta})$ for the history-dependent ASR model.

4.2. Multi-history training

To effectively perform both history-independent and dependent ASR using a unified model parameter set, we introduce multi-history training that jointly uses a dataset without histories and datasets with various histories. In our multi-history training, we suppose $\mathcal{D}^{\mathrm{chd}}$ defined in Section 3.2 is available. We define the following functions:

$$\mathcal{L}_0(\boldsymbol{\Theta}) = -\sum_{c=1}^{C}\sum_{t=1}^{T^c} \log P(\boldsymbol{W}^{t,c}|\boldsymbol{X}^{t,c}, \boldsymbol{\Theta}), \quad (19)$$

$$\mathcal{L}_q(\boldsymbol{\Theta}) = -\sum_{c=1}^{C}\sum_{t=1}^{T^c} \log P(\boldsymbol{W}^{t,c}|\boldsymbol{X}^{t,c}, \boldsymbol{W}^{t-q:t-1,c}, \boldsymbol{\Theta}), \quad (20)$$

where $\mathcal{L}_0$ is the function to evaluate the history-independent modeling and $\mathcal{L}_q(\boldsymbol{\Theta})$ is that to evaluate history-dependent modeling that considers preceding q histories. In this case, the model parameter set is optimized from

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\mathrm{argmin}} \left[\mathcal{L}_0(\boldsymbol{\Theta}) + \sum_{q=1}^{Q} \mathcal{L}_q(\boldsymbol{\Theta}) \right], \quad (21)$$

where Q represents the max number of histories, which is set to same value with that in the history text encoder. By introducing this training, we expect to improve both history-independent and dependent ASR systems compared with Eq. (4).

Table 1: *Experimental results in terms of character error rate (%).*

	Model architecture	Multi-historiy training	Use histories	Test 1	Test 2	Test 3
Baseline	Transformer	-		7.1	5.1	5.3
Conventional	Hierarchical transformer	-		7.8	5.5	5.8
			✓	6.7	4.5	4.8
Proposed	Crossmodal transformer-based joint model	$\mathcal{L}_5$		6.5	4.9	5.4
			✓	6.0	4.5	4.9
Proposed	Crossmodal transformer-based joint model	$\mathcal{L}_0, \mathcal{L}_1, \mathcal{L}_2$		5.9	4.3	4.9
			✓	5.6	4.0	4.6
Proposed	Crossmodal transformer-based joint model	$\mathcal{L}_0, \mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3, \mathcal{L}_4, \mathcal{L}_5$		5.8	4.2	4.8
			✓	**5.5**	**3.8**	**4.4**

5. Experiments

We evaluate the effectiveness of the proposed method on Japanese discourse ASR tasks using the Corpus of Spontaneous Japanese (CSJ) [24]. We divided the CSJ into a training set (512.6 hours), validation set (4.8 hours), and three test sets (1.8 hours, 1.9 hours, and 1.3 hours, respectively). The segmentation of each discourse-level speech into utterances follows that of previous studies [8, 9]. The number of discourses and that of utterances in the training set were about 3.2K and 413.2K, respectively. We used characters as the tokens. The number of characters in the training set was about 13.3M. These datasets were used for evaluating both conversation history-dependent and independent ASR systems.

5.1. Setups

We compared our proposed crossmodal transformer-based joint modeling of history-dependent and independent ASR systems with two systems. One is a transformer-based history-independent ASR system [6], which was trained from the dataset without conversation histories using Eq. (2). Another is a hierarchical transformer-based history-dependent ASR system [8], which was trained from a dataset with conversation histories using Eq. (4). As described in section 1, the hierachical transformer can perform history-independent processing by feeding the model zero padded conversation histories. For the crossmodal transformer-based joint modeling, we examined three setups by changing the training criterion in multi-history training (details are shown in Table 1).

We used 80 log Mel-scale filterbank coefficients appended with delta and acceleration coefficients as acoustic features. The frame shift was 10 ms. For the transformer and hierarchical transformer, the same setups as in a previous study were used [9]. The setups for the proposed crossmodal transformer-based joint mdeling are as follows. In the speech encoder, acoustic features passed two convolution and max pooling layers with a stride of 2, so we down-sampled them to $1/4$ along with the time axis. We stacked eight transformer encoder blocks. For the history text encoder, we stacked four transformer encoder blocks. For the crossmodal encoder, we stacked four transformer encoder blocks. For the text decoder, we stacked six transformer decoder blocks, and the output unit size in a softmax layer, which corresponds to the number of characters in the training set, was set to 3,084. For each transformer block, the dimensions of the output continuous representations were set to 512, the dimensions of the inner outputs in the position-wise feed-forward networks were set to 2,048, and the number of heads in the multi-head attentions was set to 4. In the nonlinear transformational functions, the Swish activation [25] was used.

For the mini-batch training, the mini-batch size was set to 16 and the dropout rate in the transformer blocks was set to 0.1. We used RAdam [26] for optimization. The training steps were stopped on the basis of early stopping using the validation set. We also applied SpecAugment with frequency masking and time masking [27], where the number of frequency masks and time-step masks were set to 2, frequency-masking width was randomly chosen from 0 to 20 frequency bins, and time-masking width was randomly chosen from 0 to 100 frames. We also applied label smoothing [28]. For ASR decoding, we used a beam search algorithm in which the beam size was set to 4.

5.2. Results

Table 1 shows the evaluation results in terms of character error rate (%). 'Use histories" represents history-dependent processing or history-independent processing. First, the hierarchical transformer-based history-dependent processing outperformed transformer-based history-independent processing. This indicates that it is valuable to take conversation histories into consideration to improve ASR performance. Next, the hierarchical transformer-based history-independent processing was inferior to transformer-based history-independent processing. This indicates that the hierarchical transformer is not suitable for history-independent processing because it is specialized in the cases where conversation histories are given as input. In contrast, the proposed crossmodal transformer-based joint model performed well in both history-dependent and independent processing. In fact, the proposed joint model trained with only $\mathcal{L}_5$ used almost the same training criterion as that for the hierarchical transformer. This indicates the proposed architecture is effective to jointly handle both history-dependent and independent processing. The proposed crossmodal transformer-based joint model was improved by jointly using a dataset without histories and datasets with various histories in multi-history training. The best results were attained by the proposed crossmodal transformer-based joint model trained with $\mathcal{L}_0, \mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3, \mathcal{L}_4, \mathcal{L}_5$ in both history-dependent and independent processing. This confirms that multi-history training is an effective way to produce a robust ASR model against both a variety of conversational contexts and none.

6. Conclusions

We proposed E2E joint modeling of conversation history-dependent and independent ASR systems. Key advances of the proposed method is a crossmodal transformer-based architecture, which can flexibly support two different ASR processing, and a multi-history training which can produce a robust ASR model against both a variety of conversational contexts and none. Experimental results showed that the proposed E2E joint model provides superior performance in both history-dependent and independent ASR processing compared with conventional E2E-ASR systems.

7. References

[1] G. Zweig, C. Yu, J. Droppo, and A. Stolcke, "Advances in all-neural speech recognition," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4805–4809, 2017.

[2] K. Audhkhasi, B. Ramabhadran, G. Saon, M. Picheny, and D. Nahamoo, "Direct acoustics-to-word models for English conversational speech recognition," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 959–963, 2017.

[3] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4945–4949, 2015.

[4] L. Lu, X. Zhang, K. Cho, and S. Renals, "A study of the recurrent neural network encoder-decoder for large vocabulary speech recognition," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 3249–3253, 2015.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *In Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 5998–6008, 2017.

[6] L. Dong, S. Xu, and B. Xu, "Speech-Transformer: A no-recurrence sequence-to-sequence model for speech recognition," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5884–5888, 2018.

[7] S. Kim and F. Metze, "Dialog-context aware end-to-end speech recognition," *In Proc. Spoken Language Technology Workshop (SLT)*, pp. 434–440, 2018.

[8] 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. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5661–5665, 2019.

[9] R. Masumura, N. Makishima, M. Ihori, T. Tanaka, A. Takashima, and S. Orihashi, "Hierarchical transformer-based large-context end-to-end asr with large-context knowledge distillation," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5864–5868, 2021.

[10] T. Hori, N. Moritz, C. Hori, and J. L. Roux, "Transformer-based long-context end-to-end speech recognition," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 5011–5015, 2020.

[11] T. Hori, N. Moritz, C. Hori, and J. L. Roux, "Advanced long-context end-to-end speech recognition using context-expanded transformers," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2097–2101, 2021.

[12] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A simple and performant baseline for vision and language," *arxiv:1908.03557*, 2019.

[13] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "SIMVLM: Simple visual language model pretraining with weak supervision," *arXiv:2108.10904*, 2021.

[14] C.-C. Chiu, W. Han, Y. Zhang, R. Pang, S. Kishchenko, P. Nguyen, A. Narayanan, H. Liao, S. Zhang, A. Kannan, R. Prabhavalkar, Z. Chen, T. Sainath, and Y. Wu, "A comparison of end-to-end models for long-form speech recognition," *In Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019.

[15] A. Narayanan, R. Prabhavalkar, C.-C. Chiu, D. Rybach, T. N. Sainath, and T. Strohman, "Recognizing long-form speech using streaming end-to-end models," *In Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 920–927, 2019.

[16] C.-C. Chiu, A. Narayanan, W. Han, R. Prabhavalkar, Y. Zhang, N. Jaitly, R. Pang, T. N. Sainath, P. Nguyen, L. Cao, and Y. Wu, "RNN-T models fail to generalize to out-of-domain audio: Causes and solutions," *In Proc. Spoken Language Technology Workshop (SLT)*, 2021.

[17] Z. Lu, Y. Pan, T. Doutre, L. Cao, R. Prabhavalkar, C. Zhang, and T. Strohman, "Input length matters: An empirical study of RNN-T and MWER training for long-form telephony speech recognition," *arXiv:2110.03841*, 2021.

[18] T. Tanaka, R. Masumura, M. Ihori, A. Takashima, T. A. Takafumi Moriya, S. Orihashi, and N. Makishima, "Cross-modal transformer-based neural correction models for automatic speech recognition," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 4059–4063, 2021.

[19] H. Li, W. Ding, Y. Kang, T. Liu, Z. Wu, and Z. Liu, "CTAL: Pre-training cross-modal transformer for audio-and-language representations," *In Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3966–3977, 2021.

[20] Y. Qian, X. Bian, Y. Shi, N. Kanda, L. Shen, Z. Xiao, and M. Zeng, "Speech-language pre-training for end-to-end spoken language understanding," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 7458–7462, 2021.

[21] Q. Chen, W. Wang, and Q. Zhang, "Pre-training for spoken language understanding with joint textual and phonetic representation learning," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1244–1248, 2021.

[22] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang, Z. Wei, Y. Qian, J. Li, and F. Wei, "SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing," *arXiv:2110.07205*, 2021.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *In Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171—4186, 2019.

[24] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," *In proc. International Conference on Language Resources and Evaluation (LREC)*, pp. 947–952, 2000.

[25] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *arXiv:1710.05941*, 2017.

[26] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," *In Proc. International Conference on Learning Representations (ICLR)*, 2020.

[27] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2613–2617, 2019.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *In Proc. IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.