MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

End-to-end ASR without using morphological analyzer, pronunciation dictionary and language model

Watanabe, S.; Hori, T.; Hayashi, T.; Kim, S.

TR2017-021 March 2017

Abstract

This paper introduces Japanese end-to-end ASR system based on a joint CTC/attention scheme [1], which is an extension of attention-based ASR [2] by using multi-task learning to incorporate the Connectionist Temporal Classification (CTC) objective. Unlike the conventional Japanese ASR systems based on DNN/HMM hybrid [3] or end-to-end systems with Japanese syllable characters (i.e., hiragana or katakana) [4], this method directly predicts a Japanese sentence based on a standard Japanese character set including Kanji, hiragana, and katakana characters, Roman/Greek alphabets, Arabic numbers, and so on. Thus, the method does not use any pronunciation dictionary, which requires hand-crafted work by human. In addition, since it's based on character based recognition, it does not require a morphological analyzer to chunk a character sequence to a word sequence. Finally, attention mechanism itself holds a language-model-like function in the decoder network, unlike a Japanese end-to-end system based on CTC [5]. Therefore, it does not require a separate language model module, which makes system construction and decoding process very simple.

Acoustical Society of Japan Spring Meeting (ASJ)

End-to-end Japanese ASR without using morphological analyzer, pronunciation dictionary and language model*

WATANABE, Shinji (MERL), HORI, Takaaki (MERL)
◎ HAYASHI, Tomoki (Nagoya Univ.), and KIM, Suyoun (CMU)

1 Introduction

This paper introduces Japanese end-to-end ASR system based on a joint CTC/attention scheme [1], which is an extension of attention-based ASR [2] by using multi-task learning to incorporate the Connectionist Temporal Classification (CTC) objective. Unlike the conventional Japanese ASR systems based on DNN/HMM hybrid [3] or end-to-end systems with Japanese syllable characters (i.e., hiragana ”あ”, ”か” or katakana “ ア”, ”カ”) [4], this method directly predicts a Japanese sentence based on a standard Japanese character set including Kanji, hiragana, and katakana characters, Roman/Greek alphabets, Arabic numbers, and so on. Thus, the method does not use any pronunciation dictionary, which requires hand-crafted work by human. In addition, since it's based on character-based recognition, it does not require a morphological analyzer to chunk a character sequence to a word sequence. Finally, attention mechanism itself holds a language-model-like function in the decoder network, unlike a Japanese end-to-end system based on CTC [5]. Therefore, it does not require a separate language model module, which makes system construction and decoding process very simple.

2 Joint CTC-attention (MTL)

The idea of our model is to use a CTC objective function as an auxiliary task to train the attention model encoder within the multi-task learning (MTL) framework. Figure 1 illustrates the overall architecture of our framework, where the encoder network is shared with CTC and attention models. Unlike the attention model, the forward-backward algorithm of CTC can enforce monotonic alignment between speech and label sequences. We therefore expect that our framework is more robust in acquiring appropriate alignments. Another advantage of using CTC as an auxiliary task is that the network training is converged quickly. In our experiments, rather than solely depending on data-driven attention methods to estimate the desired alignments in long sequences, the forward-backward algorithm in CTC helps to speed up the process of estimating the desired alignment without the aid of rough estimates of the alignment which requires manual effort. The proposed objective is a linear combination of attention $\mathcal{L}_{\text{Attention}}$ and CTC $\mathcal{L}_{\text{CTC}}$ based objective functions:

$$\mathcal{L}_{\text{MTL}} = \lambda\mathcal{L}_{\text{CTC}} + (1-\lambda)\mathcal{L}_{\text{Attention}}, \qquad (1)$$

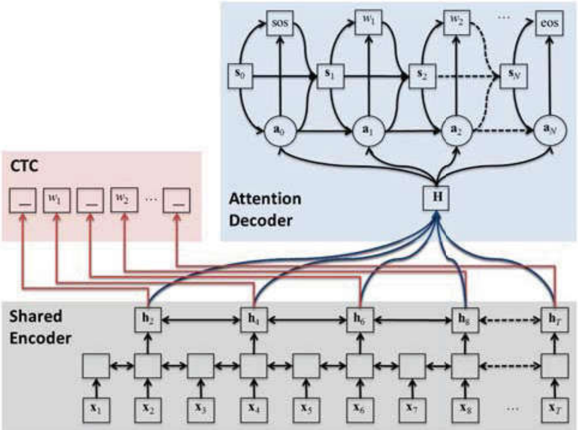

Fig. 1 Joint CTC-attention based end-to-end framework: the shared encoder is trained by both CTC and attention model objectives simultaneously. The shared encoder transforms our input sequence $\boldsymbol{x}$ into high level features $\boldsymbol{h}$, the location-based attention decoder generates the character sequence $\boldsymbol{w}$.

with a tunable parameter $\lambda : 0 \leq \lambda \leq 1$.

3 Experiments

We demonstrate speech recognition experiments by using the Corpus of Spontaneous Japanese (CSJ) [6] based on the CSJ Kaldi recipe developed by [3]. CSJ has totally 581 hours of training data and three types of evaluation data, where each evaluation task consists of 10 lectures. As input features, we used 40 mel-scale filterbank coefficients, with their first and second order temporal derivatives to obtain a total of 120-dimensional feature vector per frame. The encoder was a 4-layer Bidirectional Long Short-Term Memory (BLSTM) with 320 cells in each layer and direction, and linear projection layer is followed by each BLSTM layer. The 2nd and 3rd bottom layers of the encoder read every second hidden state in the network below, reducing the utterance length by the factor of 4. The decoder was a 1-layer LSTM with 320 cells. See [1] for detailed information. The joint CTC/attention ASR was implemented by using the Chainer deep learning toolkit [7].

Table 1 compares the character error rate (CER) for conventional attention and proposed MTL based end-to-end ASR for different amounts of training data. λ in Eq. (1) was set to 0.1. When decoding, we set the minimum and maximum lengths of out-

*形態素解析も辞書も言語モデルも要らない end-to-end 日本語音声認識, 渡部晋治 (MERL)，堀貴明 (MERL)，林知樹 (名古屋大学)，Suyoun Kim (CMU)

Table 1 Character error rate (CER) for conventional attention and proposed MTL based end-to-end ASR for different amounts of training data.

Model (hours)	task1	task2	task3
Attention (147h)	20.1	14.0*	32.7
MTL (147h)	16.9	12.7*	28.9
Attention (236h)	17.2	12.4*	25.4
MTL (236h)	13.9	10.2*	22.2
Attention (581h)	11.5	7.9*	9.0
MTL (581h)	10.9	7.8*	8.3
MTL2 (581h)	**9.5**	**7.0**	**7.8**
GMM-discr. [3] (236h for AM, 581h for LM)	11.2	9.2	12.1
DNN-hybrid [3] (236h for AM, 581h for LM)	9.0	7.2	9.6
CTC-syllable [4] (581h)	9.4	7.3	7.5

put sequences by 0.1 and 0.5 times input sequence lengths, respectively. The insertion penalty was set to 0.1. MTL2 has a larger network (5-layer encoder network), and performs a re-scoring technique by using encoder network outputs during decoding. As a reference, we also list the state-of-the-art CERs of GMM discriminative training and DNN-sMBR hybrid systems obtained from the Kaldi recipe [3] and a system based on syllable-based CTC with MAP decoding [4]. Unlike the proposed method, these methods use linguistic resources including a morphological analyzer, pronunciation dictionary, and language model. The Kaldi recipe systems use academic lectures (236h) for AM training and all training-data transcriptions for LM training. Note that the number of utterances of the attention and MTL methods except for MTL2 in task2 is different from that of the reference results and these are not directly compared.

Table 1 shows that the proposed MTL consistently improved the performance from the attention-based system. Also, by increasing the amount of training data, the proposed MTL greatly improved the performance, and finally the CER of all evaluation tasks scored less than 10.0%.

We also compare the performance of the proposed MTL2 with the conventional state-of-the-art techniques obtained by using linguistic resources. Note that since the amount of training data and experimental configurations of the proposed and reference methods are different, it is difficult to compare the performance listed in the table directly. However, since the CERs of the proposed method are comparable to those of the best reference results, we can state that the proposed method reaches the state-of-the-art performance.

4 Summary and discussion

This paper proposes a Japanese end-to-end ASR system by using joint CTC/attention. This method does not require to use linguistic resources including morphological analyzer, pronunciation dictionary, and language model, which are essential component of building conventional Japanese ASR systems. Nevertheless, the method achieved comparable performance to the state-of-the-art conventional systems for the CSJ task. In addition, the proposed method does not require GMM-HMM construction for initial alignments, DNN pre-training, lattice generation for sequence discriminative training, complex search in decoding (e.g., FST decoder or lexical tree search based decoder). Thus, the method greatly simplifies ASR building process with even smaller amounts of coding. Currently, it took 7-9 days by using a single GPU to train the network with full training data (581h), which is comparable to the whole training time of the conventional state-of-the-art system due to the simplification of building process.

Acknowledgment We would like to thank Dr. Jiro Nishitoba at Retrieva, Inc. and Dr. Shohei Hido at Preferred Networks, Inc. for their valuable discussions and comments on applications of Chainer to (end-to-end) speech recognition. We also would like to thank Dr. Naoyuki Kanda at Hitachi, Ltd. for providing the CER results of his Japanese CTC system.

References

[1] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," *arXiv preprint arXiv:1609.06773*, 2016.

[2] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent NN: First results," *arXiv preprint arXiv:1412.1602*, 2014.

[3] T. Moriya, T. Shinozaki, and S. Watanabe, "Kaldi recipe for Japanese spontaneous speech recognition and its evaluation," in *Autumn Meeting of ASJ*, no. 3-Q-7, 2015.

[4] N. Kanda, X. Lu, and H. Kawai, "Maximum a posteriori based decoding for CTC acoustic models," in *Interspeech 2016*, pp. 1868–1872, 2016.

[5] H. Ito, A. Hagiwara, M. Ichiki, T. Mishima, S. Sato, and A. Kobayashi, "End-to-end neural network modeling for japanese speech recognition," in *The Journal of the Acoustical Society of America*, vol. 140, p. 3116, 2016.

[6] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of japanese," in *LREC*, vol. 2, pp. 947–952, 2000.

[7] S. Tokui, K. Oono, S. Hido, and J. Clayton, "Chainer: a next-generation open source framework for deep learning," in *Proceedings of Workshop on Machine Learning Systems (LearningSys) in NIPS*, 2015.