Dialect-Aware Modeling for End-to-End Japanese Dialect Speech Recognition

Ryo Imaizumi*, Ryo Masumura†, Sayaka Shiota* and Hitoshi Kiya*
* Tokyo Metropolitan University, Tokyo, Japan
E-mail: imaizumi-ryo@ed.tmu.ac.jp
† NTT Media Intelligence Laboratories, NTT Corporation, Kanagawa, Japan

***Abstract*—In this paper, we present a novel model for building end-to-end Japanese-dialect automatic speech recognition (ASR) system. It is known that ASR systems modeling for the standard Japanese language is not suitable for recognizing Japanese dialects, which include accents and vocabulary different from standard Japanese. Therefore, we aim to produce dialect-specific end-to-end ASR systems for Japanese. Since it is difficult to collect a massive amount of speech-to-text paired data for each Japanese dialect, we utilize both dialect data and standard Japanese language data for constructing the dialect-specific end-to-end ASR systems. One primitive approach is a multi-condition modeling that simply merges the dialect data with the standard-language data. However, this simple multi-condition modeling causes inadequate dialect-specific characteristics to be captured because of a mismatch between the dialects and standard language. Thus, to produce reliable dialect-specific end-to-end ASR systems, we propose the dialect-aware modeling that utilizes dialect labels as auxiliary features. The main strength of the proposed method is that it effectively utilizes both dialect and standard-language data while capturing adequate dialect-specific characteristics. In our experiments using a home-made database of Japanese dialects, the proposed dialect-aware modeling outperformed the simple multi-condition modeling and achieved an error reduction of 19.2%.**

I. Introduction

Recently, deep learning has been in development in the field of automatic speech recognition (ASR). As one of the state-of-the-art deep learning-based ASR systems, end-to-end ASR has been proposed [1,2]. End-to-end ASR consists of a single network, and it directly maps acoustic features to characters. Recent studies have proposed many advanced end-to-end ASR models: sequence-to-sequence models with recurrent neural network-based approaches [3, 4] and connectionist temporal classification and attention-based approaches [5,6]. In particular, the performance of transformer-based approaches has been amongst the most powerful [7–11]. However, it is known that the performance of end-to-end ASR depends on the amount of training data [12, 13].

There are many dialects across Japan. Each of them has lots of dialect-specific accents and vocabulary. For examples, even though a dialect includes the same words as the standard Japanese language, the meaning of each can be completely different. In other cases, although the meaning of a word is the same between dialects and the standard language, the word itself is absolutely different. Basically, training data for end-to-end ASR consists of a large amount of standard-language data. Therefore, it is known that ASR systems constructed for a standard language are not suitable for recognizing its dialects [14–16]. To design reliable speech recognizers for each dialect, a large amount of dialect data is required. It is, however, difficult to collect a large amount of speech-to-text paired data for each dialect. One primitive approach to relaxing this problem is a multi-condition modeling that simply merges the dialect data with that of the standard language. However, this simple multi-condition modeling causes inadequate dialect-specific characteristics to be captured because of a mismatch between the dialects and standard language.

To produce reliable dialect-specific end-to-end ASR systems, we propose a dialect-aware modeling that utilizes dialect labels as auxiliary features for a transformer-based end-to-end ASR system. Introducing the labels to the transformer decoder part of the proposed method can mitigate falling into dialect-specific local optima [17]. The main strength of the proposed method is that it effectively utilizes both dialect and standard-language data while capturing adequate dialect-specific characteristics. Hence, the proposed method improves not only the recognition performance for dialects but also for standard language. In our experiments, a home-made database consisting of six Japanese dialects and a standard-Japanese database were used for constructing a transformer-based end-to-end ASR system. From the experimental results, we demonstrate that the proposed dialect-aware modeling outperformed the simple multi-condition modeling and achieved an error reduction of 19.2%.

The paper is organized as follows. Section II describes an end-to-end ASR system based on a transformer encoder-decoder. Then, the proposed method is presented in Section I II. Experimental conditions, the database, and results are presented in Section IV. Section V concludes our work.

II. End-to-End ASR system based on transformer encoder-decoder

This section briefly describes end-to-end ASR using a transformer-based auto-regressive generative model. This model predicts the generation probability of text $\boldsymbol{W} = \{w_1, ..., w_N\}$ given speech $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_M\}$, where w_n is the n-th token in the text and $\boldsymbol{x}_m$ is the m-th acoustic feature in the speech. N is the number of tokens in the text and M is the number of acoustic features in the speech. The auto-regressive generative models define the generation probability of $\boldsymbol{W}$ as

$$P(\boldsymbol{W}|\boldsymbol{X};\boldsymbol{\Theta}) = \prod_{n=1}^{N} P(w_n|\boldsymbol{W}_{1:n-1},\boldsymbol{X};\boldsymbol{\Theta}), \tag{1}$$

where $\boldsymbol{\Theta}$ represents model parameter sets, and $\boldsymbol{W}_{1:n-1} = \{w_1, ..., w_{n-1}\}$.

A. Network structure

In our transformer-based end-to-end ASR system, $P(w_n|\boldsymbol{W}_{1:n-1},\boldsymbol{X};\boldsymbol{\Theta})$ can be computed using a speech encoder and a text decoder, both of which are composed of a couple of transformer blocks. The model parameter sets are split into those for the speech encoder $\theta_{\rm enc}$, and those for the text decoder $\theta_{\rm dec}$.

Speech encoder: The speech encoder converts input acoustic features into hidden representations $\boldsymbol{H}^{(I)}$ using I transformer encoder blocks. The i-th transformer encoder block composes i-th hidden representations $\boldsymbol{H}^{(i)}$ from the lower layer inputs $\boldsymbol{H}^{(i-1)}$ as indicated by

$$\boldsymbol{H}^{(i)} = {\rm TransformerEncoderBlock}(\boldsymbol{H}^{(i-1)};\theta_{\rm enc}), \tag{2}$$

where TransformerEncoderBlock() is a transformer encoder block that consists of a scaled dot-product multi-head self-attention layer and a position-wise feed-forward network. The hidden representation $\boldsymbol{H}^{(0)} = \{\boldsymbol{h}_1^{(0)}, ..., \boldsymbol{h}_{M'}^{(0)}\}$ is produced by

$$\boldsymbol{h}_{m'}^{(0)} = {\rm AddPostionalEncoding}(\boldsymbol{h}_{m'}), \tag{3}$$

where AddPositionalEncoding() is a function that adds a continuous vector in which position information is embedded. $\boldsymbol{H} = \{\boldsymbol{h}_1, ..., \boldsymbol{h}_{M'}\}$ is produced by

$$\boldsymbol{H} = {\rm ConvolutionPooling}(\boldsymbol{x}_1, ..., \boldsymbol{x}_{M'};\theta_{\rm enc}), \tag{4}$$

where ConvolutionPooling() is a function composed of convolution layers and pooling layers. M' is the subsampled sequence length, which depends on the function.

Text decoder: The text decoder computes the generation probability of a token from preceding tokens and the hidden representations of the speech. The predicted probabilities of the n-th token w_n are calculated as

$$P(w_n|\boldsymbol{W}_{1:n-1},\boldsymbol{X};\boldsymbol{\Theta}) = {\rm Softmax}(\boldsymbol{u}_{n-1}^{(J)};\theta_{\rm dec}), \tag{5}$$

where Softmax() is a softmax layer with a linear transformation. The input hidden vector $\boldsymbol{u}_{n-1}^{(J)}$ is computed from J transformer decoder blocks. The j-th transformer decoder block composes j-th hidden representation $\boldsymbol{u}_{n-1}^{(j)}$ from the lower inputs $\boldsymbol{U}_{1:n-1}^{(j-1)} = \{\boldsymbol{u}_1^{(j-1)}, ..., \boldsymbol{u}_{n-1}^{(j-1)}\}$ as

$$\boldsymbol{u}_{n-1}^{(j)} = {\rm TransformerDecoderBlock}(\boldsymbol{U}_{1:n-1}^{(j-1)},\boldsymbol{H}^{(I)};\theta_{\rm dec}), \tag{6}$$

where TransformerDecoderBlock() is a transformer decoder block that consists of a scaled dot-product multi-head self-attention layer, a scale dot product multi-head source-target attention layer, and a position-wise feed-forward network. The hidden representation $\boldsymbol{U}_{1:n-1}^{(0)} = \{\boldsymbol{u}_1^{(0)}, ..., \boldsymbol{u}_{n-1}^{(0)}\}$ is produced by

$$\boldsymbol{u}_{n-1}^{(0)} = {\rm AddPostionalEncoding}(\boldsymbol{w}_{n-1}), \tag{7}$$

$$\boldsymbol{w}_{n-1} = {\rm Embedding}(w_{n-1};\theta_{\rm dec}), \tag{8}$$

where Embedding() is a linear layer that embeds an input token into a continuous vector.

B. Supervised learning

In end-to-end ASR, a model parameter set can be optimized from the utterance-level labeled data (speech-to-text paired data) as

$$\mathcal{D} = \{(\boldsymbol{X}^1,\boldsymbol{W}^1), ..., (\boldsymbol{X}^T,\boldsymbol{W}^T)\}, \tag{9}$$

where T is the number of utterances in the training data set. The objective function based on maximum likelihood estimation is defined as

$$\mathcal{L}_{\rm mle}(\theta_{\rm enc},\theta_{\rm dec}) = -\sum_{t=1}^{T}\sum_{n=1}^{N^t} \log P(w_n^t|\boldsymbol{W}_{1:n-1}^t,\boldsymbol{X}^t;\theta_{\rm enc},\theta_{\rm dec}), \tag{10}$$

where w_n^t is the n-th token for the t-th utterance and $\boldsymbol{W}_{1:n-1}^t = \{w_1^t, ..., w_{n-1}^t\}$. N^t is the number of tokens in the t-th utterance.

C. Challenges in Japanese dialect speech recognition

While the performance of transformer-based end-to-end ASR has been amongst the most powerful, it is known that an end-to-end ASR system constructed for a standard language is not suitable for recognizing dialect. There are many dialects across Japan. Each of them has lots of dialect-specific accents and vocabulary. For examples, “very” in English translates into “totemo” in the standard Japanese language, but in the case of the dialect of the Aomori region, “very” translates into “tange.” In this way, although the meaning of a word is the same between the dialects and standard language, the pronunciation is absolutely different. In other cases, even though a dialect has the same word as the standard language, the meaning of each can be completely different. Thus, a method is required to compensate for the mismatch between dialects and standard language.

Solving such a mismatch problem is similar to the tasks of domain adaptation. Many domain adaptation methods have been proposed to capture unseen information or mismatches from original tasks [18,19]. One domain adaptation approach focuses on the use of auxiliary features. So far, the approaches have improved extremely in terms of performance [20,21]. Dialect speech recognition can be regarded as a similar problem to domain adaptation. However, there has been no study that has tried to recognize Japanese dialect speech using end-to-end networks.

III. Proposed method

This section describes our proposed dialect-aware modeling. The proposed method constructs a transformer-based ASR system using dialect labels as auxiliary features. In the method, an output word sequence can be predicted by using a dialect

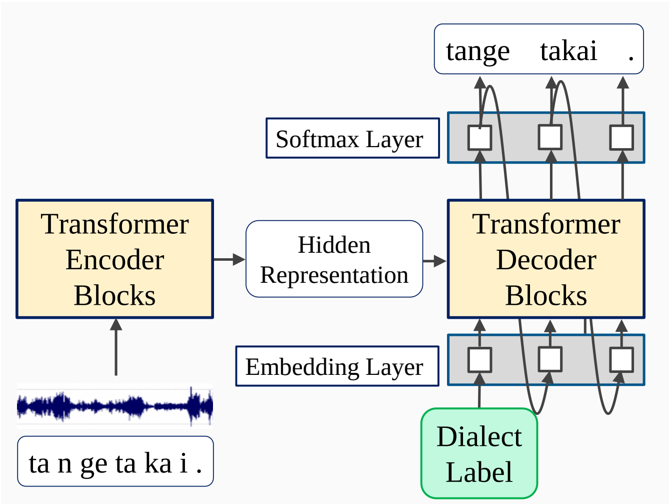

Fig. 1. Transformer encoder-decoder model using dialect label as auxiliary feature

label d and input speech. The generation probability of $\boldsymbol{W}$ is defined as

$$P(\boldsymbol{W}|\boldsymbol{X}, d; \boldsymbol{\Theta}) = \prod_{n=1}^{N} P(w_n|\boldsymbol{W}_{1:n-1}, d, \boldsymbol{X}; \boldsymbol{\Theta}). \quad (11)$$

Figure 1 shows the structure of the proposed model. The left side of Fig. 1 depicts a speech encoder part, and the right side shows a text decoder part. The speech encoder adopts the same modeling as equations (2) - (4). In the text decoder, the predictive probability of the n-th token w_n is calculated using equation (5) and the dialect label d as

$$P(w_n|\boldsymbol{W}_{1:n-1}, d, \boldsymbol{X}; \boldsymbol{\Theta}) = \mathrm{Softmax}(\boldsymbol{u}_{n-1}^{(J)}; \theta_{\mathrm{dec}}). \quad (12)$$

Equations (6) and (8) are re-defined using the dialect label d as

$$\boldsymbol{u}_{n-1}^{(j)} = \mathrm{TransformerDecoderBlock}(\boldsymbol{U}_{1:n-1}^{(j-1)}, \boldsymbol{H}^{I}; \theta_{\mathrm{dec}}), \quad (13)$$

$$\boldsymbol{U}_{1:n-1}^{(0)} = \{\boldsymbol{d}, \boldsymbol{u}_1^{(0)}, \cdots, \boldsymbol{u}_{n-1}^{(0)}\}, \quad (14)$$

$$\boldsymbol{u}_{n-1}^{(0)} = \mathrm{AddPostionalEncoding}(\boldsymbol{w}_{n-1}), \quad (15)$$

$$\boldsymbol{u}_{n-1}^{(0)} = \mathrm{Embedding}(w_{n-1}; \theta_{\mathrm{dec}}), \quad (16)$$

$$\boldsymbol{d} = \mathrm{Embedding}(d; \theta_{\mathrm{dec}}). \quad (17)$$

By inputting the dialect label, the dialect-aware modeling can be performed. The predicted probability is calculated from the embedding layer using the dialect label d as shown in equations (12) - (17). The model parameter sets can be optimized from a set of speech, dialect label, and text as

$$\mathcal{D} = \{(\boldsymbol{X}^1, d^1, \boldsymbol{W}^1), ..., (\boldsymbol{X}^T, d^T, \boldsymbol{W}^T)\}. \quad (18)$$

The objective function used in the proposed method is defined as

$$\mathcal{L}_{\mathrm{mle}}(\theta_{\mathrm{enc}}, \theta_{\mathrm{dec}}) = -\sum_{t=1}^{T}\sum_{n=1}^{N^t} \log P(w_n^t|\boldsymbol{W}_{1:n-1}^t, d^t, \boldsymbol{X}^t; \theta_{\mathrm{enc}}, \theta_{\mathrm{dec}}). \quad (19)$$

By optimizing with dialect labels, speech text, and obtaining the generation probabilities, the information of each dialect is enhanced, and the dialect-specific characteristics are clarified.

TABLE I
NUMBER OF UTTERANCES FOR EACH DIALECT AND STANDARD LANGUAGE

	Region	Training	Validation	Test	All
Dialect	Aomori	10,741	676	676	12,093
	Hiroshima	18,670	566	567	19,803
	Kumamoto	9,328	719	719	10,766
	Nagoya	18,611	551	550	19,713
	Sapporo	15,955	678	678	17,311
	Sendai	16,512	535	535	17,582
Standard	CSJ	162,243	1,292	2,573	166,108

TABLE II
NUMBERS OF MALE / FEMALE SPEAKERS FOR EACH DIALECT

Region	Male	Female
Aomori	34	36
Hiroshima	41	44
Kumamoto	31	41
Nagoya	38	43
Sapporo	42	44
Sendai	47	47

IV. EXPERIMENT

We conducted Japanese dialect ASR experiments to confirm the effectiveness of the proposed method.

A. Database

A home-made speech database of Japanese dialects and a database of standard Japanese were used in all experiments. The dialect database consisted of six dialects: Aomori, Hiroshima, Kumamoto, Nagoya, Sapporo, and Sendai [22]. For the standard language database, the Corpus of Spontaneous Japanese (CSJ) [23] consisting of academic lectures and simulated public speeches was used. The numbers of utterances for each dialect and CSJ are shown in Table I. CSJ contains three test sets: Eval 1, Eval 2, and Eval 3. In the experiments, Eval 2 was used as the development data, and Eval 1 and Eval 3 were used as the test data. The content of these test sets was academic lectures. As shown in Table II, the gender ratios of the speakers in the dialect database were almost the same for each dialect. Each dialect utterance was recorded by using an iPhone 5 or an Xperia Z1. The length of each dialect utterance was about 7 seconds, and the content of the dialect database was daily conversations. All transcriptions of the dialect database were hand-labeled. Both databases were sampled at 16 kHz and quantized to 16 bit.

B. Experiment conditions

The transformer-based encoder-decoder modeling was performed to construct end-to-end ASR systems. The transformer-based network consisted of eight encoder blocks and six decoder blocks. All functions used in the transformer networks were implemented in accordance with [10]. Regarding the composition of the transformer blocks, the dimension of the continuous vector was 256, the dimension of inner outputs

TABLE III
CERs (%) OF CONVENTIONAL AND PROPOSED METHODS FOR EACH COMBINATION OF DATABASES FOR TRAINING AND TESTING

		Test Data		
	Train Data	Dialect only	Standard only	Dialect + Standard
Conventional Method	Dialect only	52.9	100 ↑	86.2
	Standard only	52.5	14.3	35.4
	Dialect + Standard	9.3	16.6	12.5
Proposed Method	Dialect + Standard	7.1	13.8	10.1

TABLE IV
CERs (%) AND RELATIVE IMPROVEMENT (%) FOR EACH DIALECT USING CONVENTIONAL AND PROPOSED METHODS

Region	Conventional Method	Proposed Method	Relative Improvement
Aomori	5.9	2.7	54.2
Hiroshima	7.5	5.8	22.7
Kumamoto	4.0	2.2	45.0
Nagoya	10.8	8.7	19.4
Sapporo	17.3	16.0	7.5
Sendai	10.2	7.3	28.4

in the position-wise feed-forward networks was 2,048, and the number of attention heads was set to 4. For the speech encoder, we used 40-order log mel-scale filterbank coefficients appended with delta and acceleration coefficients as acoustic features. The frame length and the frame shift were 25 ms and 10 ms, respectively. The acoustic features were down-sampled to 1/4 along the time-axis via two convolutional layers and max pooling ones with a stride of two. In the text decoder, the dimension of word embeddings was 256, and the optimizer used was the rectified Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-9}$ [24]. The mini-batch size was set to 16 utterances. The dropout rate in the transformer blocks was set to 0.1. For the ASR decoding, we used a beam search algorithm in which the beam size was set to 20. In the proposed method, the dialect labels of the six dialects of the dialect database were used as the auxiliary features was set to six dialect along the dialect database. The dialect label was put in the embedding layer and treated as 256 dimensions. Transformer-based end-to-end ASR without any dialect labels was regarded as the conventional method. As the evaluation index, the character error rate (CER) was used:

$$\mathrm{CER} = (1 - \frac{\mathrm{COR} - \mathrm{INS}}{\mathrm{TOTAL}}) \times 100\ (\%), \tag{20}$$

where COR and INS were the numbers of correct characters and inserted characters, respectively. TOTAL was the total number of characters.

C. Result

Table III shows the CERs of the conventional method and the proposed one for each combination of databases for training and testing. In the case of using only the dialect database for testing, the CERs of the conventional method using dialect only and standard language only were extremely high. This means that there are two serious problems; each dialect data was not enough to train the end-to-end ASR model. The other problem was that the end-to-end ASR system constructed with the standard language was not suitable for recognizing the dialects. However, the conventional method using both dialect and standard language obtained a much lower CER than the conventional method with a single database. This means that the multi-condition modeling was able to relax the two problems. Furthermore, the CER of the proposed method had the lowest value among the methods using only the dialect database for testing. The results demonstrate that the proposed modeling using dialect labels can compensate for the mismatch between dialects and standard language adequately.

In the case of using only the standard-language database for testing, the CER of the conventional method using dialect only was over 100% due to the large number of insertion errors. Compared with the conventional method using only standard language with that using both databases, the CER of the conventional method using only standard language was lower. This indicates that the performance of simple multi-condition modeling was insufficient because of mismatches between the dialect and standard language. In contrast, the proposed method had the lowest CER in this testing case as well. These results demonstrate that the proposed method can effectively use both the dialect database and the standard language one. Consequently, in the case of using both for testing, the proposed method achieved an error reduction of 19.2%, compared with the simple multi-condition modeling.

Table IV illustrates the CERs of simple multi-condition modeling and the proposed method for each region. The training condition was the same as in the case of using only dialect data for testing, and both databases were used for training. From the results, the CERs of the proposed method for Aomori and Kumamoto showed an error reduction of around 50%, compared with those of the simple multi-condition modeling. On the other hand, the error reduction rate for Sapporo was the smallest. To investigate the trend in CERs for each region, additional experiments in which the amount of training data for each region was the same were performed. The results demonstrated that the trend in CERs for each dialect was not dependent on the data amount.

V. Conclusion

In this paper, we proposed a dialect-aware modeling method that utilizes dialect labels as auxiliary features for a transformer-based end-to-end ASR system. The proposed modeling could compensate for the mismatch between dialects and standard language; thus, both types pf data were effectively used for conducting the end-to-end ASR systems. The experimental results showed that the proposed dialect-aware modeling outperformed the simple multi-condition modeling.

As future work, we will investigate the trend in error reductions for each region. Additionally, we will consider estimating dialect labels automatically.

REFERENCES

[1] Suyoun Kim, Takaaki Hori, and Shinji Watanabe. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In *proc. ICASSP*, pages 4835–4839, 2017.

[2] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and Mandarin. In *proc. ICML*, pages 173–182, 2016.

[3] Albert Zeyer, Parnia Bahar, Kazuki Irie, Ralf Schlüter, and Hermann Ney. A comparison of transformer and lstm encoder decoder models for asr. In *proc. ASRU*, pages 8–15, 2019.

[4] Liang Lu, Xingxing Zhang, and Steve Renais. On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition. In *proc. ICASSP*, pages 5060–5064, 2016.

[5] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253, 2017.

[6] Takaaki Hori, Shinji Watanabe, Yu Zhang, and William Chan. Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM. In *proc. INTERSPEECH*, pages 949–953, 2017.

[7] Ryo Masumura, Mana Ihori, Akihiko Takashima, Takafumi Moriya, Atsushi Ando, and Yusuke Shinohara. Sequence-level consistency training for semi-supervised end-to-end automatic speech recognition. In *proc. ICASSP*, pages 7054–7058, 2020.

[8] Linhao Dong, Shuang Xu, and Bo Xu. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *proc. ICASSP*, pages 5884–5888, 2018.

[9] Shigeki Karita, Nelson Enrique Yalta Soplin, Shinji Watanabe, Marc Delcroix, Atsunori Ogawa, and Tomohiro Nakatani. Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration. In *proc. INTERSPEECH*, pages 1408–1412, 2019.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *proc. NIPS*, pages 5998–6008, 2017.

[11] Sheng Li, Dabre Raj, Xugang Lu, Peng Shen, Tatsuya Kawahara, and Hisashi Kawai. Improving transformer-based speech recognition systems with compressed structure and speech attributes augmentation. In *proc. INTERSPEECH*, pages 4400–4404, 2019.

[12] Tomoki Hayashi, Shinji Watanabe, Yu Zhang, Tomoki Toda, Takaaki Hori, Ramon Astudillo, and Kazuya Takeda. Back-translation-style data augmentation for end-to-end ASR. In *proc. SLT*, pages 426–433, 2018.

[13] Adithya Renduchintala, Shuoyang Ding, Matthew Wiesner, and Shinji Watanabe. Multi-modal data augmentation for end-to-end ASR. In *proc. INTERSPEECH*, pages 2394–2398, 2018.

[14] Suwon Shon, Ahmed Ali, and James Glass. Convolutional neural networks and language embeddings for end-to-end dialect recognition. In *proc. Odyssey*, pages 98–104, 2018.

[15] Yue Zhao, Jianjian Yue, Xiaona Xu, Licheng Wu, and Xiali Li. End-to-end-based tibetan multitask speech recognition. *IEEE Access*, 7:162519–162529, 2019.

[16] Sheng Li, Xugang Lu, Chenchen Ding, Peng Shen, Tatsuya Kawahara, and Hisashi Kawai. Investigating radical-based end-to-end speech recognition systems for chinese dialects and japanese. In *proc. INTERSPEECH*, pages 2200–2204, 2019.

[17] Sei Ueno, Takafumi Moriya, Masato Mimura, Shinsuke Sakai, Yusuke Shinohara, Yoshikazu Yamaguchi, Yushi Aono, and Tatsuya Kawahara. Encoder transfer for attention-based acoustic-to-word speech recognition. In *proc. INTERSPEECH*, pages 2424–2428, 2018.

[18] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *proc. ICCV*, pages 1406–1415, 2019.

[19] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *proc. CVPR*, pages 4893–4902, 2019.

[20] Lahiru Samarakoon, Brian Mak, and Albert YS Lam. Domain adaptation of end-to-end speech recognition in low-resource settings. In *proc. SLT*, pages 382–388, 2018.

[21] Anirudh Mani, Shruti Palaskar, Nimshi Venkat Meripo, Sandeep Konam, and Florian Metze. ASR error correction and domain adaptation using machine translation. In *proc. ICASSP*, pages 6344–6348, 2020.

[22] Ryo Imaiuzmi, Ryo Masumura, Sayaka Shiota, and Hitoshi Kiya. Japanese dialect speech classification using sequence-to-one neural networks. in japanese. In *proc. SP2019-57*, volume 119, pages 41–46, 2020.

[23] Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. Spontaneous speech corpus of japanese. In *proc. LREC*, pages 947–952, 2000.

[24] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *proc. ICLR*, 2020.