Dialect-aware Semi-supervised Learning for End-to-End Multi-dialect Speech Recognition

Sayaka Shiota* Ryo Imaizumi† Ryo Masumura* and Hitoshi Kiya*
* Tokyo Metropolitan University, Department of Computer Science, Japan
† NTT Corporation, NTT Computer and Data Science Laboratories, Japan

Abstract—**In this paper, we propose dialect-aware semi-supervised learning for end-to-end automatic speech recognition (ASR) models considering multi-dialect speech. Some multi-domain ASR tasks require a large amount of training data containing additional information (e.g., language or dialect), whereas it is difficult to prepare such data with accurate transcriptions. Semi-supervised learning is a method of using a massive amount of untranscribed data effectively, and it can be applied to multi-domain ASR tasks to relax the missing transcriptions problem. However, semi-supervised learning has usually used generated pseudo-transcriptions only. The problem is that simply combining a multi-domain model with semi-supervised learning makes use of no additional information even though the information can be obtained. Therefore, in this paper, we focus on semi-supervised learning based on a multi-domain model that takes additional domain information into account. Since the accuracy of pseudo-transcriptions can be improved by using the multi-domain model and additional information, our proposed semi-supervised learning is expected to provide a reliable ASR model. In experiments, we performed Japanese multi-dialect ASR as one type of multi-domain ASR. From the results, a model trained with the proposed method yielded the lowest character error rate compared with other models trained with the conventional semi-supervised method.**

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 (E2E) ASR has been proposed [1], [2]. It is known that the performance of E2E ASR depends on the amount of training data [3], [4]. To use a large amount of training data, databases with different domains such as recording environments and spoken styles are used in combination. Thus, many methods have been reported that consider this domain-mismatch problem, e.g., multi-dialect speech recognition and multilingual speech recognition [5], [6]. Multi-task learning, a multi-domain task, has been proposed to use additional domain information (e.g., language or dialect) effectively [7], [8]. As an example of a multi-task model, multi-dialect ASR has been reported to provide high performance since dialect information can be used as one form of additional domain information [9].

It is assumed that the training of multi-task models requires a large amount of sets of speech data, domain information, and accurate transcriptions. However, it is difficult to prepare accurate transcriptions corresponding to the data. In contrast, semi-supervised learning is a method that uses a massive amount of untranscribed data effectively. In the procedure of semi-supervised learning for ASR, first, a teacher model is trained from a small amount of accurately transcribed data, and pseudo-transcriptions are generated by the teacher model using untranscribed data. A large amount of untranscribed data can be used as speech-to-text pair data, so semi-supervised learning provides high performance in many ASR tasks [10]–[12]. However, since the conventional semi-supervised learning is performed for a single task, additional information is ignored even though the information can be used for a multi-task model.

In this paper, we propose a semi-supervised learning method using the multi-task ASR model to make use of additional information. In the proposed method, a teacher multi-task ASR model considering additional information is estimated from a small amount of accurate paired data. The multi-task-based teacher model can generate reliable pseudo-transcriptions by using the additional information. Thus, the performance of the student ASR model is also improved by using reliable pseudo-transcriptions. In experiments, we performed Japanese multi-dialect ASR as one type of multi-task ASR, and a DID2ASR model [13] was used. From the results, the model trained with the proposed method yielded the lowest character error rate compared with other models trained with the conventional semi-supervised method.

This paper is organized as follows. Section II describes the transformer-based ASR model architecture and semi-supervised ASR. Then, the model architecture for the proposed learning and dialect-aware semi-supervised ASR are presented in Section III. Experimental conditions and results are shown in Section IV. Finally, Section V concludes our work.

II. Conventional method

A. Transformer-based encoder-decoder ASR

This section describes the transformer-based encoder-decoder ASR, one of the state-of-the-art E2E ASR models [14]–[17]. 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 model defines the

978-616-590-477-3 ©2022 APSIPA

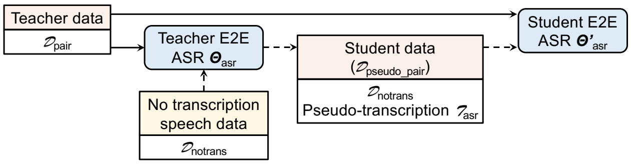
Teacher data
$\mathcal{D}_{\text{pair}}$
Teacher E2E ASR $\boldsymbol{\Theta}_{\text{asr}}$
Student data ($\mathcal{D}_{\text{pseudo\_pair}}$)
$\mathcal{D}_{\text{notrans}}$
Pseudo-transcription $\mathcal{T}_{\text{asr}}$
Student E2E ASR $\boldsymbol{\Theta}'_{\text{asr}}$
No transcription speech data
$\mathcal{D}_{\text{notrans}}$

Fig. 1. Training process of semi-supervised learning in E2E ASR

generation probability of $\boldsymbol{W}$ as

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

where $\boldsymbol{\Theta}_{\text{asr}}$ represents model parameter sets, and $\boldsymbol{W}_{1:n-1} = \{w_1, ..., w_{n-1}\}$. In the transformer-based E2E ASR system, $P(w_n|\boldsymbol{W}_{1:n-1},\boldsymbol{X};\boldsymbol{\Theta}_{\text{asr}})$ can be computed using a speech encoder and a text decoder. The speech encoder converts input acoustic features into hidden representations using transformer encoder blocks. The text decoder computes the generation probability of a token from generated tokens and the hidden representations of the speech by using transformer decoder blocks. In this E2E ASR system, a whole model parameter set can be optimized from speech-to-text paired data:

$$\mathcal{D}_{\text{pair}} = \{(\boldsymbol{X}^1, \boldsymbol{W}^1), ..., (\boldsymbol{X}^T, \boldsymbol{W}^T)\}, \quad (2)$$

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

$$\begin{aligned}\mathcal{L}_{\text{mle}}(\boldsymbol{\Theta}_{\text{asr}}) = -\sum_{t=1}^{T}\sum_{n=1}^{N^t} \\ \log P(w_n^t|\boldsymbol{W}_{1:n-1}^t,\boldsymbol{X}^t;\boldsymbol{\Theta}_{\text{asr}}),\end{aligned} \quad (3)$$

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.

B. Semi-supervised ASR

This section details training for semi-supervised E2E ASR models. In semi-supervised learning, a small amount of paired data is used to train a teacher model, which is used for decoding untranscribed speech data $\mathcal{D}_{\text{notrans}}$ and generating pseudo-transcriptions [18]–[20].

$$\mathcal{D}_{\text{notrans}} = \{\boldsymbol{X}^{T+1}, ..., \boldsymbol{X}^{T+L}\}, \quad (4)$$

where L is the number of utterances in the speech data without corresponding transcriptions. The generated pseudo-transcriptions defined as $\mathcal{T}_{\text{asr}} = \{\boldsymbol{W}^{T+1}, ..., \boldsymbol{W}^{T+L}\}$ can be used as paired data of $\mathcal{D}_{\text{notrans}}$. The pair data $\mathcal{D}_{\text{pseudo\_pair}}$ is defined from $\mathcal{D}_{\text{notrans}}$ and $\mathcal{T}_{\text{asr}}$.

$$\mathcal{D}_{\text{pseudo\_pair}} = \{(\boldsymbol{X}^{T+1}, \boldsymbol{W}^{T+1}), ..., (\boldsymbol{X}^{T+L}, \boldsymbol{W}^{T+L})\}. \quad (5)$$

Fig. 1 shows the training procedure of semi-supervised learning for an E2E ASR model. First, we train a teacher E2E ASR model with parameters $\boldsymbol{\Theta}_{\text{asr}}$ using the teacher data $\mathcal{D}_{\text{pair}}$ as in Subsection II.A. Then, the trained teacher E2E ASR is used to recognize the untranscribed speech data $\mathcal{D}_{\text{notrans}}$ and generate automatic transcription $\mathcal{T}_{\text{asr}}$. Next, the semi-supervised ASR model is trained using $\mathcal{D}_{\text{pair}}$ and $\mathcal{D}_{\text{pseudo\_pair}}$, and the objective function is defined as

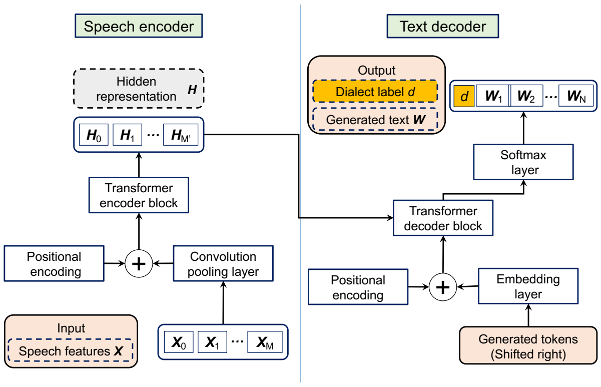
Speech encoder
Hidden representation $\boldsymbol{H}$
$\boldsymbol{H}_0$ $\boldsymbol{H}_1$ ··· $\boldsymbol{H}_{M'}$
Transformer encoder block
Positional encoding
Convolution pooling layer
Input
Speech features $\boldsymbol{X}$
$\boldsymbol{X}_0$ $\boldsymbol{X}_1$ ··· $\boldsymbol{X}_M$
Text decoder
Output
Dialect label d
Generated text $\boldsymbol{W}$
d $\boldsymbol{W}_1$ $\boldsymbol{W}_2$ ··· $\boldsymbol{W}_N$
Softmax layer
Transformer decoder block
Positional encoding
Embedding layer
Generated tokens (Shifted right)

Fig. 2. Network architecture of multi-task learning systems performing DID and MD-ASR in series (DID2ASR)

$$\mathcal{L}'_{\text{asr}}(\boldsymbol{\Theta}'_{\text{asr}}) = -\sum_{t=1}^{T+L}\sum_{n=1}^{N^t} \log P(w_n^t|\boldsymbol{W}_{1:n-1}^t,\boldsymbol{X}^t;\boldsymbol{\Theta}'_{\text{asr}}). \quad (6)$$

The problem with E2E ASR with the conventional semi-supervised learning is that it uses only transcriptions and ignores additional domain information.

III. Proposed Method

A. Model architecture (DID2ASR)

DID2ASR is a model proposed for multi-task learning of dialect identification and multi-dialect ASR [13]. In DID2ASR, the generation probability of a dialect label d is estimated first, and then the generation probability of text $\boldsymbol{W}$ is estimated as follows.

$$P(\boldsymbol{W}, d|\boldsymbol{X};\boldsymbol{\Theta}_{\text{d2a}}) = P(\boldsymbol{W}|\boldsymbol{X}, d;\boldsymbol{\Theta}_{\text{d2a}})P(d|\boldsymbol{X};\boldsymbol{\Theta}_{\text{d2a}}) \quad (7)$$

Fig. 2 shows the network architecture for DID2ASR. The architecture of the speech encoder is the same as the general transformer-based E2E ASR described in Section II. As is shown, the text decoder of DID2ASR is essentially the same as the general transformer-based E2E ASR, although a softmax function is used to calculate the generation probability of the dialect label and text. The model parameter set can be optimized from a set of speech, dialect label, and text as

$$\mathcal{D}_{\text{set}} = \{(\boldsymbol{X}^1, d^1, \boldsymbol{W}^1), ..., (\boldsymbol{X}^T, d^T, \boldsymbol{W}^T)\}. \quad (8)$$

The objective function used in the proposed method is defined as

$$\mathcal{L}_{\text{d2a}}(\boldsymbol{\Theta}_{\text{d2a}}) = -\sum_{t=1}^{T} \log P(\boldsymbol{W}^t, d^t|\boldsymbol{X}^t;\boldsymbol{\Theta}_{\text{d2a}}). \quad (9)$$

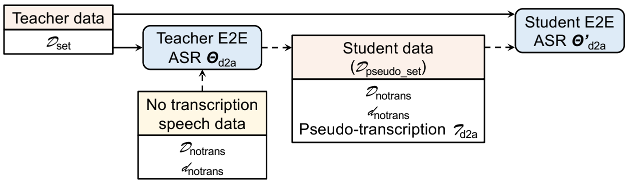
Teacher data
$\mathcal{D}_{\text{set}}$
Teacher E2E ASR $\boldsymbol{\Theta}_{\text{d2a}}$
No transcription speech data
$\mathcal{D}_{\text{notrans}}$
d_{notrans}
Student data ($\mathcal{D}_{\text{pseudo\_set}}$)
$\mathcal{D}_{\text{notrans}}$
d_{notrans}
Pseudo-transcription $\mathcal{T}_{\text{d2a}}$
Student E2E ASR $\boldsymbol{\Theta}'_{\text{d2a}}$

Fig. 3. Semi-supervised learning with dialect labels in DID2ASR for transcription generation

While this model has the characteristic of the performance of ASR being affected by the accuracy of the dialect identification, it has been reported that the DID2ASR model yielded reliable performance when the dialect label is known.

B. Dialect-aware semi-supervised learning

We propose a semi-supervised learning that takes dialect labels into account by using DID2ASR, which is called dialect-aware semi-supervised learning. In dialect-aware semi-supervised learning, a small amount of set data $\mathcal{D}_{\text{set}}$ is used to train a DID2ASR-based teacher model. Then, a large amount of untranscribed data $\mathcal{D}_{\text{set\_notrans}}$ is decoded by the teacher model with the given dialect label.

$$d_{\text{notrans}} = \{d^{T+1}, ..., d^{T+L}\}. \tag{10}$$

$$\boldsymbol{W} = \arg\max_{\boldsymbol{W}} P(\boldsymbol{W}|\boldsymbol{X}_{\text{notrans}}, d_{\text{notrans}}). \tag{11}$$

By using the teacher model with the dialect label, reliable pseudo-transcriptions $\mathcal{T}_{d2a} = \{\boldsymbol{W}_{T+1}, ..., \boldsymbol{W}_{T+L}\}$ can be generated. As with Subsection Ⅱ.B, the generated pseudo-transcriptions can be used as a set of data combined with $\mathcal{D}_{\text{notrans}}$, d_{notrans} and $\mathcal{T}_{d2a}$. The pseudo set of data $\mathcal{D}_{\text{pseudo\_set}}$ is defined as :

$$\begin{aligned} &\mathcal{D}_{\text{pseudo\_set}} = \\ &\{(\boldsymbol{X}^{T+1}, \boldsymbol{W}^{T+1}, d^{T+1}), ..., (\boldsymbol{X}^{T+L}, \boldsymbol{W}^{T+L}, d^{T+L})\}. \end{aligned} \tag{12}$$

To train a DID2ASR-based student model with a pseudo set of data $\mathcal{D}_{\text{pseudo\_set}}$, the proposed method finally provides a reliable multi-dialect ASR model.

Fig. 3 shows the training procedure of dialect-aware semi-supervised learning for the DID2ASR model. Dialect-aware semi-supervised learning is trained using $\mathcal{D}_{\text{set}}$ and $\mathcal{D}_{\text{pseudo\_set}}$, and the objective function is defined as

$$\mathcal{L}'_{\text{d2a}}(\boldsymbol{\Theta}'_{\text{d2a}}) = -\sum_{t=1}^{T+L} \log P(\boldsymbol{W}^t, d^t|\boldsymbol{X}^t; \boldsymbol{\Theta}'_{\text{d2a}}). \tag{13}$$

Using a dialect-aware supervised ASR model is expected to improve the accuracy of pseudo-transcriptions. Thus, the performance of the semi-supervised ASR model is also improved by using high accuracy pseudo-transcriptions.

TABLE I
DATA AMOUNT FOR EACH DIALECT AND STANDARD LANGUAGE [HOURS]

Region	Teacher	Student	Validation	Test
Aomori	1.92	19.77	1.07	1.01
Hiroshima	1.87	33.63	1.01	1.01
Kumamoto	1.55	17.77	1.22	1.22
Nagoya	1.82	29.77	1.12	1.12
Sapporo	1.75	30.43	1.18	1.18
Sendai	2.10	33.78	1.15	1.15
Standard	290	-	-	-

IV. EXPERIMENT

A. Database

A home-made speech database of Japanese dialects [9] 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. 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. The gender ratios of the speakers for each dialect were almost the same. For the standard Japanese language database, the Corpus of Spontaneous Japanese (CSJ) [21], consisting of academic lectures and simulated public speeches, was used. Standard language was used only for training an initial teacher model. The number of male speakers was about double that of female speakers. Both databases were sampled at 16 kHz and quantized to 16 bit. The data amount for each data set are shown in Table I. The training data of the dialect database was divided into teacher data and student data. The teacher data was a set of speech, manually transcribed speech, and manually labeled dialect labels, and the student data was pairs of speech and manually labeled dialect labels.

B. Model details

As shown in Section Ⅱ, both proposed models were based on the transformer-based ASR system. The transformer-based E2E ASR described in Subsection Ⅱ.A was used as the conventional method. As described in Subsection Ⅲ.B, the model architectures of the conventional transformer model and DID2ASR model were almost the same. The unified training conditions were set as follow. The transformer network consisted of eight encoder blocks and six decoder blocks. All functions used in the transformer networks were implemented in accordance with [14]. Regarding the composition of the transformer blocks, the dimension of the continuous vector was 256, the dimension of the inner outputs in the position-wise feed-forward networks was 2,048, and the number of attention heads was set to 4. The parameters for the speech encoder and the text decoder were the same as in [9]. For DID2ASR, dialect labels were put in the embedding layer and treated as a 256-dimensional vector for *aom, hir, kum, nag, sap, sen*, and one standard language label; *jap*. A dialect label d was regarded as one token.

TABLE II
CERs AT EACH ITERATION OF SUPERVISED AND SEMI-SUPERVISED LEARNING WITH TRANSFORMER MODEL AND DID2ASR MODEL, AND DIALECT-AWARE SEMI-SUPERVISED LEARNING WITH DID2ASR MODEL.

Model	Learning	Iteration	CER
Transformer	Supervised	Initial	22.9
	Semi-supervised	First	18.2
		Second	17.9
		Third	17.6
DID2ASR	Supervised	Initial	23.4
	Semi-supervised	First	20.2
		Second	18.8
		Third	18.1
	Dialect-aware semi-supervised (Proposed)	First	18.1
		Second	17.6
		Third	17.4

TABLE III
CERs (%) AT EACH ITERATION OF SUPERVISED AND SEMI-SUPERVISED LEARNING OF PSEUDO-TRANSCRIPTIONS FOR SPEECH DATA WITHOUT TRANSCRIPTION

Model	Learning	Iteration	CER
Transformer	Supervised	First	12.0
	Semi-supervised	Second	8.6
		Third	7.7
DID2ASR	Supervised	First	14.1
	Semi-supervised	Second	9.8
		Third	8.6
	Dialect-aware supervised	First	10.1
	Dialect-aware semi-supervised	Second	7.4
		Third	6.5

C. Training detail

Semi-supervised learning has been reported to improve performance by iteratively generating pseudo-transcriptions [22]–[24]. Therefore, in the experiments, we generated pseudo-transcriptions by iterating three times for all conditions and evaluated the results. The initial state of the iteration was the time when supervised learning with the standard language data and a small amount of dialect data was performed, and three iterations of semi-supervised learning were conducted. Supervised learning was based on the teacher data of speech, manual transcriptions, and dialect labels. Semi-supervised learning used both teacher and student data. For semi-supervised learning, three learning methods were used as comparison methods with the above two models. First, semi-supervised learning was used, which is used in the conventional transformer model. Second, the DID2ASR model was used to estimate dialect labels and generate automatic transcriptions, and the transcriptions were used for ASR in semi-supervised learning. Third, the DID2ASR model was used to generate automatic transcriptions with known dialect labels, and it performed dialect-aware semi-supervised learning using the transcriptions. A seven-dimensional vector consisting of six dialects and standard Japanese language was outputted from the softmax layer.

TABLE IV
CERs (%) FOR EACH ITERATION OF SEMI-SUPERVISED LEARNING IN DID2ASR MODEL WHEN DIALECT IS KNOWN AT TESTING

Model	Learning	Iteration	CER
DID2ASR	Semi-supervised	First	19.7
		Second	18.2
		Third	17.7
	Dialect-aware semi-supervised (Proposed)	First	17.5
		Second	16.8
		Third	16.6

D. Result

The character error rates (CERs) at each iteration of supervised and semi-supervised learning for the comparison conditions presented in Section IV.C are shown in Table II. Comparing the transformer model and the DID2ASR model with supervised learning, the CERs were 22.9% and 23.4%, respectively. The reason DID2ASR with supervised learning had the lowest CER is that DID2ASR is affected by the accuracy of the dialect identification as described in Subsection Ⅲ.A. In fact, the accuracy of dialect identification for DID2ASR with supervised learning was 65.9%. Next, for semi-supervised learning, the CER of the transformer model was 18.2%, which was lower than the CER of the DID2ASR model, 20.2%. This confirmed that even in semi-supervised learning, the ASR performance is affected by the accuracy of the dialect identification. In comparison, the CER of the proposed dialect-aware semi-supervised learning was the lowest at 18.1%. The most important difference between DID2ASR with semi-supervised learning and DID2ASR with dialect-aware semi-supervised learning was the accuracy of the generated pseudo-transcriptions. Since the proposed method could use more reliable pseudo-transcriptions, the CER of the proposed method became lower than that of the DID2ASR model trained by the conventional semi-supervised learning. Next, we focused on the iterations for semi-supervised learning as shown Table II. The CER decreased gradually through the iterations. From this perspective, the proposed method could obtain the lowest CER. These results showed that the proposed learning can improve the performance of ASR effectively.

We investigated the reason for our improvements in detail, and the CERs of the pseudo-transcriptions are shown in Table III. The first iteration showed the CER is for the case when a pseudo-transcription was generated by the ASR model constructed using the supervised data, and the second and third iterations showed the CERs are for the cases when the pseudo-transcriptions were generated by iterating. Comparing the methods in each iteration, the CER of the Transformer model was lower than that of the DID2ASR model without the dialect in all conditions. In comparison, the CERs of the proposed learning with the DID2ASR model were the lowest in all conditions. This result confirms that one of the reasons for the improvement of the proposed method is the improvement in the accuracy of the pseudo-transcriptions.

Finally, we performed the experiment for the case where accurate dialect labels for test data were also given. The CERs of each condition are shown in Table IV. Compared with Table II, the trend was the same; however, the CERs of both DID2ASR semi-supervised learning and DID2ASR dialect-aware semi-supervised learning were significantly lower. In particular, for the proposed DID2ASR dialect-aware semi-supervised learning, the CER was the lowest at 16.6% for the third iteration in all experiments. From these results, it was shown that the proposed method could achieve the best performance with the most additional information when the dialect labels were known.

V. Conclusion

In this paper, we proposed dialect-aware semi-supervised learning for E2E ASR models considering multi-dialect speech. A multi-task model, DID2ASR, was used to generate highly accurate pseudo-transcriptions. Using the pseudo-transcriptions and the dialect labels, we demonstrated that the proposed method can achieve high performance in ASR. This result showed that the proposed method can be useful for multi-domain tasks by utilizing additional information in semi-supervised learning.

As future work, we will experiment with other networks such as the CTC/attention hybrid system using other dialects. Also, we will experiment with dynamic transcription generation using student data.

References

[1] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *proc. ICASSP*, 2017, pp. 4835–4839.

[2] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and Mandarin," in *proc. ICML*, 2016, pp. 173–182.

[3] T. Hayashi, S. Watanabe, Y. Zhang, T. Toda, T. Hori, R. Astudillo, and K. Takeda, "Back-translation-style data augmentation for end-to-end ASR," in *proc. SLT*, 2018, pp. 426–433.

[4] A. Renduchintala, S. Ding, M. Wiesner, and S. Watanabe, "Multi-modal data augmentation for end-to-end ASR," in *proc. INTERSPEECH*, 2018, pp. 2394–2398.

[5] S. Dalmia, R. Sanabria, F. Metze, and A. W. Black, "Sequence-based multi-lingual low resource speech recognition," in *proc. ICASSP*, 2018, pp. 4909–4913.

[6] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Weinstein, and K. Rao, "Multilingual speech recognition with a single end-to-end model," in *proc. ICASSP*, 2018, pp. 4904–4908.

[7] T. Moriya, T. Ochiai, S. Karita, H. Sato, T. Tanaka, T. Ashihara, R. Masumura, Y. Shinohara, and M. Delcroix, "Self-distillation for improving ctc-transformer-based ASR systems." in *proc. INTERSPEECH*, 2020, pp. 546–550.

[8] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *Transaction on speech and audio processing*, vol. 4, no. 1, p. 31, 1996.

[9] R. Imaiuzmi, R. Masumura, S. Shiota, and H. Kiya, "Dialect-aware modeling for end-to-end japanese dialect speech recognition." in *proc. APSIPA ASC*, 2020, pp. 297–301.

[10] D. S. Park, Y. Zhang, Y. Jia, W. Han, C.-C. Chiu, B. Li, Y. Wu, and Q. V. Le, "Improved noisy student training for automatic speech recognition," in *proc. INTERSPEECH*, 2020, pp. 2817–2821.

[11] Y. Higuchi, N. Moritz, J. L. Roux, and T. Hori, "Momentum pseudo-labeling for semi-supervised speech recognition," in *proc. INTERSPEECH*, 2021, pp. 726–730.

[12] R. Masumura, M. Ihori, A. Takashima, T. Moriya, A. Ando, and Y. Shinohara, "Sequence-level consistency training for semi-supervised end-to-end automatic speech recognition," in *proc. ICASSP*, 2020, pp. 7054–7058.

[13] R. Imaizumi, R. Masumura, S. Shiota, and H. Kiya, "End-to-end japanese multi-dialect speech recognition and dialect identification with multi-task learning," *APSIPA Transactions on Signal and Information Processing (accepted)*, 2022.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Å. Kaiser, and I. Polosukhin, "Attention is all you need," in *proc. NIPS*, 2017, pp. 5998–6008.

[15] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *proc. ICASSP*, 2018, pp. 5884–5888.

[16] S. Karita, N. Enrique Yalta Soplin, S. Watanabe, M. Delcroix, A. Ogawa, and T. Nakatani, "Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration," in *proc. INTERSPEECH*, 2019, pp. 1408–1412.

[17] S. Li, D. Raj, X. Lu, P. Shen, T. Kawahara, and H. Kawai, "Improving transformer-based speech recognition systems with compressed structure and speech attributes augmentation." in *proc. INTERSPEECH*, 2019, pp. 4400–4404.

[18] F. Weninger, F. Mana, R. Gemello, J. Andrés-Ferrer, and P. Zhan, "Semi-supervised learning with data augmentation for end-to-end ASR," in *proc. INTERSPEECH*, 2020, pp. 2802–2806.

[19] J. Kahn, A. Lee, and A. Hannun, "Self-training for end-to-end speech recognition," in *proc. ICASSP*, 2020, pp. 7084–7088.

[20] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *proc. CVPR*, 2020.

[21] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese." in *proc. LREC*, 2000, pp. 947–952.

[22] A. Xiao, C. Fuegen, and A. Mohamed, "Contrastive semi-supervised learning for ASR," in *proc. ICASSP*, 2021, pp. 3870–3874.

[23] K. Singh, V. Manohar, A. Xiao, S. Edunov, R. Girshick, V. Liptchinsky, C. Fuegen, Y. Saraf, G. Zweig, and A. Mohamed, "Large scale weakly and semi-supervised learning for low-resource video ASR," in *proc. INTERSPEECH*, 2020, pp. 3770–3774.

[24] Q. Xu, T. Likhomanenko, J. Kahn, A. Hannun, G. Synnaeve, and R. Collobert, "Iterative pseudo-labeling for speech recognition," in *proc. INTERSPEECH*, 2020, pp. 1006–1010.