LETTER

End-to-End Multilingual Speech Recognition System with Language Supervision Training

Danyang LIU[†,††], **Ji XU**[†,†††a)], ***Nonmembers***, ***and*** **Pengyuan ZHANG**[†,††], ***Member***

SUMMARY End-to-end (E2E) multilingual automatic speech recognition (ASR) systems aim to recognize multilingual speeches in a unified framework. In the current E2E multilingual ASR framework, the output prediction for a specific language lacks constraints on the output scope of modeling units. In this paper, a language supervision training strategy is proposed with language masks to constrain the neural network output distribution. To simulate the multilingual ASR scenario with unknown language identity information, a language identification (LID) classifier is applied to estimate the language masks. On four Babel corpora, the proposed E2E multilingual ASR system achieved an average absolute word error rate (WER) reduction of 2.6% compared with the multilingual baseline system.

key words: *multilingual speech recognition, language-adaptive training, hybrid attention/CTC*

1. Introduction

The end-to-end (E2E) framework has been widely applied in the field of automatic speech recognition (ASR). Implementing ASR systems under the E2E framework adds flexibility by eliminating the need for pronunciation dictionaries [1]. Therefore, this paper focuses on the joint modeling of multilingual ASR under the E2E framework. Although acoustic model information and language model information can be shared among multiple languages under the E2E framework, it also causes new problems. Given that pronunciation and grammar rules differ among languages, multilingual joint modeling may result in confusion among languages compared with E2E monolingual systems. Therefore, it is necessary to apply language-adaptive training to the E2E multilingual ASR system to improve its language discriminability.

Some previous studies have explored the language-adaptive training of E2E multilingual ASR systems. The language-conditioned method was adopted in some studies [2]–[4] by injecting the language feature vectors into the model at different locations. To assist in the multilingual modeling process, an auxiliary LID task was introduced to the E2E multilingual ASR framework [2], [5]. Other approaches, such as adding language identity tags at the beginnings of transcripts [6] and introducing adapter modules in the E2E framework [7], have also been investigated. However, all the above methods of language-adaptive training require language information as a priori knowledge to assist in system construction. In addition, the above approaches do not directly constrain the output scope of multilingual modeling units for language-specific predictions.

This paper proposes a language supervision training strategy with language masks to constrain the neural network output distribution and weaken the probabilities of the modeling units of the non-target languages.

2. Language-Conditioned Methods

Language-conditioned language-adaptive training methods have been applied in attention-based [2], CTC-based [3], and RNN-T-based [4] E2E multilingual ASR systems. As shown in Fig. 1, the language-conditioned methods can be categorized into encoder- and decoder-conditioned methods under the attention-based encoder-decoder framework. The encoder-conditioned method is implemented by appending the spectral features with language identity information to improve the language discriminability of the front-end acoustic modeling:

$$\mathbf{h}^{enc} = Encoder(\mathbf{x}, \mathbf{v}(\mathbf{x})), \tag{1}$$

where $\mathbf{v}(\mathbf{x})$ is the language feature vector of the input acoustic feature sequence $\mathbf{x} = (x_t)_{t=1}^{T}$, and $\mathbf{h}^{enc}$ is the hidden state of the encoder module. In the decoder-conditioned method,

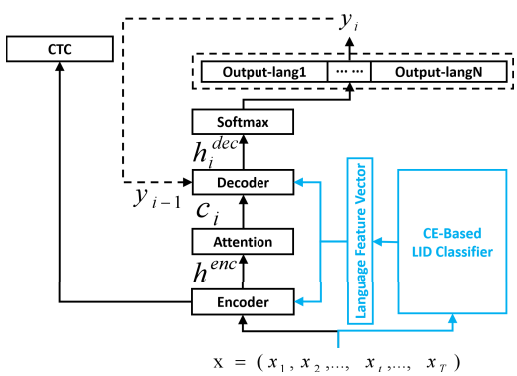

Fig. 1 The joint attention/CTC multilingual ASR framework based on language-conditioned methods: the encoder-conditioned method is implemented by inputting the language feature vector into the encoder module; the decoder-conditioned method is implemented by inputting the language feature vector into the decoder module.

Manuscript received November 29, 2019.
Manuscript revised February 6, 2020.
Manuscript publicized March 19, 2020.

[†]The authors are with the Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, China.

[††]The authors are also with School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, China.

a) E-mail: xuji@hccl.ioa.ac.cn

DOI: 10.1587/transinf.2019EDL8214

language information is input into the decoder module to assist with the back-end modeling unit prediction:

$$h_i^{dec} = Decoder(y_{i-1}, c_i, \mathbf{v}(\mathbf{x})), \tag{2}$$

$$p(y_i|y_{1:i-1}, \mathbf{x}) = softmax(h_i^{dec}), \tag{3}$$

where y_i and c_i are the output modeling unit and context vector of step i, respectively, and h_i^{dec} is the hidden state of the decoder module. In this paper, the language feature vector $\mathbf{v}(\mathbf{x})$ is generated from a cross-entropy-based (CE-based) LID classifier [8] trained with one-hot language identity labels.

3. Language Supervision Training Strategy

Figure 2 shows the proposed joint attention/CTC multilingual ASR framework [9] based on language supervision training strategy, where the language mask is estimated from a mean squared error (MSE)-based LID classifier. The flowchart of the recognition strategy and the flowchart of the training strategy (i.e. Fig. 2) are the same except that the historical information (y_{i-1}) is ground truth in the training process and is recognized label in the recognition process. In multilingual ASR systems, the multilingual modeling unit is a union of multiple languages. During the modeling unit prediction, probability distributions are generated for the entire set of multilingual units, including nontarget language modeling units. The nontarget language modeling units, however, may result in confusion during the language-specific training process. To eliminate the impact of the nontarget output nodes, a language supervision training strategy is applied with language masks. The output of the masked decoder module is defined as follows:

$$\mathbf{v}^{mask}(\mathbf{x}) = (\mathbf{v}_m(\mathbf{x}))_{m=1}^{M}, \mathbf{v}_m(\mathbf{x}) = \begin{cases} 1, m \in U(\mathbf{x}); \\ 0, m \notin U(\mathbf{x}), \end{cases} \tag{4}$$

$$p(y_i|y_{1:i-1}, \mathbf{x}) = renorm(\mathbf{v}^{mask}(\mathbf{x}) \odot softmax(h_i^{dec})), \tag{5}$$

where $\mathbf{v}^{mask}(\mathbf{x})$ is the language mask, which has the same size as the multilingual modeling unit, M in (4) is the total number of output nodes, which also has the same size as the multilingual modeling unit, and $U(\mathbf{x})$ is the modeling unit of the language to which $\mathbf{x}$ belongs. In (5), the masked multilingual modeling unit probability distributions are smoothed by the renorm function to ensure that the output probabilities sum to 1.

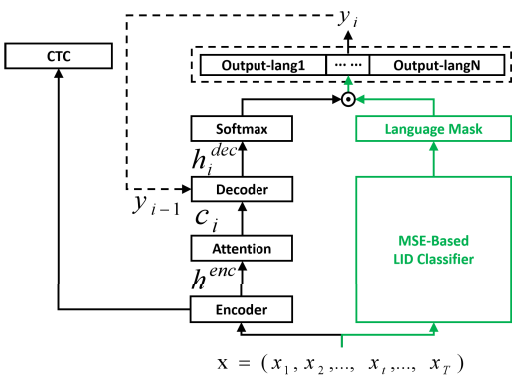

Fig. 2 The proposed joint attention/CTC multilingual ASR framework based on language supervision training strategy.

4. Language Masks Estimation

To accomplish the language supervision training of E2E multilingual ASR system, an LID classifier needs to be constructed in advance to estimate the language mask information. It is designed to simulate the multilingual ASR scenario, in which language identity information is not provided as prior knowledge during the testing process. The MSE function [10] is applied to minimize the error between the real distributions and the estimated distributions of the language masks. The MSE function is defined as follows:

$$\pounds_{mse} = \frac{1}{M}\sum_{m=1}^{M}(\mathbf{v}_m - \hat{\mathbf{v}}_m)^2, \tag{6}$$

where $\mathbf{v}_m$ and $\hat{\mathbf{v}}_m$ are the real and estimated probabilities of output node m, respectively. The ASR results of the E2E multilingual system are generated on utterance level information. Thus, the language mask vector is generated at the utterance level and is calculated by averaging all frame information of the utterance:

$$\hat{\mathbf{v}}^{mask}(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{v}}^{mask}(\mathbf{x}_t) \tag{7}$$

where $\hat{\mathbf{v}}^{mask}(\mathbf{x}) = (\hat{\mathbf{v}}_m)_{m=1}^{M}$ is the estimated language mask for input $\mathbf{x}$, and T is the number of frames in $\mathbf{x}$. It is worth mentioning that the statistics pooling component [11] is designed within the LID classifier to improve the accuracy of the language information prediction.

5. Experiments and Results

5.1 Speech Datasets and Experimental Configurations

In this paper, four IARPA Babel corpora are adopted to conduct the experiments: Tagalog (Tag), Cebuano (Ceb), Tok Pisin (Tok), and Haitian Creole (Hai). These four languages are close in terms of language family and have some pronunciation similarities, which is benefit for information sharing in multilingual joint modeling. Statistics of the four corpora are shown in Table 1. The system performance is measured using the 10-hour official IARPA Babel development sets, which are randomly divided into development sets and evaluation sets. During the model iterative training process, the optimal model is selected by calculating the recognition accuracy of the development set on the model.

The Espnet E2E Speech Processing Toolkit with a PyTorch back-end was adopted to perform all the ASR experiments [9]. The joint attention/CTC architecture was adopted to construct both monolingual and multilingual models,

Table 1 Multilingual corpora statistics.

Language	BpeTypes	Task	Length (h)	#utterances
Tag	100	Training	77.28	71,467
		Development	4.81	4,280
		Evaluation	4.99	4,391
Ceb	98	Training	38.13	34,532
		Development	4.87	4,474
		Evaluation	4.66	4,473
Tok	95	Training	35.79	31,167
		Development	4.48	3,879
		Evaluation	4.59	3,923
Hai	101	Training	63.18	47,623
		Development	4.92	4,002
		Evaluation	5.17	4,038
Total	106	Training	214.39	184,789
		Development	19.09	16,635
		Evaluation	19.40	16,825

which were configured with a 4-layer CNN/BLSTM-based encoder network, location-aware attention [12], and a 2-layer LSTM-based decoder network. The cell dimensions of both the encoder and decoder were set to 512. The 40-dimensional mel frequency cepstral coefficient (MFCC) features, the language feature vectors, and the language masks were generated using the Kaldi Toolkit. The LID models were configured with a 6-layer, 2048-dimensional time delay deep neural network (TDNN), and the statistics pooling component was added after the third layer. During the training and decoding processes, the weighted coefficient of the CTC loss function was set to 0.2 and 0.3, respectively. To ensure consistency on both monolingual and multilingual tasks, no additional language model is used during the decoding process.

5.2 Language Identity Information

In this paper, two LID classifiers are implemented with the CE criterion and the MSE criterion to generate the language feature vectors and the language masks, respectively. Before the model training process, the corpus of the four languages was randomly copied to keep the amount of data of four languages balanced. The output distributions of both classifiers are averaged on utterance level information. Figure 3 shows the modeling unit distributions of the utterance-level language masks, where (a) is the real distribution of the evaluation sets and (b) is the estimated distribution of the evaluation sets generated from the MSE-based LID classifier. The estimated language masks are similar to the distributions of the real masks. For the CE-based LID classifier, due to insufficient corpora, the LID accuracy reaches only 85.7% on average. The posterior distributions of the CE-based model are generated as language feature vectors and used to conduct language-conditioned multilingual language-adaptive training.

5.3 E2E Multilingual ASR Experiments

The baseline E2E monolingual model and E2E multilingual model are implemented with subword-level modeling

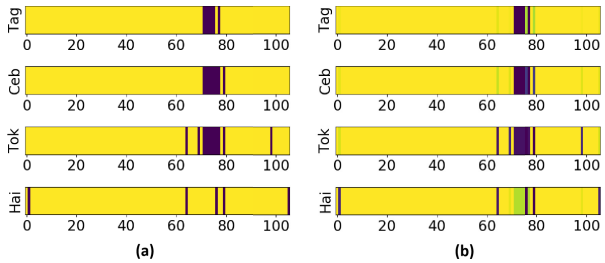

Fig. 3 The language mask distributions of the evaluation sets: (a) the real distributions; (b) the estimated distributions. The sections in bright colors represent the areas of the modeling unit for the four languages.

Table 2 ASR performance of multilingual models with/without language-adaptive training.

Language	Model	Dev		Eval	
		TER%	WER%	TER%	WER%
Tag	Mono	44.9	**55.9**	45.2	55.8
	Multi	46.5	58.3	46.3	57.5
	Enc	46.5	58.6	45.6	57.2
	Dec	45.8	57.5	45.0	56.4
	Mask	45.1	56.8	44.6	55.8
	Enc+Dec	45.7	57.7	44.9	56.4
	Enc+Mask	44.1	**55.9**	43.8	**55.2**
Ceb	Mono	89.6	97.2	87.4	95.9
	Multi	53.9	67.3	52.8	66.9
	Enc	54.9	68.6	52.6	67.0
	Dec	53.8	67.0	52.1	65.9
	Mask	52.6	66.1	51.0	65.3
	Enc+Dec	54.0	**65.9**	51.7	66.1
	Enc+Mask	53.6	66.6	50.8	**64.9**
Tok	Mono	41.1	48.1	47.6	55.2
	Multi	39.9	47.5	47.3	54.6
	Enc	40.3	47.8	46.2	53.5
	Dec	39.6	47.1	46.2	53.9
	Mask	38.4	**45.8**	44.8	52.7
	Enc+Dec	39.2	47.0	45.5	53.3
	Enc+Mask	38.5	45.9	44.5	**51.7**
Hai	Mono	50.9	68.5	58.0	74.4
	Multi	47.8	65.3	53.4	71.3
	Enc	46.8	64.6	52.4	70.1
	Dec	47.1	64.6	52.4	70.3
	Mask	46.4	64.0	51.9	69.7
	Enc+Dec	46.4	63.9	52.5	70.0
	Enc+Mask	45.3	**62.6**	51.5	**68.2**
Avg	Mono	56.6	67.4	59.6	70.3
	Multi	47.0	59.6	50.0	62.6
	Enc	47.1	59.9	49.2	62.0
	Dec	46.6	59.1	48.9	61.6
	Mask	45.6	58.2	48.1	60.6
	Enc+Dec	46.3	58.6	48.7	61.5
	Enc+Mask	45.4	**57.8**	47.7	**60.0**

units generated from the multilingual corpora by the byte-pair-encoding (BPE) algorithm. The size of modeling units (BpeTypes) is shown in Table 1. The monolingual modeling units are obtained by removing the modeling units of the nontarget language from the multilingual modeling units. The results from Table 2 indicate that multilingual joint modeling improves the ASR performance compared to that of the monolingual systems (62.6%/70.3%).

To apply language-adaptive training, the encoder-conditioned method (Enc) and decoder-conditioned (Dec)

method are both implemented for comparison with the proposed language supervision training strategy (Mask). Table 2 shows that all the language-adaptive training methods improve the performance of the multilingual ASR system. In general, the three types of language-adaptive training methods work along different aspects to improve the language discriminability of the attention-based encoder-decoder E2E model. The encoder-conditioned method acts on the front-end acoustic modeling to help distinguish languages by appending language feature vectors to the spectral features. The decoder-conditioned method and the language supervision training strategy both act on the back-end modeling unit prediction. The language supervision training strategy improves the performance by directly weakening the prediction probabilities of the nontarget language modeling units (60.6%/62.6%). It is worth mentioning that the performance of the language supervision training strategy with estimated language masks is quite close to that of the language supervision training strategy with real language masks (60.6%/60.4%). Finally, the encoder-conditioned method is combined with the decoder-conditioned method (61.5%/61.6%) and the language supervision training strategy (60.0%/60.6%), which provide further performance improvement. Notably, even though the language supervision training strategy works independently, it outperforms the combination of the encoder-conditioned method and decoder-conditioned method (60.6%/61.5%).

6. Conclusion

In this paper, a language supervision training strategy is proposed to conduct language-adaptive training under the joint attention/CTC multilingual ASR framework. The encoder-conditioned method and the decoder-conditioned method are also implemented and compared with the proposed method. To simulate the multilingual ASR scenario with unknown language identity information, the CE-based and MSE-based LID classifiers are implemented to generate the language feature vectors and language masks to complement the language-adaptive training. The results show that the combination of the encoder-conditioned method and the language supervision training strategy provides the best performance, achieving an absolute WER reduction of 2.6%/10.3%, on average compared with the multilingual and monolingual baseline systems (60.0%/62.6%/70.3%). In the future, the language adaptive training methods will be investigated with language-specific attentions to model the attention information language-independently.

Acknowledgments

This work is partially supported by the National Key Research and Development Program (Nos. 2019QY1805) and the National Natural Science Foundation of China (Nos. 11590774, 11590770).

References

[1] B. Li, Y. Zhang, T. Sainath, Y. Wu, and W. Chan, "Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes," Proc. ICASSP 2019, Brighton, United Kingdom, pp.5621–5625, May 2019.

[2] 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," Proc. ICASSP 2018, Calgary, AB, Canada, pp.4904–4908, April 2018.

[3] M. Miiller, S. Stiiker, and A. Waibel, "Multilingual adaptation of rnn based asr systems," Proc. ICASSP 2018, Calgary, AB, Canada, pp.5219–5223, April 2018.

[4] A. Kannan, A. Datta, T.N. Sainath, E. Weinstein, B. Ramabhadran, Y. Wu, A. Bapna, Z. Chen, and S. Lee, "Large-Scale Multilingual Speech Recognition with a Streaming End-to-End Model," Proc. Interspeech 2019, Graz, pp.2130–2134, 2019.

[5] C. Shan, C. Weng, G. Wang, D. Su, M. Luo, D. Yu, and L. Xie, "Investigating end-to-end speech recognition for mandarin-english code-switching," Proc. ICASSP 2019, Brighton, United Kingdom, pp.6056–6060, May 2019.

[6] H. Seki, S. Watanabe, T. Hori, J.L. Roux, and J.R. Hershey, "An end-to-end language-tracking speech recognizer for mixed-language speech," Proc. ICASSP 2018, Calgary, AB, Canada, pp.4919–4923, April 2018.

[7] K. Rao, H. Sak, and R. Prabhavalkar, "Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer," Proc. ASRU 2017, Okinawa, Japan, pp.193–199, Dec. 2017.

[8] S. Wiesler, J. Li, and J. Xue, "Investigations on hessian-free optimization for cross-entropy training of deep neural networks," Proc. Interspeech 2013, Lyon, France, pp.3317–3321, Aug. 2013.

[9] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "Espnet: End-to-end speech processing toolkit," Proc. Interspeech 2018, Hyderabad, pp.2207–2211, 2018.

[10] M. Tuchler, A.C. Singer, and R. Koetter, "Minimum mean squared error equalization using a priori information," IEEE Trans. Signal Processing, vol.50, no.3, pp.673–683, March 2002.

[11] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," Proc. ICASSP 2018, Stockholm, Sweden, pp.5329–5333, Aug. 2018.

[12] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," Computer Science, vol.10, no.4, pp.429–439, 2015.