The CCDL System Description for OLR-ASR Tasks of OLR2021

Zhuxin Chen, Duisheng Chen, Xiaorong Huang, Yue Lin

NetEase Games AI Lab, China

{chenzhuxin,gzchenduisheng,huangxiaorong,gzlinyue}@corp.netease.com

Abstract

In this report, we present our CCDL system for the OLR-ASR tasks of OLR2021. For constrained condition, we trained an end-to-end multilingual ASR system for all of the target languages and six hybrid monolingual ASR systems for Mandarin, Cantonese, Shanghainese, Indonesia, Japanese and Korean. Model fusion is based on the confidence of language identification and ASR systems. For unconstrained condition, we additionally trained three end-to-end monolingual systems for Mandarin, Japanese and Indonesian. Finally, we achieved CER of 13.1% in constrained condition on the progress subset.

Index Terms: automatic speech recognition, OLR2021

1. Introduction

The OLR 2021 challenge [1] intends to improve the performance of language recognition systems and speech recognition systems within multilingual scenarios. The challenge this year contains four tasks: (1) constrained LID, (2) unconstrained LID, (3) constrained multilingual ASR, (4) unconstrained multilingual ASR.

In this challenge, we focused on the constrained multilingual ASR task and submitted the final results of task 3 and task 4. Since the test recordings were provided without language tag, we divided it into two sub-tasks. We first determined the language through language identification, and then transcribed the speech using the ASR system of corresponding language.

2. Constrained System

2.1. Data preparation

2.1.1. Training Datasets

For LID system, we utilized all available training datasets, including OLR16-OL7, OLR17-OL3, OLR17-test, OLR17-dev, OLR18-test, OLR19-test, OLR19-dev, OLR20-dialect and OLR20-test [1] .

For ASR system, we only utilized the labeled training datasets, including OLR16-OL7, OLR17-OL3, OLR20-dialect and OLR20-test. Unlike the OLR-ASR baseline [1], for each language, we reserved 100 sentences for validation and others for training. We removed all of punctuation marks and special labels, except the unintelligible speech tag as described in [1].

2.1.2. Data Augmentation

The speed perturbation and spectral augmentation [2] were used in the same way as in the baseline system [1]. In addition, the MUSAN [1] corpus and RIRs [2] datasets were also used to do augmentation.

[1]http://www.openslr.org/17

[2]http://www.openslr.org/28

2.2. E2E Multilingual ASR

Due to the lack of lexicons and text data, building the hybrid ASR system is challenging. Thus, we first trained an end-to-end (E2E) multilingual ASR system by combining all target languages. Our system was based on the OLR-ASR baseline with the following modifications:

- We used conformer [3] as the neural network architecture, which contained 12-layer encoder and 6-layer decoder with 2048-dimensional each layer. And the attention sub-layer was 1024-dimensional with 16 attention heads.
- We used a mixture of characters and sub-words as the output units of model. For the languages of Indonesian, Korean, Russian, Kazakh, Tibetan, Uyghur and Vietnamese, we used sub-words as output units. We used SentencePiece [3] to train tokenizer for each language. The vocab size was set to 500 for Indonesian, Russian, Kazakh, Tibetan, Uyghur and Vietnamese, but 1500 for Korean. For the language of Mandarin, Hokkien, Shanghainese, Sichuanese, Cantonese and Japanese, we used characters as output units directly.
- We performed the model average according to the loss of validation set, not just used the last 10 epochs. The number of epochs was also increased to 40.
- We modified the data preparation as described in session 2.1.

2.3. Hybrid monolingual ASR

Since hybrid HMM-DNN acoustic model is proved to be more promising than conformer-based end-to-end structures in the particular under-resource condition [4], hybrid ASR systems were also employed for model fusion. Due to the limit of lexicons and text data, we only employed the hybrid ASR systems for the languages of Mandarin, Cantonese, Shanghainese, Indonesia, Japanese and Korean. All of systems were trained separately with the same architecture.

For acoustic model, we used a CNN-TDNNF architecture which consists of 4 convolutional layers and 17 factored time delayed neural network layers [5]. The input features were 40-dimensional MFCC features with 3-dimensional pitch features. I-vectors were also used for speaker adaptation. The model was trained with chain model in Kaldi using LF-MMI criterion [6].

For language model, we trained 5-gram for all of systems. We used the crawled text from web to train the language model of Mandarin, Indonesian, Japanese and Korean. For Cantonese and Shanghainese, due to the lack of crawled text, we performed the mixture of Mandarin language model and the specific language model that trained from the corresponding text in the training set.

[3]https://github.com/google/sentencepiece

Table 1: *The results of CER (%) in constrained condition on the progress subset.*

system	Total	zh-cn	Minnan	Shanghai	Sichuan	ct-cn	id-id	ja-jp	ko-kr	ru-ru	vi-vn	Kazak	Tibet	Uyghu
OLR-ASR baseline	39.1	115.8	69.3	35.9	34.4	47.0	8.4	67.0	32.5	34.8	30.3	35.0	52.7	21.0
E2E Multilingual ASR	14.2	17.6	51.8	29.5	26.8	36.9	5.5	34.7	15.2	10.4	5.8	23.2	5	4
Hybrid Monolingual ASR	–	13.8	–	27.9	–	33.5	2.7	22	10.9	–	–	–	–	–
Model Fusion	13.1	13.3	51.9	28.2	27	33.2	2.8	24.7	11.4	10.4	5.8	23.2	5	4

2.4. Language Identification

We used two methods to determine the language tag of recordings. The first method was a language identification classifier based on ResNetSE34 [7, 8]. We used the output of classifier to distinguish languages directly. The second method was based on the edit-distance between the E2E multilingual ASR system and hybrid monolingual ASR system. Considering that the E2E multilingual ASR system mainly suffered the problem of language model due the limit of training data, the edit-distance between with matched hybrid system tended be less than the score of unmatched system. Thus, we combined two methods to distinguish languages.

2.5. Model fusion

The fusion strategy adopted in our system was based on the confidence of language identification and ASR systems. For the recording with high confidence of language identification and hybrid ASR systems, we used the transcription of hybrid ASR system as result. Otherwise, we used the transcription of E2E multilingual ASR system directly.

3. Unconstrained System

3.1. Training Datasets

For unconstrained condition, we trained three E2E ASR systems for Mandarin, Japanese and Indonesian separately. In addition to the training set in constrained condition, we utilized WenetSpeech [4] dataset for Mandarin system, Mozilla Common Voice [5] Japanese and CSJ [6] datasets for Japanese system, Mozilla Common Voice Indonesian and MagicData Indonesian Scripted Speech [7] and Google TTS [8] datasets for Indonesian system.

3.2. E2E Monolingual ASR

We used WeNet [9] toolkit to build the systems. The neural network architecture was based on conformer, which contained 12-layer encoder and 6-layer decoder with 2048-dimensional each layer. The attention sub-layer was 512-dimensional and used 8 attention heads.

4. Results

Table 1 shows the results of CER in constrained condition on the progress subset. As can be seen, our E2E multilingual ASR system significantly outperforms the OLR-ASR baseline, reducing the CER about 64% relatively. We also evaluated our hybrid monolingual ASR systems without the distinction of language, as shown in the third line of Table 1. Finally, the fusion system achieved a CER of 13.1% on the progress subset.

[4]https://github.com/wenet-e2e/WenetSpeech

[5]https://commonvoice.mozilla.org/zh-CN/datasets

[6]https://ccd.ninjal.ac.jp/csj/

[7]https://magichub.com/datasets/indonesian-scripted-speech-corpus-daily-use-sentence

[8]https://github.com/Wikidepia/indonesian_datasets/tree/master/speech/gtts

5. References

[1] B. Wang, W. Hu, J. Li, Y. Zhi, Z. Li, Q. Hong, L. Li, D. Wang, L. Song, and C. Yang, "Olr 2021 challenge: Datasets, rules and baselines," *arXiv preprint arXiv:2107.11113*, 2021.

[2] 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," *arXiv preprint arXiv:1904.08779*, 2019.

[3] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[4] T. Alumäe and J. Kong, "Combining Hybrid and End-to-End Approaches for the OpenASR20 Challenge," in *Proc. Interspeech 2021*, 2021, pp. 4349–4353.

[5] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks." in *Interspeech*, 2018, pp. 3743–3747.

[6] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi." in *Interspeech*, 2016, pp. 2751–2755.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[8] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[9] B. Zhang, D. Wu, Z. Yao, X. Wang, F. Yu, C. Yang, L. Guo, Y. Hu, L. Xie, and X. Lei, "Unified streaming and non-streaming two-pass end-to-end model for speech recognition," *arXiv preprint arXiv:2012.05481*, 2020.