Ant Multilingual Identification System for OLR 2021

Anqi Lyu, Zhiming Wang

Tiansuan Lab, Security BG, Ant Group, Shanghai, China

{lyuanqi.laq, zhiming.wang}@antgroup.com

Abstract

This paper presents a comprehensive description of the Ant multilingual identification system for the 6th Oriental Language Recognition(OLR 2021) Challenge. A conformer-based[1] language identification(LID) model is proposed; it is confirmed that if the LID model is pretrained as an automatic speech recognition(ASR) task to integrate the phonetic information and then further finetuned until its optimum, better performance is achieved in comparison with that from scratch; model's fusion and ensemble is a good strategy to improve performance indicator. In the leaderboard for evaluating the progress subset, our submitted fusion system ranked the top in task 1, and the second in task 2[1].

Index Terms: language identification; language recognition; OLR 2021; conformer; phonetic information

1. Introduction

As globalization is becoming a trend, multilingual communication is common in real scenarios. Multilingual speech technologies are increasingly important and get a lot of attention from research community. Among them, language identification is one of the crucial tasks that aims to predict the language category of the given utterance.

The Oriental Language Recognition(OLR) Challenge is organized annually to encourage in-depth research on multilingual speech domains. The challenge in 2021 [2] covers four tasks: (1) constrained LID on close domain data, only the data provided by the organizer can be used, with the exception of non-speech data; (2) unconstrained LID on wild data, any accessible data (except evaluation data) is allowed; (3) constrained multilingual ASR; (4) unconstrained multilingual ASR. Our team has participated in tasks 1 and 2, which both involve multilingual identification[2].

This paper presents a comprehensive description of the Ant multilingual identification system for OLR 2021 Challenge. The system has the following vital contributions: based on better global and local modeling capability of conformer blocks[1], a conformer-based LID model is proposed; inspired by [3] and [4], the LID model is pretrained as an ASR task to integrate the phonetic information that helps to enhance language identification, and then further finetuned until its optimum; utilize the strategy of model's fusion and ensemble for improved performances.

The subsequent parts are organized as follows: in Section 2, we describe the data information for the first two tasks, especially for the unconstrained task; Section 3 illustrates our multilingual identification system detailly; in Section 4 are experimental settings and results; conclusion is given in Section 5.

[1]http://olrchallenge.org.cn/index.php?p=leaderboard

[2]Our team name is X-Voice, which means to explore the unknown voice world.

2. Data Profile

As described in [2], the challenge organizer provides the following datasets: OLR16-OL7, OLR17-OL3, OLR17-dev, OLR17-test, OLR18-test, OLR19-dev, OLR19-test, OLR20-dialect, and OLR20-test, which contain 17 different languages among them.

2.1. Data for Task 1

For the 1st task, only 13 languages(i.e., Indonesian, Japanese, Russian, Korean, Vietnamese, Mandarin, Cantonese, Sichuanese, Shanghainese, Hokkien, Tibetan, Kazakh and Uyghur) are considered, and the corresponding original audio samples are from the datasets mentioned above.

2.2. Data for Task 2

For task 2, all of the official datasets are used, except for the Cantonese data. Since there is no data constraint for this task, we use extra open-source datasets to strengthen diversities, as are shown in Table 1. Some of them, such as WenetSpeech, are of large size; and out of consideration of data balance, only part of their corresponding data are randomly picked and employed in our experiments.

Table 1: *Extra open-source datasets used in Task 2*

Data Source	Included Languages
VoxLingua107 [5]	13 languages: Indonesian, Japanese, Russian, Korean, Vietnamese, Thai, Malay, Telugu, Hindi, English, Kazakh, Tibetan, Mandarin, Uyghur, Sichuan, Shanghai and Hokkien
OpenSLR [3]	6 languages: Uyghur, Korean, Malay, Telugu, Hindi and Kazakh (that is, SLR22, SLR40, SLR63, SLR66, SLR97, SLR102 and SLR103 are covered)
CommonVoice [4]	7 languages: Indonesian, Hindi, Japanese, Russian, Thai, Uyghur and Vietnamese
Librispeech [6]	English
WenetSpeech [7]	Mandarin

2.3. Data Augmentation

Four strategies for data augmentation are performed to improve model robustness:

- Speed perturbation, with the speed factors of 0.9, 1.0, 1.1;

[3]http://www.openslr.org

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

- Mixture of noise, which are from MUSAN(that is SLR17 [1]), and following the challenge's data protocol, only non-speech noises are used in task 1;
- Reverberation injection, using simulated room impulse responses from SLR28 [1];
- SpecAugment [8], is applied with one frequency mask with maximum frequency mask($F = 10$), and one time mask with maximum time mask($T = 5$).

For both tasks, data are randomly divided into training and development sets. With the help of given speaker information, there is no speaker overlap in the split for a better cross-validation indicator of model's performance [3]. In addition, data with transcriptions in the two separate datasets are further selected to train and evaluate the ASR system.

Except for the differences in data constitution, both tasks of 1 and 2 share the same system framework as discussed in Section 3.

3. System Description

3.1. Input Features

Audios are 16,000 Hz. 80-dimensional logarithm mel filter banks are generated within a 25ms sliding window using a hop step size of 10ms; cepstral mean normalization(CMN) is performed within a 3-second sliding window. No voice activity detection(VAD) is used.

3.2. Backbone Model

The training procedure is divided into two stages: first, an end-to-end ASR encoder-decoder model is pretrained to integrate lexical phonetic information; then with the shared encoder components of ASR model, the LID encoder is further finetuned to achieve the optimal.

3.2.1. Pretrained ASR encoder-decoder networks

The ASR encoder-decoder networks are the conformer architecture[1], consisting of subsample layers(factor is 4), 12-block conformer as encoder, 3-block transformer as decoder. They are trained end-to-end with joint CTC and attention loss using the Wenet toolkit [9].

For each language, a sentencepiece [10] model is trained to define modeling units; then all of them from different languages are combined together to constitute an overall dictionary. For most of the languages, the number of tokens is less than 2000, while for some languages such as Mandarin, more than 2000 tokens is allowed due to their large and various character space.

3.2.2. Finetuned LID encoder

At the finetuning stage, with the shared conformer encoder of ASR networks, an attentive statistical pooling layer [11] is used to map frame level representations into segment level output vector, which is then projected by a linear layer with batch normalization and nonlinear activation, to the utterance level embedding. That is the LID encoder, and following that is a language softmax classifier. The model is finetuned with cross entropy(CE) loss. Other loss function like AAM-Softmax [12] is an alternative.

3.3. Scoring Methods

The embeddings are computed for original training data and augmented audio samples by the front three strategies in 2.3. For more discriminative embedding features, the dimension of the embeddings is reduced to a fixed size with Linear Discriminative Analysis(LDA) method; we find the number of language categories minus 1 is the best choice for LDA dimensional size, maybe attributed to the orthogonal lower dimension subspace. The dimensional reduction embedding features are averaged into one enrollment embedding vector for each language.

We have experimented both cosine similarity and logistic regression(LR) as back-end scoring methods. Our experiments indicated that cosine similarity is superior to LR; hence it is chosen to compute the score of each trial. As for score normalization, simple min-max normalization as in Eq.(1) is performed for each score vector x, which consists of scores for one test segment with all enrollment languages, independently.

$$x^{'} = \frac{x - min_i(x_i)}{max_i(x_i) - min_i(x_i)} \tag{1}$$

4. Experimental Settings and Results

4.1. Experimental Settings

For the shared 12-block conformer encoder for ASR and LID models, output dimension is 256, linear dimension is 2048, and number of attention heads is 4. And for LID model, the attentive pooling layer is 1536-dimensional multilayer perceptron(MLP), and the projection layer has 256 hidden units.

Both ASR and LID models are trained with ADAM optimizer. The learning rate schedule follows that from transformer[13] with 25000 warm up steps, and the peak learning rate 1e-3 for ASR pretraining, 1e-4 for LID finetuning. For LID training task, as in [14], randomly selected T frames of audio segments are taken as a mini-batch at each iteration; to be specific, $T \sim Uniform(200, 400)$ in general unless otherwise specified.

As in [2], we adopt average cost function(C_{avg}) as primary metric, and equal error rate(EER) as auxiliary one; smaller values of them correspond to better performances.

4.2. Main Results

4.2.1. Analysis of a single model

For fair comparison, the baseline E-TDNN x-vector model released by the organizer [5] and our proposed conformer-based LID model are trained from scratch, and their performances are reported in Table 2. As is shown here, a conformer-based LID model trained from scratch outperforms the baseline by a large margin, which means conformer architecture is competitive in language recognition task. It is believed that conformer block is good at global modeling capability by attention mechanism, and benefits from local convolution.

When pretrained as an ASR task, the performance of conformer-based LID model is further improved, which confirms the assumption that the phonetic information integrated by ASR task helps to enhance language identification.

4.2.2. Fusion and Ensemble

For a single system, the weights of the top $K(K = 5)$ checkpoints with lower validation losses on the development datasets are averaged into those of one model, which avoids local fluctuation and brings in more better generalization. For multiple

[5] https://github.com/Snowdar/asv-subtools/tree/master/recipe/olr2021-baseline

Table 2: *Performances of conformer-based LID model in Task 1*

Models	Dev	
	C_{avg}	EER(%)
baseline	0.0608	5.8030
Conformer moldel from scratch	0.0165	1.6620
Conformer model with ASR pretraining	0.0030	0.3015

systems, the scores from different encoder networks are linearly weighted into a regression value, and the optimal weights are tuned on the development datasets with grid search method. We observe that, with some techniques, ensemble model always leads to better performance for mutual complementation, which is also our final submission to the challenge.

To be specific, for task 1, four different conformer-based LID models are selected: (A)trained with cross entropy loss; (B)trained with AAM-Softmax [12] loss; (C)based on A, with additive babble noises from training datasets; (D)based on A, $T \sim Uniform(300, 400)$. Their independent evaluations and fusion performance on the development datasets are shown in Table 3. In the leaderboard for evaluating the progress subset, our submitted fusion system achieved C_{avg} 0.0062 and EER 0.6512 %, ranking the top.

Table 3: *Model's independent and fusion evaluations in Task 1*

Models	Fusion weight	Dev	
		C_{avg}	EER(%)
A	0.10	0.0026	0.2546
B	0.45	0.0024	0.2479
C	0.15	0.0021	0.2211
D	0.30	0.0021	0.2144
fusion	-	0.0018	0.1809

For task 2, two different conformer-based LID models are chosen based on the above A and B methods respectively. Their independent and fusion evaluations are reported in Table 4. In the leaderboard for evaluating the progress subset, our submitted fusion system achieved C_{avg} 0.0166 and EER 1.85 %, ranking the second.

Table 4: *Model's independent and fusion evaluations in Task 2*

Models	Fusion weight	Dev	
		C_{avg}	EER(%)
A	0.4	0.0058	0.8298
B	0.6	0.0058	0.7587
fusion	-	0.0053	0.7208

5. Conclusion

In this paper, we describe the Ant multilingual identification system for OLR 2021 challenge. A conformer-based LID model is proposed and it is pretrained as an ASR task to integrate the phonetic information that improves language identification; model's fusion and ensemble results in better performance. In the leaderboard for evaluating the progress subset, our submitted fusion system ranked the first in task 1, and the second in task 2.

6. References

[1] 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.

[2] 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.

[3] R. Duroselle, M. Sahidullah, D. Jouvet, and I. Illina, "Language recognition on unknown conditions: the loria-inria-multispeech system for ap20-olr challenge," 2021.

[4] D. Wang, S. Ye, and X. Hu, "The royal flush system for ap20-olr challenge," http://cslt.riit.tsinghua.edu.cn/mediawiki/images/0/06/Royal-Flush_system_description.pdf.

[5] J. Valk and T. Alumäe, "Voxlingua107: a dataset for spoken language recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 652–658.

[6] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[7] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. Zeng *et al.*, "Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition," *arXiv preprint arXiv:2110.03370*, 2021.

[8] 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.

[9] B. Zhang, D. Wu, C. Yang, X. Chen, Z. Peng, X. Wang, Z. Yao, X. Wang, F. Yu, L. Xie *et al.*, "Wenet: Production first and production ready end-to-end speech recognition toolkit," *arXiv preprint arXiv:2102.01547*, 2021.

[10] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," *arXiv preprint arXiv:1808.06226*, 2018.

[11] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *arXiv preprint arXiv:1803.10963*, 2018.

[12] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[14] Z. Wang, F. Xu, K. Yao, Y. Cheng, T. Xiong, and H. Zhu, "Antvoice neural speaker embedding system for ffsvc 2020," *Proc. Interspeech 2021*, pp. 1069–1073, 2021.