AP19-OLR Challenge: Three Tasks and Their Baselines

Zhiyuan Tang[†], Dong Wang[†‡*] and Liming Song[§]
[†] Center for Speech and Language Technologies, Tsinghua University
[‡] Beijing National Research Center for Information Science and Technology
[§] SpeechOcean
Corresponding email: wangdong99@mails.tsinghua.edu.cn

***Abstract*—This paper introduces the fourth oriental language recognition (OLR) challenge AP19-OLR, including the data profile, the tasks and the evaluation principles. The OLR challenge has been held successfully for three consecutive years, along with APSIPA Annual Summit and Conference (APSIPA ASC). The challenge this year still focuses on practical and challenging tasks, precisely (1) short-utterance LID, (2) cross-channel LID and (3) zero-resource LID.**

The event this year includes more languages and more real-life data provided by SpeechOcean and the NSFC M2ASR project. All the data is free for participants. Recipes for x-vector system and back-end evaluation are also conducted as baselines for the three tasks. The participants can refer to these online-published recipes to deploy LID systems for convenience. We report the baseline results on the three tasks and demonstrate that the three tasks are worth some efforts to achieve better performance.

I. Introduction

There are thousands of languages around the world grouping many language families, such as the oriental language families which often include Austroasiatic languages (e.g.,Vietnamese, Cambodia) [1], Tai-Kadai languages (e.g., Thai, Lao), Hmong-Mien languages (e.g., some dialects in south China), Sino-Tibetan languages (e.g., Chinese Mandarin), Altaic languages (e.g., Korea, Japanese) and Indo-European languages (e.g., Russian) [2], [3], [4]. With so many languages, the development of communication technology and movement of world-wide population make multilingual phenomena more and more common, and in turn, more advanced speech technologies have been developed to further boost the communication in multilingual environment, e.g., instant and simultaneous interpretation with machine.

The language identification (LID) technology plays a great role in the development of multilingual interaction between human and machine, and it is often located at the front end of other speech processing systems, mostly speech recognition (ASR). To better meet the needs of multilingual ASR, the building of LID system may encounter many difficult issues, such as high real-time requirement, cross-channel speech signals and very noisy background.

Considering the languages for which we build the LID system, there may also exist a huge difference in linguistic resources between two different languages, such as expert knowledge about the language and amounts of digital resources for machine learning. Some languages spoken by decreasing number of population may even face the risk of extinction. That requires better language technologies to process these low-resource or even zero-resource languages, including spoken language identification technology. Different languages also interact and influence each other, leading to complicated linguistic evolution and lots of research [5], [6], [7].

The oriental language recognition (OLR) challenge is organized annually, aiming at improving the research on multilingual phenomena and advancing the development of language recognition technologies. The challenge has been conducted three times since 2016, namely AP16-OLR [8], AP17-OLR [9] and AP18-OLR [10], each attracting dozens of teams around the world.

AP18-OLR involved 10 languages and focused on three challenging tasks: (1) short-utterance (1 second) LID, which was inherited from AP17-OLR; (2) LID for confusing language pairs; (3) open-set LID where the test data involved unknown interference languages. In the first task, the system submitted by the XMUspeech team achieved the best performance (C_{avg}=0.0462, EER%=4.59). In the second and third tasks, the systems submitted by the NetEase AI-Speech team achieved the best performance with C_{avg}=0.0032, EER%=0.33 and C_{avg}=0.0119, EER%=3.16 respectively. From these results, one can see that for the short-utterance condition, the task remains challenging. More details about the past three challenges can be found on the challenge website.[1]

Based on the experience of the last three challenges and the calling from industrial application, we propose the fourth OLR challenge. This new challenge, denoted by AP19-OLR, will be hosted by APSIPA ASC 2019. It involves more languages and focuses on more practical and challenging tasks: (1) short-utterance (1 second) LID, as in the past two challenges, (2) cross-channel LID, which reveals the real-life demand of speech technology such as machine interpretation, and (3) zero-resource LID, where no resources are provided for training before inference, but only several utterances of each language are provided for language reference.

In the rest of the paper, we will present the data profile and the evaluation plan of the AP19-OLR challenge. To assist participants to build their own submissions, baseline recipes are constructed based on the x-vector system. The Kaldi recipes of these baselines can be downloaded from the challenge website.

[1] http://olr.cslt.org

TABLE I
AP16-OL7 AND AP17-OL3 DATA PROFILE

AP16-OL7			AP16-OL7-train/dev			AP16-OL7-test		
Code	Description	Channel	No. of Speakers	Utt./Spk.	Total Utt.	No. of Speakers	Utt./Spk.	Total Utt.
ct-cn	Cantonese in China Mainland and Hongkong	Mobile	24	320	7559	6	300	1800
zh-cn	Mandarin in China	Mobile	24	300	7198	6	300	1800
id-id	Indonesian in Indonesia	Mobile	24	320	7671	6	300	1800
ja-jp	Japanese in Japan	Mobile	24	320	7662	6	300	1800
ru-ru	Russian in Russia	Mobile	24	300	7190	6	300	1800
ko-kr	Korean in Korea	Mobile	24	300	7196	6	300	1800
vi-vn	Vietnamese in Vietnam	Mobile	24	300	7200	6	300	1800
AP17-OL3			AP17-OL3-train/dev			AP17-OL3-test		
Code	Description	Channel	No. of Speakers	Utt./Spk.	Total Utt.	No. of Speakers	Utt./Spk.	Total Utt.
ka-cn	Kazakh in China	Mobile	86	50	4200	86	20	1800
ti-cn	Tibetan in China	Mobile	34	330	11100	34	50	1800
uy-id	Uyghur in China	Mobile	353	20	5800	353	5	1800

Male and Female speakers are balanced.
The number of total utterances might be slightly smaller than expected, due to the quality check.

II. DATABASE PROFILE

Participants of AP19-OLR can request the following datasets for system construction. All these data can be used to train their submission systems.

- AP16-OL7: The standard database for AP16-OLR, including AP16-OL7-train, AP16-OL7-dev and AP16-OL7-test.
- AP17-OL3: A dataset provided by the M2ASR project, involving three new languages. It contains AP17-OL3-train and AP17-OL3-dev.
- AP17-OLR-test: The standard test set for AP17-OLR. It contains AP17-OL7-test and AP17-OL3-test.
- AP18-OLR-test: The standard test set for AP18-OLR. It contains AP18-OL7-test and AP18-OL3-test.
- THCHS30: The THCHS30 database (plus the accompanied resources) published by CSLT, Tsinghua University [11].

Besides the speech signals, the AP16-OL7 and AP17-OL3 databases also provide lexicons of all the 10 languages, as well as the transcriptions of all the training utterances. These resources allow training acoustic-based or phonetic-based language recognition systems. Training phone-based speech recognition systems is also possible, though large vocabulary recognition systems are not well supported, due to the lack of large-scale language models.

A test dataset AP19-OLR-test will be provided at the date of result submission, which includes three parts corresponding to the three LID tasks.

A. AP16-OL7

The AP16-OL7 database was originally created by Speechocean, targeting for various speech processing tasks. It was provided as the standard training and test data in AP16-OLR. The entire database involves 7 datasets, each in a particular language. The seven languages are: Mandarin, Cantonese, Indonesian, Japanese, Russian, Korean and Vietnamese. The data volume for each language is about 10 hours of speech signals recorded in reading style. The signals were recorded by mobile phones, with a sampling rate of 16 kHz and a sample size of 16 bits.

For Mandarin, Cantonese, Vietnamese and Indonesia, the recording was conducted in a quiet environment. As for Russian, Korean and Japanese, there are 2 recording sessions for each speaker: the first session was recorded in a quiet environment and the second was recorded in a noisy environment. The basic information of the AP16-OL7 database is presented in Table I, and the details of the database can be found in the challenge website or the description paper [8].

B. AP17-OL7-test

The AP17-OL7 database is a dataset provided by SpeechOcean. This dataset contains 7 languages as in AP16-OL7, each containing 1800 utterances. The recording conditions are the same as AP16-OL7. This database is used as part of the test set for the AP17-OLR challenge.

C. AP17-OL3

The AP17-OL3 database contains 3 languages: Kazakh, Tibetan and Uyghur, all are minority languages in China. This database is part of the Multilingual Minorlingual Automatic Speech Recognition (M2ASR) project, which is supported by the National Natural Science Foundation of China (NSFC). The project is a three-party collaboration, including Tsinghua University, the Northwest National University, and Xinjiang University [12]. The aim of this project is to construct speech recognition systems for five minor languages in China (Kazakh, Kirgiz, Mongolia, Tibetan and Uyghur). However, our ambition is beyond that scope: we hope to construct a full set of linguistic and speech resources and tools for the five languages, and make them open and free for research purposes. We call this the M2ASR Free Data Program. All the data resources, including the tools published in this paper, are released on the web site of the project.[2]

The sentences of each language in AP17-OL3 are randomly selected from the original M2ASR corpus. The data volume for each language in AP17-OL3 is about 10 hours of speech signals recorded in reading style. The signals were recorded by mobile phones, with a sampling rate of 16 kHz and a sample size of 16 bits. We selected 1800 utterances for each language as the development set (AP17-OL3-dev), and the rest is used as the training set (AP17-OL3-train). The test set of each language involves 1800 utterances, and is provided separately and

[2]http://m2asr.cslt.org

denoted by AP17-OL3-test. Compared to AP16-OL7, AP17-OL3 contains much more variations in terms of recording conditions and the number of speakers, which may inevitably increase the difficulty of the challenge task. The information of the AP17-OL3 database is summarized in Table I.

D. AP18-OLR-test

The AP18-OLR-test database is the standard test set for AP18-OLR, which contains AP18-OL7-test and AP18-OL3-test. Like the AP17-OL7-test database, AP18-OL7-test contains the same target 7 languages, each containing 1800 utterances, while AP18-OL7-test also contains utterances from several interference languages. The recording conditions are the same as AP17-OL7-test. Like the AP17-OL3-test database, AP18-OL3-test contains the same 3 languages, each containing 1800 utterances. The recording conditions are also the same as AP17-OL7-test.

E. AP19-OLR-test

The AP19-OLR-test database is the standard test set for AP19-OLR, which includes 3 parts responding to the 3 LID tasks respectively, precisely AP19-OLR-short, AP19-OLR-channel and AP19-OLR-zero.

- AP19-OLR-short: This subset is designed for the short-utterance LID task, which contains the ten target languages as in AP18-OLR-test and each language has 1800 utterances.
- AP19-OLR-channel: This subset is designed for the cross-channel LID task, which contains six of the ten target languages as in AP18-OLR-test, but was recorded in wild environment. The six languages are Tibetan, Uyghur, Japanese, Russian, Vietnamese and Mandarin. Each language has 1800 utterances.
- AP19-OLR-zero: This subset is designed for the zero-resource LID task. The three languages are not in the ten traditional languages, but other resource-limited languages, namely Catalan, Greek and Telugu. Each language has 10 utterances for reference and 1800 for identification test.

To help the participants develop systems against the three tasks, development set AP19-OLR-dev is also provided. Specifically, for task 1, the short-utterance test set from AP18-OLR-test can be reused. For task 2 and 3, a new smaller development set is provided respectively, while the three target languages in the third development set are different from those in the final test set.

III. AP19-OLR CHALLENGE

The evaluation plan of AP19-OLR keeps mostly the same as in AP18-OLR, except some modification for the new challenge tasks.

Following the definition of NIST LRE15 [13], the task of the LID challenge is defined as follows: Given a segment of speech and a language hypothesis (i.e., a target language of interest to be detected), the task is to decide whether that target language was in fact spoken in the given segment (yes or no), based on an automated analysis of the data contained in the segment. The evaluation plan mostly follows the principles of NIST LRE15.

The AP19-OLR challenge includes three tasks as follows:

- Task 1: Short-utterance LID is a close-set identification task, which means the language of each utterance is among the known traditional 10 target languages. The utterances are as short as 1 second.
- Task 2: Cross-channel LID, where test data in different channels for the known 10 target languages will be provided.
- Task 3: Zero-resource LID, where no resources are provided for training before inference, but several reference utterances are provided for each language.

A. System input/output

The input to the LID system is a set of speech segments in unknown languages. For task 1 and task 2, those speech segments are within the 10 known target languages. For task 3, the target languages of the speech segments are the same as the reference utterances. The task of the LID system is to determine the confidence that a language is contained in a speech segment. More specifically, for each speech segment, the LID system outputs a score vector $< \ell_1, \ell_2, ..., \ell_{10} >$, where ℓ_i represents the confidence that language i is spoken in the speech segment. The scores should be comparable across languages and segments. This is consistent with the principles of LRE15, but differs from that of LRE09 [14] where an explicit decision is required for each trial.

In summary, the output of an OLR submission will be a text file, where each line contains a speech segment plus a score vector for this segment, e.g.,

	lang_1	lang_2	...	lang_9	lang_{10}
seg_1	0.5	-0.2	...	-0.3	0.1
seg_2	-0.1	-0.3	...	0.5	0.3
...			...		

B. Test condition

- No additional training materials. The only resources that are allowed to use are: AP16-OL7, AP17-OL3, AP17-OLR-test, AP18-OLR-test, and THCHS30.
- All the trials should be processed. Scores of lost trials will be interpreted as -inf.
- The speech segments in each task should be processed independently, and each test segment in a group should be processed independently too. Knowledge from other test segments is not allowed to use (e.g., score distribution of all the test segments).
- Information of speakers is not allowed to use.
- Listening to any speech segments is not allowed.

C. Evaluation metrics

As in LRE15, the AP19-OLR challenge chooses C_{avg} as the principle evaluation metric. First define the pair-wise loss that composes the missing and false alarm probabilities for a particular target/non-target language pair:

$$C(L_t, L_n) = P_{Target} P_{Miss}(L_t) + (1 - P_{Target}) P_{FA}(L_t, L_n)$$

where L_t and L_n are the target and non-target languages, respectively; P_{Miss} and P_{FA} are the missing and false alarm

probabilities, respectively. P_{target} is the prior probability for the target language, which is set to 0.5 in the evaluation. Then the principle metric C_{avg} is defined as the average of the above pair-wise performance:

$$C_{avg} = \frac{1}{N} \sum_{L_t} \left\{ \begin{array}{l} P_{Target} \cdot P_{Miss}(L_t) \\ + \sum_{L_n} P_{Non-Target} \cdot P_{FA}(L_t, L_n) \end{array} \right\}$$

where N is the number of languages, and $P_{Non-Target} = (1-P_{Target})/(N-1)$. We have provided the evaluation script for system development.

IV. Baseline systems

We construct the baseline systems for the three tasks respectively. All the experiments are conducted with Kaldi [15]. The purpose of these experiments is to present a reference for the participants, rather than a competitive submission. The recipes can be downloaded from the website of the challenge.

A. X-vector system

We use the x-vector system as described in [16], [17]. The raw feature of the system is 40-dimensional filterbanks. The energy VAD is used to filter out nonspeech frames. The network configuration is outlined in Table II as shown in [16]. The DNN is trained to classify the N languages in the training data. After training, embeddings called 'x-vectors' are extracted from the affine component of layer $segment6$. Excluding the $softmax$ output layer and $segment7$ there is a total of 4.2 million parameters.

TABLE II
THE EMBEDDING DNN ARCHITECTURE. X-VECTORS ARE EXTRACTED AT LAYER $segment6$, BEFORE THE NONLINEARITY. THE N IN THE SOFTMAX LAYER CORRESPONDS TO THE NUMBER OF TRAINING LANGUAGES.

Layer	Layer context	Total context	Input × output
frame1	$[t-2, t+2]$	5	200×512
frame2	$\{t-2, t, t+2\}$	9	1536×512
frame3	$\{t-3, t, t+3\}$	15	1536×512
frame4	$\{t\}$	15	512×512
frame5	$\{t\}$	15	512×1500
stats pooling	$[0, T)$	T	$1500T \times 3000$
segment6	$\{0\}$	T	3000×512
segment7	$\{0\}$	T	512×512
softmax	$\{0\}$	T	$512 \times N$

We train the x-vector system with a combined dataset including AP16-OL7, AP17-OL3 and AP17-OLR-test, and the target number of the system refers to the number of all languages, i.e. 10. This basic system is used for all three tasks with different back-end evaluation, either producing scores directly from the output of the original system for different target languages (task 1 and 2), or extracting x-vectors for each utterance for later identification (task 3).

B. Performance results

The primary evaluation metric in AP19-OLR is C_{avg}. Besides that, we also present the performance in terms of equal error rate (EER). These metrics evaluate system performance from different perspectives, offering a whole picture of the capability of the tested system. The performance is evaluated on both the AP19-OLR-dev and AP19-OLR-test databases. Table III shows the utterance-level C_{avg} and EER results for the three tasks respectively. For task 1, we choose the short-utterance subset of AP18-OLR-test to be the development set.

TABLE III
C_{avg} AND EER RESULTS OF THREE TASKS

	Dev set		Test set	
Task	C_{avg}	EER%	C_{avg}	EER%
short-utterance	0.1271	12.37	0.1257	12.22
cross-channel	0.3868	43.13	0.3720	38.44
zero-resource	0.3393	34.47	0.2027	21.94

1) Short-utterance LID: The first task identifies short-duration utterances. The test set is AP19-OLR-short which contains candidate speech segments with 1 second duration. As the languages in AP19-OLR-short are the same as in the training set of our x-vector system, the output of the original system by propagating the test set can be seen as the confidence that each speech segment is belonging to a specific language. From the 'short-utterance' results in Table III, we find that short-duration utterances are hard to recognize.

2) Cross-channel LID: The second task identifies six languages which are also included in the training set of above x-vector system, so the scores referring to those six languages for each utterance can also be produced as task 1 does. Cross-channel speech signals are much more difficult for the baseline system to recognize, that can be seen from the 'cross-channel' results in Table III.

3) Zero-resource LID: The evaluation process for task 3 can be seen as identifying languages based on a reference dataset where the languages may never be seen before. First we extract x-vectors for each segment in the reference set, and then accumulate the utterance-level x-vectors for each language to produce language-level x-vectors. Each language-level x-vector can represent that specific language. We also extract x-vectors for each segment in the test set. Finally, we compare the utterance-level x-vectors from the test set to the language-level x-vectors from the reference set respectively, then decide which language the test segments belong to. The metric used for the comparison in this paper is 'cosine' distance. From the 'zero-resource' results in Table III, it can be seen that resource-limited LID keeps challenging. The difference of target languages between development set and test set results in the gap of the performance.

V. Conclusions

We presented the data profile, task definitions and evaluation principles of the AP19-OLR challenge. To assist participants to construct a reasonable starting system, we published baseline system based on the x-vector model. We showed that the tasks defined by AP19-OLR are rather challenging and are worthy of careful study. All the data resources are free for the participants, and the recipes of the baseline systems can be freely downloaded.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Projects 61633013.

REFERENCES

[1] P. Sidwell and R. Blench, "14 the austroasiatic urheimat: the southeastern riverine hypothesis," *Dynamics of human diversity*, p. 315, 2011.

[2] S. R. Ramsey, *The languages of China*. Princeton University Press, 1987.

[3] M. Shibatani, *The languages of Japan*. Cambridge University Press, 1990.

[4] B. Comrie, G. Stone, and M. Polinsky, *The Russian language in the twentieth century*. Oxford University Press, 1996.

[5] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7304–7308.

[6] R. Fér, P. Matějka, F. Grézl, O. Plchot, and J. Černockỳ, "Multilingual bottleneck features for language recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[7] D. Wang and T. F. Zheng, "Transfer learning for speech and language processing," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2015, pp. 1225–1237.

[8] D. Wang, L. Li, D. Tang, and Q. Chen, "Ap16-ol7: A multilingual database for oriental languages and a language recognition baseline," in *APSIPA ASC*. IEEE, 2016.

[9] Z. Tang, D. Wang, Y. Chen, and Q. Chen, "Ap17-OLR challenge: Data, plan, and baseline," in *APSIPA ASC*. IEEE, 2016.

[10] Z. Tang, D. Wang, and Q. Chen, "Ap18-olr challenge: Three tasks and their baselines," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 596–600.

[11] D. Wang and X. Zhang, "THCHS-30: A free chinese speech corpus," *arXiv preprint arXiv:1512.01882*, 2015.

[12] D. Wang, T. F. Zheng, Z. Tang, Y. Shi, L. Li, S. Zhang, H. Yu, G. Li, S. Xu, A. Hamdulla *et al.*, "M2asr: Ambitions and first year progress," in *OCOCOSDA*, 2017.

[13] "The 2015 NIST language recognition evaluation plan (LRE15)," NIST, 2015, ver. 22-3.

[14] "The 2009 NIST language recognition evaluation plan (LRE09)," NIST, 2009, ver. 6.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, and P. Schwarz, "The kaldi speech recognition toolkit," in *Proceedings of IEEE 2011 workshop on Automatic Speech Recognition and Understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[16] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[17] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 105–111.