AP18-OLR Challenge: Three Tasks and Their Baselines

Zhiyuan Tang[†‡], Dong Wang[†‡*] and Qing Chen[§]
[†] Center for Speech and Language Technologies, Tsinghua University
[‡] Beijing National Research Center for Information Science and Technology
[§] SpeechOcean

***Abstract*—The third oriental language recognition (OLR) challenge AP18-OLR is introduced in this paper, including the data profile, the tasks and the evaluation principles. Following the events in the last two years, namely AP16-OLR and AP17-OLR, the challenge this year focuses on more challenging tasks, including (1) short-duration utterances, (2) confusing languages, and (3) open-set recognition.**

The same as the previous events, the data of AP18-OLR is also provided by SpeechOcean and the NSFC M2ASR project. Baselines based on both the i-vector model and neural networks are constructed for the participants' reference. We report the baseline results on the three tasks and demonstrate that the three tasks are truly challenging. All the data is free for participants, and the Kaldi recipes for the baselines have been published online.

I. Introduction

Oriental languages can be grouped into several language families, such as 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 the worldwide population movement and communication, more and more multilingual phenomena are clear, e.g., code switching between languages in an utterance where the secondary languages may just appear as a single word. The oriental languages themselves also influence each other via the multilingual interaction, leading to complicated linguistic evolution. This complicated multilingual phenomena attracted lots of research recently [5], [6], [7].

To consistently boost the research on multilingual phenomena, the center for speech and language technologies (CSLT) at Tsinghua University and SpeechOcean organize the oriental language recognition (OLR) challenge annually, aiming at developing advanced language identification (LID) techniques. The challenge has been conducted two times since 2016, namely AP16-OLR [8] and AP17-OLR [9]. They were very successful, especially AP17-OLR, in which 31 teams from 5 countries participated.

AP17-OLR defined three test conditions according to the duration of the test utterances: 1-second condition, 3-second condition and full-utterance condition. For the full-utterance condition, the system submitted by the NUS-I2R-NTU team achieved the best performance (C_{avg}=0.0034, EER%=0.37), and for the 1-second condition, the team SASI got the best performance (C_{avg}=0.0765, EER%=7.91). From these results, one can see that LID on long utterances have been solved to a large extent, however for the short-utterance condition, the task remains challenging. This is essentially the base when we designed the OLR tasks this year. More details about the past two challenges can be found on the challenge website.[1]

Based on the experience of the last two years, we propose the third OLR challenge. This new challenge, denoted by AP18-OLR, will be hosted by APSIPA ASC 2018. It involves the same 10 languages as in AP17-OLR, but focuses on more challenging tasks: (1) short-duration utterances (1 second) LID, which inherits from AP17-OLR; (2) LID for confusing language pairs; (3) open-set LID, where the test data involves unknown interference languages.

In the rest of the paper, we will present the data profile and the evaluation plan of the AP18-OLR challenge. To assist participants to build their own submissions, two types of baseline systems are constructed, based on the i-vector model and various DNN models respectively. The Kaldi recipes of these baselines can be downloaded from the challenge web site.

II. Database profile

Participants of AP18-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.
- THCHS30: The THCHS30 database (plus the accompanied resources) published by CSLT, Tsinghua University [10].

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 will be provided at the date of result submission. This test set involves two parts: AP18-OL7-test and

[1] http://www.olrchallenge.org

TABLE I
AP17-OL3 AND AP16-OL7 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.

AP18-OL3-test. The former involves utterances from the 7 target languages of AP16-OL7, but also 8 unknown interference languages. The details of these databases are described as follows.

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), 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 [11]. 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 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.

III. AP18-OLR CHALLENGE

The evaluation plan of AP18-OLR keeps mostly the same as in AP16-OLR and AP17-OLR, except some modification for the new challenge tasks.

Following the definition of NIST LRE15 [12], the task of the LID challenge is defined as follows: Given a segment of

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

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 AP18-OLR challenge includes three tasks as follows:

- Task 1: Short-utterance identification task: This is a close-set identification task, which means the language of each utterance is among the known 10 target languages. The utterances are as short as 1 second.
- Task 2: Confusing-language identification task: This task identifies the language of utterances from 3 highly confusing languages (Cantonese, Korean and Mandarin).
- Task 3: Open-set recognition task: In this task, the test utterance may be in none of the 10 target languages.

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, while for task 3, the speech segment may be a non-target language. 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 [13] 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 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 AP18-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:

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

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 constructed two kinds of baseline LID systems, based on the i-vector model and various DNN models respectively. All the experiments were conducted with Kaldi [14]. 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 web page of the challenge.

A. i-vector system

The i-vector baseline systems were constructed based on the i-vector model [15], [16]. The static acoustic features involved 19-dimensional Mel frequency cepstral coefficients (MFCCs) and the log energy. This static features were augmented by their first and second order derivatives, resulting in 60-dimensional feature vectors. The UBM involved $2,048$ Gaussian components and the dimensionality of the i-vectors was 400. Linear discriminative analysis (LDA) was employed to promote language-related information. The dimensionality of the LDA projection space was set to 150.

With the i-vectors (either original or after LDA/PLDA transform), the score of a trail on a particular language can be simply computed as the cosine distance between the test i-vector and the mean i-vector of the training segments that belong to that language. This is denoted to be 'cosine distance scoring'.

B. DNN systems

For the DNN baseline, two kinds of DNN architectures were designed, namely the traditional time-delay neural network (TDNN) [17] and recurrent neural network with long short-term memory units (LSTM-RNN) [18].

The raw feature of all the two DNN systems is 40-dimensional Fbanks, with a symmetric 2-frame window to splice neighboring frames. For the TDNN LID, there are 6 hidden layers, and the activation function is rectified linear unit (ReLU). The number of units of each TDNN layer is set to be 650. The number of cells of the LSTM is set to be 512.

C. Performance results

The primary evaluation metric in AP18-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 verification/identification capability of the tested system. At present, the performance is evaluated on the development set which is actually the AP17-OLR-test database. We present the utterance-level C$_{avg}$ and EER results for the three tasks respectively.

1) Short-utterance LID: The first task identifies short-duration utterances. AP17-OLR-test contains three subset sets with different durations (1 second, 3 second and regular length). Besides the performance of the baseline systems on the 1 second condition, we also report the performance on the regular length for reference. The results of i-vector and DNN systems are showed in Table II. From the results, we find that short-duration utterances are hard to recognize for both the i-vector system and the DNN systems, while DNN systems show more robustness.

TABLE II
C$_{avg}$ and EER results on 1 second and full-length conditions

	1 second		Full-Length	
System	C_{avg}	EER%	C_{avg}	EER%
i-vector	0.1888	18.75	0.0578	5.92
i-vector + LDA	0.1784	18.04	0.0598	6.12
i-vector + PLDA	0.1746	17.51	0.0596	5.86
TDNN	0.1282	14.04	0.1034	11.31
LSTM	0.1452	15.92	0.1154	12.76

2) Confusing-language LID: The second task focuses on languages that are hard to distinguish (confusing languages). To find the most indistinguishable languages among the 10 target ones, we investigated all possible language pairs (totally 45 pairs) and selected the most confusing ones. The experiments showed that different LID systems perform differently on different language pairs.

By analyzing the results of all the systems on all the language pairs, we found Cantonese, Korean and Mandarin are the most difficult to discriminate from each other by both the i-vector and the DNN systems. The results on the pairs of the three languages are shown in Table III. Finally, we put the three languages together to form the confusing-language set, and baseline performance is reported in Table IV. The test utterances are the full-length set in AP17-OLR-test (this will be also the case in the official test set).

The results indicate that the three languages are truly confusing and difficult to distinguish from each other. This seems not surprising, as they are spoken by people in neighboring areas. Additionally, acoustic analysis shows that the inspiration and aspiration of the pronunciation of these three languages are similar, and social linguistic analysis shows that these languages influence each other in a significant way. This is quite obvious as Chinese and Cantonese share almost the same characters, and Korean borrowed many characters from Chinese, known as ‘hanja’.

TABLE III
C_{avg} and EER results on pairs of Cantonese (Ca), Korean(Kr) and Mandarin (zh)

	Ca/Kr		Ca/zh		Kr/zh	
System	C_{avg}	EER%	C_{avg}	EER%	C_{avg}	EER%
i-vector	0.1684	17.00	0.1381	13.90	0.1172	11.40
i-vector + LDA	0.1569	15.72	0.1753	17.07	0.1246	12.40
i-vector + PLDA	0.1584	16.07	0.1850	18.36	0.1219	12.19
TDNN	0.3478	36.98	0.4360	61.72	0.1663	16.72
LSTM	0.3080	36.60	0.4513	66.10	0.2131	21.45

TABLE IV
C_{avg} and EER results on Cantonese, Korean and Mandarin

System	C_{avg}	EER%
i-vector	0.1446	14.13
i-vector + LDA	0.1550	15.20
i-vector + PLDA	0.1563	15.64
TDNN	0.3752	37.86
LSTM	0.3902	40.56

3) Open-set language recognition: In this task, utterances from non-target languages will be added to the test set. To have a quick glimpse of the influence of the interference languages, we chose $1,264$ utterances in non-target languages, and combine them with $3,000$ utterances randomly selected from AP17-OLR-test, leading to $42,640$ trials where $12,640$ non-target ones are from the interference languages. The results with and without interference languages are shown in Table V. From the results, it can be seen that interference utterance indeed impacts the performance of LID systems, particularly in terms of EER. In the official test set that will be released, there will be a significant proportion of utterances of non-target languages.

TABLE V
C$_{avg}$ and EER results with and without interference languages

	Without interference		With interference	
System	C_{avg}	EER%	C_{avg}	EER%
i-vector	0.0556	5.93	0.0596	7.27
i-vector + LDA	0.0613	6.17	0.0643	7.40
i-vector + PLDA	0.0563	5.73	0.0606	7.13
TDNN	0.1024	11.33	0.1056	13.53
LSTM	0.1125	12.77	0.1159	14.77

V. Conclusions

We presented the data profile, task definitions and evaluation principles of the AP18-OLR challenge. To assist participants to construct a reasonable starting system, we published two types of baseline systems based on the i-vector model and various DNN models respectively. We showed that the tasks defined by AP18-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 and 61371136.

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

[11] 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.

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

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

[14] 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.

[15] N. Dehak, P. G. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[16] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *INTERSPEECH*, 2011, pp. 857–860.

[17] K. J. Lang, A. H. Waibel, and G. E. Hinton, "A time-delay neural network architecture for isolated word recognition," *Neural networks*, vol. 3, no. 1, pp. 23–43, 1990.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.