Zero-resource Language Recognition

Jiawei Yu†, Jinsong Zhang†*
† Beijing Advanced Innovation Center for Language Resources, Beijing Language and Culture University, Beijing, China
E-mail: vyujiawei@gmail.com, jinsong.zhang@blcu.edu.cn

***Abstract*—Language recognition (LRE) can be categorized into two configurations: in the close-set setting, a test segment is classified into one of several pre-defined languages, and in the open-set setting, a segment that is not in any of the pre-defined languages will be labelled as 'unknown'. In real applications, there is another scenario: we hope to register a new language with several utterances and then this language will be recognized by the system, although this language is not involved when the system is constructed. We call this zero-resource LRE (ZR-LRE).**

In this paper, we explore the language embedding approach and apply it to tackle the ZR-LRE problem. Specifically, we first train an embedding space of languages based on i-vector or d-vector, and then new languages can be registered and recognized within this space. The experiments were conducted on the AP18-OLR database including 10 languages for training the embedding space and another 8 languages as zero-resource (ZR) languages for registration and recognition. To explore the influence of different test condition to the performance of ZR-LRE, we evaluated various configurations that involve different numbers of enrollment utterances and different duration of test utterances. The results show that embedding based on i-vectors is suitable for ZR-LRE, which achieved an Equal Error Rate (EER) of 8.7%.

I. Introduction

Language Recognition(LRE) is the task of automatically identifying or verifying the language being spoken in a given speech utterance [1]. It plays an essential role in multilingual speech pre-processing which is typically followed by speech recognition systems and automatic translation systems [2]. Generally, LRE task include two type: close-set LRE and open-set LRE. Most researches on LRE have been focused on close-set LRE where all test utterances correspond to one of small set target languages. In other words, the languages of the training set and the test set are the same. However, the open-set LRE is that test utterances are not likely to be strictly limited to a small set of target languages but may also correspond to some unknown languages. A system is required to recognize in-set languages and effectively reject out-of-set languages [3].

Specifically, we are given a list of n target language classes, $\{L_1, L_2, ..., L_n\}$. In the close-set scenario, $\{L_1, L_2, ..., L_n\}$ are N different explicitly specified language and they are shared by training set and test set. In the open-set scenario, $\{L_1, L_2, ..., L_n\}$ are still explicitly specified languages, and we can set L_{n+1} denoting any of the yet unseen languages. This means that $\{L_1, L_2, ..., L_n\}$ are shared by training set and test set, but the test set will have more unseen languages as the disturbing languages. The disturbing languages can be one or more languages, and they must be different languages from $\{L_1, L_2, ..., L_n\}$. No matter how many languages L_{n+1} have, however, these languages will only have one label which is 'unknown' languages [1].

Another LRE task, which we call zero-resource LRE (ZR-LRE) , is different from the close-set and open-set LRE. It is that we can register a new language with several utterances and then this language will be recognized by the system, although this language is not involved when the system is constructed. As we can see, ZR-LRE system is more applicable in real-life scenarios, where speech may come from any languages. And it will be very helpful to some low-resource languages which is hard to get enough data to model these languages.

The framework of ZR-LRE system is very similar to speaker recognition system, which they all have an enrollment step. Specifically, assuming that the languages of the training set are $\{L_1, L_2, ..., L_n\}$, and the languages of the test set are $\{L_1^*, L_2^*, ..., L_m^*\}$ where n and m are the total number of languages included in the training set and the test set, respectively. The training set and test set are different languages, and the languages of the enrollment set are the same as the test set, both $\{L_1^*, L_2^*, ..., L_m^*\}$. So we can use the trained language model to extract the embedding for the enrollment set and the test set, then compare the similarity between the enrollment language and a test utterance to make a decision.

A key question in the ZR-LRE system is whether we can build a language space with sufficient generalization for an i-vector or d-vector based ZR-LRE system. So we did a series of experiments on the dataset AP18-OLR [4]. The AP18-OLR training set contains 10 languages and test set contains another 8 languages. The experimental results show that the i-vector and d-vector based ZR-LRE systems can achieve 8.7% and 12.41% in Equal Error Rate (EER), respectively, when the duration of test utterances are full length (about 7 seconds). This results indicate that the language space established by the i-vector and d-vector methods has a good generalization.

II. Zero-resource Language Recognition System

This section presents the structure of the ZR-LRE system and the model used in our study.

A. System Description

In LRE applications, two well-known problems have received much research attention in the speech community: language identification and language verification [1]. Similarly, we also have these two applications in ZR-LRE. Let us assume that we are given a set of M enrollment languages $\{L_m|m = 1, 2, , M\}$ in a system, where M is total number of

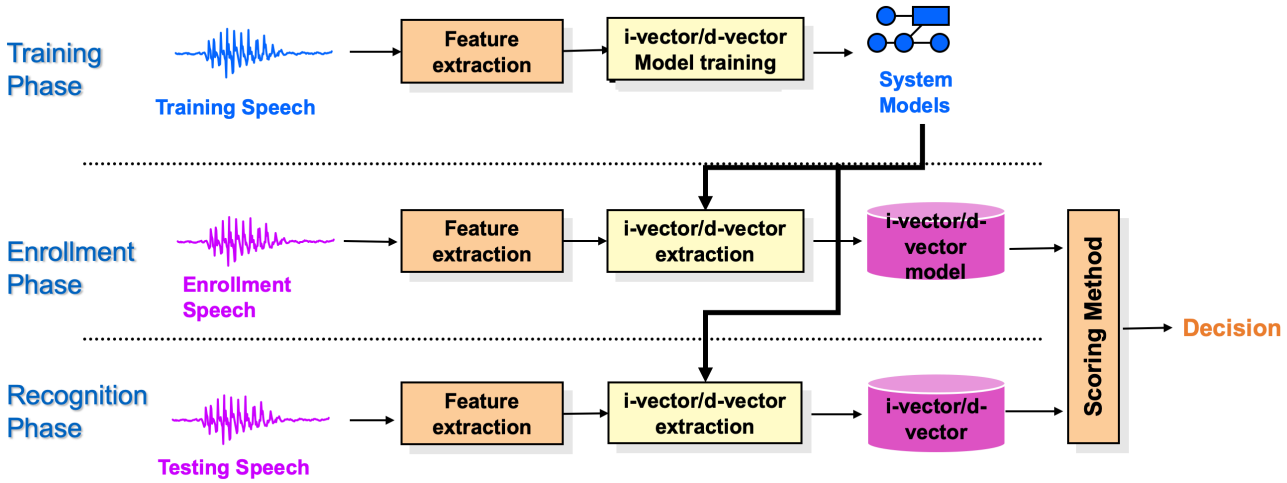

Fig. 1. Block diagram of ZR-LRE system.

enrollment languages. And let X denotes the language vector extracted from a test utterance.

- ZR language verification: Does X belong to language L_i or to other languages (i.e., one of the other $M-1$ languages)?
- ZR language identification: Which of the M languages does X belong to ?

In the ZR language verification problem, the objective is to decide whether to accept or to reject the identity of a claimed enrollment language requiring a binary yes/no answer. In the ZR language identification problem, the objective is to classify a test utterance X into one of the pre-defined set of enrollment languages. To classify X, the decision of the most similar language L^* that maximizes the a posteriori probability is given by

$$L^* = \arg\max_{1\le m\le M} p(L_m|X) = \arg\max_{1\le m\le M} p(X|L_m)p(L_m) \quad (1)$$

Where the language likelihood $P(X|Lm)$ and a language priori probability $P(L_m)$ are assumed known.

Fig. 1. shows a block diagram of the ZR-LRE system proposed in this paper. In the training phase, we perform feature extraction on the training set, and then use it to train i-vector or d-vector models. In the enrollment phase, after completing the Feature extraction for the enrollment set, we can extract the i-vector or d-vector using the trained system model. In the recognition phase, we also extract i-vector or d-vector for the test utterance, and then we can compare the language model obtained by the enrollment phase and the language vector obtained by the recognition phase to classify the testing utterance and make a final decision. The scoring methods used in this paper include Cosine Distance, Linear Discriminant Analysis (LDA) and Probability Linear Discriminant Analysis (PLDA).

B. I-vector/D-vector Model

In this study, we use i-vectors and d-vectors to represent utterances and as a front-end to the ZR-LRE system.

I-vector have been widely used in the state-of-the-art LRE systems. It aims to extract a fixed and low dimension representation from a given utterance based on a factor analysis model. As described in [5], an utterance is projected into a low-dimensional total variability space which contains both channel-dependent and speaker-dependent information, as an i-vector. Given an utterance, the channel-dependent and speaker-dependent GMM vector M can be written as:

$$M = m + Tw \quad (2)$$

where m is the speaker- and channel-independent supervector, usually taken from the universal background model (UBM) [6], T is a rectangular matrix of low rank, referred to as the total variability matrix (TVM), and w is a random vector with a standard normal distribution $N(0, I)$. The vector w contains the total factors and is referred to as the i-vector.

D-vector is proposed by Ehsan Variani et al., motivated by the powerful feature extraction capabilities and the success of deep neural networks(DNNs) applied to speaker recognition [7], [8], [9]. D-vector is like i-vector, trying to look for a more abstract and compact representation of the speech acoustic frames but using a DNN rather than a generative Factor Analysis model. The DNN of extracting d-vector is trained to map frame-level features in a given context to the corresponding language identity target, And the number of DNN's outputs corresponds to the number of language in the training set. Once the DNN has been trained successfully, the frame-level speaker features can be extracted from accumulated output activations of the last hidden layer of the DNN. Then the utterance-level speaker features, which is d-vector, were derived by averaging the frame-level features.

C. Back-end

When we get i-vector and d-vector, the back-end uses the following three scoring metrics: cosine distance, LDA, PLDA.

Cosine scoring is a dot product between test i-vector/d-vector and enrollment language model based on i-vector/d-

vector. The formula for scoring is as follows:

$$Score^{m}_{w_{test}} = \frac{w^{T}_{test}\mu_{m}}{\|w_{test}\|\|\mu_{m}\|} \quad (3)$$

where m is the test utterance, w is the test utterance vector, and μ is the enrollment language model mean.

LDA is a very popular technique for dimension reduction in the machine learning field. It can help maximizing the discrimination between the different classes. In the context of language recognition, each class represents a different language. The LDA procedure consists of finding the basis that maximizes the between classes variability while minimizing the intra-class variability [10].

PLDA is the probabilistic version of LDA. It can decompose the total variation of embedding space(i-vector/d-vector) into language and session variation. So PLDA provides a powerful mechanism in extracting language-specific information from all other sources of undesired variability in i-vector/d-vector space [11].

III. Experimental Setup

For comparison, our baseline is a close-set LRE system. Both the ZR-LRE system and the baseline system use the same i-vector and d-vector approaches. However the test condition is different, For baseline system, we set up two evaluation sets, one with 1000 utterances and one with 2000 utterances. At the same time, we also evaluate the performance of baseline system When the duration of the test utterances is different: 1 seconds(1s), 3 seconds(3s) and full length(about 7 seconds); For ZR system, we set two enroll conditions: 10 enrollment utterances and 20 enrollment utterances for each language. Similarly, We evaluated the performance of ZR-LRE system in different duration for enroll set and test set. The duration of enroll set and test set we use is the same as above, both 1s, 3s and full_length. All the experiments were conducted with Kaldi toolkit [12].

A. Database

The experiments were conducted with the AP18-OLR database which used for third oriental language recognition (OLR) challenge [4]. This challenge has been arranged for three times, with the aim of promoting the research on LID techniques for oriental languages [**?**], [**?**], [4]. The training set and the evaluation set are presented as follows.

1) Training set: The training set consists of 10 different languages: Kazakh in China (ka-cn) , Tibetan in China (ti-cn), Uyghur in China (uy-id), Cantonese in China Mainland and Hongkong (ct-cn), Mandarin in China (zh-cn), Indonesian in Indonesia (id-id), Japanese in Japan (ja-jp), Russian in Russia (ru-ru), Korean in Korea (ko-kr), and Vietnamese in Vietnam (vi-vn). The duration of training data for each language is about 10 hours and the speeches were recorded with mobile phone, and the sample rate is 16kHz. The total number of utterances are 92285 in training set. This dataset was used for training the i-vector and d-vector system, LDA model and PLDA model.

2) Baseline test set: It consist of two test condition. the total number of utterances in one test set is 1000 and the other is 2000. The language is the same as the training set, and the characteristics of the test utterances are the same as those of the training set.

3) ZR-LRE test set: It consist of 8 languages that are completely different from the training set. They are Arabic in The United Arab Emirates (AR-AE), English in United State of America (EN-US), French in France (FR-FR), Hindustani in India (HI-IN), Italian in Italy (IT-IT), Malay in Malaysia (MS-MY), Thai in Thailand (TH-TH) and Urdu in Pakistan (UR-PK). The total duration of test set is 2.43 hours and the characteristics of the test utterances are the same as the training set.

B. Model settings

1) I-vector model: The i-vector system follows the procedure described in [13]. The raw feature involved 13-dimensional MFCCs with a frame-length of 25ms and a frame-shift of 10ms. And an energy-based voice activity detection (VAD) was used by i-vector model. The UBM is a 2048 component full-covariance GMM, and the dimensionality of the i-vector space was 400. The dimensionality of the LDA projection space was set to 150. The i-vectors sent to PLDA are all length normalized.

2) D-vector model: The d-vector system was constructed based on a time delay neural network (TDNN) [14]. The input feature was 40-dimensional FBanks, The configuration of TDNN shows below. The TDNN model was composed of 6 layers and the dimension of each layer is 650. The activation function was p-norm and the spliced indices in the consecutive layers were [$t-2$, $t-1$, t, $t+1$, $t+2$; $t-1$, t, $t+1$; $t-1$, t,$t+1$; $t-3$, t, $t+3$; $t-6$, $t-3$, t] where t is the current frame. A total of 21 frames have been spliced together. The output is a softmax layer and the size is 10 corresponding to the number of languages in the training data. when we get d-vector, the scoring metric of back-end is same as i-vector approach, including cosine distance, LDA and PLDA.

IV. Results

In order to evaluate our proposed ZR-LRE system, we separate the experiments into two parts. First we evaluate the performance of baseline system, then we evaluate the proposed system according to different test condition.

3) close-set LRE: The results of the baseline close-set system in terms of equal error rate (EER%) are reported in Table I. In the back-end of i-vector, the performance of PLDA is the best, while the performance of cosine distance is the worst. In the back-end of d-vector, LDA performs the best, and the cosine distance method is far worse than the LDA method. Moreover, we can see that Whether the total number of utterances in the test set is 1000 or 2000, the d-vector and LDA approach performs the best, which indicates that LDA plays an important role for d-vector system. This observation is consistent with the conclusions obtained in the [15]. In the short duration test set, the d-vector and LDA method continues

to achieve the best results. Compared to the d-vector method, the i-vector based method generally has poor performance.

TABLE I
EER(%) RESULTS OF THE CLOSE-SET LRE SYSTEMS.

			EER%		
Total Numbers[a]	Systems	Scoring	test_1s[b]	test_3s[b]	test_all[b]
1000	I-vector	Cosine	13.90	4.50	2.10
		LDA	13.20	4.00	2.00
		PLDA	**12.30**	**3.70**	**1.70**
	D-vector	Cosine	7.70	6.20	6.10
		LDA	**0.50**	**0.20**	**0.10**
		PLDA	1.90	0.90	0.60
2000	I-vector	Cosine	13.71	4.00	2.05
		LDA	12.86	3.95	2.20
		PLDA	**12.01**	**3.60**	**2.00**
	D-vector	Cosine	8.20	7.10	7.05
		LDA	**0.80**	**0.10**	**0.10**
		PLDA	1.50	0.80	0.60

[a] The Total Number represents the total number of utterances in the test set.
[b] Test_1s, test_3s and test_all represent that the test sentence is 1 second, 3 second and full length (about 7 seconds).

4) ZR-LRE: Table II shows the experimental results of the ZR-LRE system proposed in this paper. The number 10 and 20 in the 'Enrollment' column represents the number of utterance enrolled in each language. The average length of these enrollment utterances is 7 seconds. Firstly, it is not difficult to find that the performance of the ZR-LRE system is improved when the number of enrollment utterances increase. In addition, when the number of enrollment utterances is 10 and 20, the i-vector and cosine based method obtains the best performance in the case of test_all, and obtains 10.64% and 8.7% in EER respectively. Second, as the duration of the utterance is shortened, the d-vector method is gradually as good as the i-vector method, even better than the i-vector method when the number of enrollment utterances is 10. Another interesting observation is that unlike the close-set LRE. In the ZR-LRE system, when the i-vector back-end uses cosine distance, the performance of ZR-LRE system is better than LDA and PLDA back-end. In the d-vector system, the cosine distance back-end also performed well. However, the PLDA back-end does not provide any contribution, and its performance is the worst. A possible reason is that the mean i-vector or d-vector of each enrolled language does not satisfy the Gaussian prior, and there are not many utterances enrolled in the ZR-LRE system. The data used to train PLDA model is insufficient, so the performance of PLDA back-end is bad.

Table III shows the results of different duration of enrollment utterances in ZR-LRE system. One interesting thing we found, in table II we can see, when the duration of the enrollment utterances is fixed and the duration of the test utterances is gradually shortened, the system performance will significantly degraded. And when the duration of test utterances is 1 second, the EER is about 20%. However, when the duration of the test utterances is fixed, the duration of the enrollment utterances is gradually shortened, and the performance degradation of the system will not as obvious as in the previous case. The reason may be that although

TABLE II
EER(%) RESULTS OF THE ZR-LRE SYSTEMS.

			EER%		
Enrollment[a]	Systems	Scoring	test_1s	test_3s	test_all
10	I-vector	Cosine	**21.71**	**14.02**	**10.64**
		LDA	26.60	19.76	16.47
		PLDA	34.46	29.73	26.60
	D-vector	Cosine	**19.93**	17.40	15.88
		LDA	20.27	**15.37**	**13.85**
		PLDA	28.89	25.59	22.97
20	I-vector	Cosine	**18.03**	**10.69**	**8.70**
		LDA	23.91	14.86	12.77
		PLDA	33.79	30.34	27.45
	D-vector	Cosine	19.66	17.57	16.76
		LDA	**18.21**	**14.22**	**12.41**
		PLDA	30.80	26.90	24.91

[a] The Enrollment represents the number of utterances enrolled in each language.

TABLE III
EER(%) RESULTS ON THE DIFFERENT DURATION OF ENROLLMENT UTTERANCES ZR-LRE SYSTEMS.

			EER%		
Enrollment[a]	Systems	Scoring	enroll_1s[b]	enroll_3s[b]	enroll_all[b]
10	I-vector	Cosine	**13.21**	**11.82**	**10.64**
		LDA	23.60	19.76	16.47
		PLDA	30.46	26.73	26.60
	D-vector	Cosine	**16.39**	16.22	15.88
		LDA	16.81	**15.21**	**13.85**
		PLDA	23.90	23.48	22.97
20	I-vector	Cosine	**12.21**	**9.59**	**8.70**
		LDA	18.93	14.58	12.77
		PLDA	30.17	27.53	27.45
	D-vector	Cosine	16.85	16.76	16.76
		LDA	**13.50**	**12.50**	**12.41**
		PLDA	22.83	25.27	24.91

[a] The Enrollment represents the number of utterances enrolled in each language.
[b] Enroll_1s, Enroll_3s and Enroll_all represent that the enrollment utterances is 1 second, 3 second and full length (about 7 seconds).

the duration of the enrollment utterances has shortened, the number of enrollment utterances has not changed, which is 10 and 20. If we enroll 10 one second utterances, it is roughly equivalent to enroll a 10-second utterances. And the quantity of data is still sufficient, so the performance of the short duration ZR-LRE system is not much reduced.

V. CONCLUSIONS

This paper presents a novel ZR-LRE task. It expands a new application for language recognition, which register a new language with a few words and then this language will be recognized by the system, although this language is not involved when the system is constructed. Our experimental results show that the ZR-LRE system can effectively recognize the OOS language by the i-vector and d-vector approaches. When the duration of the test utterances is long (greater than 3 seconds), the i-vector approach achieves better performance, and if the test utterances is short (about 1 second) and the number of enrollment utterances is small (less than 10 utterances), the d-vector approach performs better. Future work will investigate more powerful techniques such as x-vector [16], PTN [17] and LRE with auxiliary information [18], [19].

ACKNOWLEDGMENT

This study was supported by Advanced Innovation Center for Language Resource and Intelligence (KYR17005), the Fundamental Research Funds for the Central Universities (16ZDJ03, 18YJ030006, 19YCX113), and the project of "Intelligent Speech technology International Exchange". The work was also supported by the National Natural Science Foundation of China under Projects 61633013. Jinsong Zhang is the corresponding author.

REFERENCES

[1] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: from fundamentals to practice," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, 2013.

[2] A. Waibel, P. Geutner, L. M. Tomokiyo, T. Schultz, and M. Woszczyna, "Multilinguality in speech and spoken language systems," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1297–1313, 2000.

[3] H. Behravan, T. Kinnunen, and V. Hautamäki, "Out-of-set i-vector selection for open-set language identification." in *Odyssey*, vol. 2016, 2016, pp. 303–310.

[4] 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.

[5] N. Dehak, P. J. 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, 2010.

[6] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[7] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.

[8] L. Li, Y. Chen, Y. Shi, Z. Tang, and D. Wang, "Deep speaker feature learning for text-independent speaker verification," *arXiv preprint arXiv:1705.03670*, 2017.

[9] M. Zhang, X. Kang, Y. Wang, L. Li, Z. Tang, H. Dai, and D. Wang, "Human and machine speaker recognition based on short trivial events," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5009–5013.

[10] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Twelfth annual conference of the international speech communication association*, 2011.

[11] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[13] A. McCree, G. Sell, and D. Garcia-Romero, "Augmented data training of joint acoustic/phonotactic dnn i-vectors for nist lre15," *Proc. of IEEE Odyssey*, 2016.

[14] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[15] D. Wang, L. Li, Z. Tang, and T. F. Zheng, "Deep speaker verification: Do we need end to end?" in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 177–181.

[16] 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.

[17] Z. Tang, D. Wang, Y. Chen, L. Li, and A. Abel, "Phonetic temporal neural model for language identification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 134–144, 2017.

[18] L. Li, Z. Tang, D. Wang, A. Abel, Y. Feng, and S. Zhang, "Collaborative learning for language and speaker recognition," in *National Conference on Man-Machine Speech Communication*. Springer, 2017, pp. 58–69.

[19] Z. Tang, D. Wang, Y. Chen, Y. Shi, and L. Li, "Phone-aware neural language identification," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–6.