Data-Driven Phone Selection for Language Identification via Bidirectional Long Short-Term Memory Modeling

Xiao Song[1,2], Qiang Cheng[3], Jingping Xing[4], and Yuexian Zou[1](✉)

[1] ADSPLAB/Intelligent Lab, SECE, Peking University, Shenzhen, China
zouyx@pkusz.edu.cn
[2] PKU Shenzhen Institute, Shenzhen, China
[3] Shenzhen Press Group, Shenzhen, China
[4] Shenzhen Securities Information Co., Ltd, Shenzhen, China

Abstract. In this paper, we propose a new phone selection method to select more suitable phones with higher score for language identification (LID), which is more similar to target language. A data-driven approach is developed for the phone selection to avoid using complex semantic knowledge which benefits from significant reduction in the manual cost of learning different languages. Recently, bidirectional long short-term memory (BLSTM) can provides more accurate content frame alignments with sequence information from longer duration, which has improved automatic speech recognition (ASR) performance. In principle, the output of BLSTM based ASR contains more candidates in form of phone lattice, which can reduces adverse effect of many practical factors, such as variations of channels, noises and accents. Therefore, initial phones sequences are extracted from phone lattice firstly which are generated by speech recognition results of BLSTM based ASR system. Second, asymmetrical distance between each phone and target language is proposed and then applied to weight the initial phones sequences. Accordingly, language-related phones are selected from the weighted phones. Finally, the selected phones are used to re-score input sentences for the LID system. Intensive experiments have been conducted on AP16-OLR Challenge to validate the effectiveness of our proposed method. It can be seen from results, these selected phones are more effective to LID than the rest phones. Our method gives improvement up to 39.96% in terms of C_{avg} compared with method without using phone selection.

Keywords: Phone selection · Data-driven
Bidirectional long short-term memory · Language identification

1 Introduction

There are many state-of-the-art language identification (LID) systems still include acoustic modeling [1, 2], even though several high level approaches based on phonotactic and prosody were widely used as meaningful complementary information sources [3, 4]. Recently, method of using i-vector front-end features followed by a classification stage was proposed to compensate speaker and session variability [5, 6].

K. Li et al. (Eds.): ISICA 2017, CCIS 873, pp. 301–312, 2018.
https://doi.org/10.1007/978-981-13-1648-7_26

In addition, deep neural networks (DNNs) for LID [7] was shown high preference than that of i-vector based approach. Moreover, long short-term memory (LSTM) with the ability to model sequential data makes it a suitable candidate for acoustic LID systems [8]. The methods discussed previous were only implemented with proprietary DNNs and LSTMs, which are not available for research teams. Besides, the effects of well-known practical factors, such as variations of channel, noise and accent, were not considered in LID systems, which may degrade the performance of the LID systems.

In this paper, we propose an approach of data-driven phone selection method for improving LID results via a non-proprietary bidirectional LSTM (BLSTM). The details include:

(1) The BLSTM is used to build a non-proprietary neural network for modeling acoustic features. While an automatic speech recognition (ASR) system is generated based on both N-gram language model (LM) and the BLSTM acoustic model (AM). The BLSTM provides more accurate content frame alignments with sequence information from longer duration. Obviously, the output of BLSTM based ASR contains more candidates in form of phone lattice [9]. Accordingly, the recognition result can reduce the adverse effects of channel, noises and accents. It is well-known that phone lattice keeping both the scores and time alignment information is an essential part of many state-of-the-art LID system.

(2) A new data-driven phone selection method is proposed to improve the LID results. First, we extract initial phones sequences (speech recognition results) from the phone lattice. Next, different from [10, 11], where [10] used a training data selection method to develop compact models in their hierarchical LID framework, [11] proposed a phonetic unit selection method to represent speech information spoken in a different language, we use asymmetrical distance to determine the phone selection to obtain more "important" phones from the initial phones sequences. When the asymmetrical distance is generated, Gaussian mixture model (GMM) is employed to replace all phones in target language. The GMM can better model and represent the phone set than directly using them in the target language. Besides, the asymmetrical distance is used to weight the initial phones sequences. Finally, selecting some important language-related phones from these weighted phones re-score the input sentences, which produces better LID results. Two other considerations about the data-driven method lie in avoiding to use complex semantic knowledge and to reduce computational cost of learning different languages.

(3) The performance of our method is evaluated on AP16-OLR7 dataset which is provided by the oriental language recognition challenge on APSIPA 2016 (AP16-OLR challenge). It is the benchmark corpus for evaluation of LID [12]. We implement the data-driven phone selection method for LID system (denoted as DDPS-LID) and compare it with the conventional LID system without using phone selection.

The paper is organized as follows. In Sect. 2, we briefly introduce the principle of the LID system with data-driven phone selection (DDPS-LID). In Sect. 3, we present our proposed data-driven phone selection method (DDPS). In Sect. 4, we describe the dataset and evaluation metrics. Experiments and results are showed in Sect. 5. Finally, the conclusions are presented in Sect. 6.

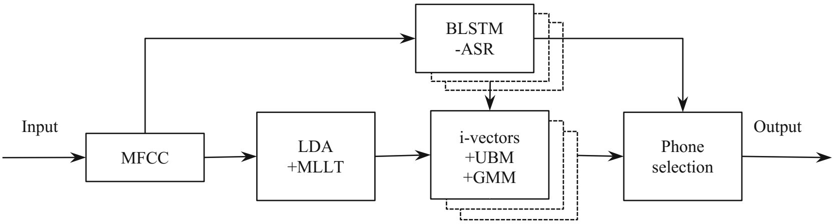

Fig. 1. LID system with phone selection.

2 Proposed LID System

Figure 1 shows the proposed LID system integrated with phone selection module (DDPS-LID). It's noted that acoustic features are needed to compute first. The well-known Mel-frequency cepstral coefficients (MFCC) features are widely used for LID system. In the LID task, the target languages are known in prior, the scores from LID system are comparable across languages and then the highest score refers to the final result of the LID system. Therefore, the score for each language hypothesis will be computed by LID system.

Shown in the top of Fig. 1, BLSTM is used for acoustic modeling and N-gram is used for language modeling. The acoustic model (AM) and language model (LM) are language independent. Considering the limitation of training data, we use phoneme as the modeling units in the AM and LM. The output of the BLSTM-ASR is taken as one of the input to phone selection module.

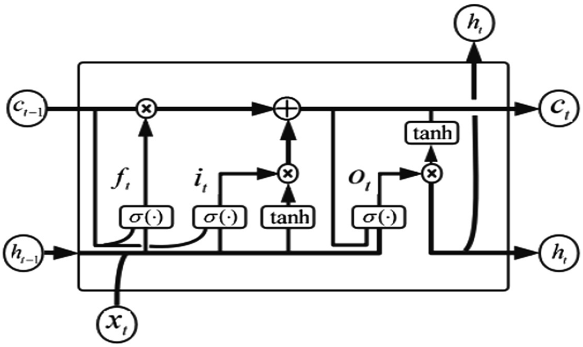

Fig. 2. Bidirectional Long short-term memory (BLSTM) block.

In BLSTM, recurrent connections and special network units called memory blocks contain memory cells with self-connections storing the temporal state of the network which changes with the input to the network at each time step. Figure 2 illustrates a single BLSTM block. In addition, they have multiplicative units called gates to control the flow of information into the memory cell and out of the cell to the rest of the network.

Mathematically, a set of cells can be described by the following forward operations iteratively over time t = 1, 2,…, T:

$$\mathrm{i}_t = \sigma(W^{(xi)}x_t + W^{(hi)}h_{t-1} + W^{(ci)}c_{t-1} + b^{(i)}) \tag{1}$$

$$f_t = \sigma(W^{(xf)}x_t + W^{(hf)}h_{t-1} + W^{(cf)}c_{t-1} + b^{(f)}) \tag{2}$$

$$c_t = f_t c_{t-1} + i_t \tanh(W^{(xc)}x_t + W^{(hc)}h_{t-1} + b^{(c)}) \tag{3}$$

$$o_t = \sigma(W^{(xo)}x_t + W^{(ho)}h_{t-1} + W^{(co)}c_t + b^{(o)}) \tag{4}$$

$$h_t = o_t \tanh(c_t) \tag{5}$$

Where i_t, o_t, f_t, c_t, and h_t are vectors, with the same dimensionality, which represent 5 different types of information at time t of the input gate, output gate, forget gate, cell activation, and hidden layer, respectively [13]. $\sigma(\cdot)$ is the logistic sigmoid function, *W's* are the weight matrices connecting different gates, and *b'*s are the corresponding bias vectors. All the weight matrices are full except the weight matrix $W^{(ci)}$ is diagonal.

Shown in the bottom of Fig. 1, after the Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) module, new features can be obtained. Then the joint Universal Background Model (UBM), i-vectors and GMM are determined accordingly, which are taken as another input to phone selection module.

For each language, those previous steps as the input to the phone selection module, key module, are completed separately. It is noted that, with unknown input utterances, the BLSTM-ASR module generates the phones sequences across each language hypothesis in the form of phone lattice, which are termed as the initial phones sequences. Therefore, initial frame-by-frame scores can be determined from the initial phones sequences. Based on language hypothesis, the distance between each initial phone and the language can be calculated. Then, the phone selection module utilizes the distances calculated above to weight the initial frame-by-frame scores of the initial phones sequences. Finally, select some phones with higher scores to re-score the input sentence, which is the final output score of the input sentence.

3 Proposed Data-Driven Phone Selection Method

As shown in Fig. 1, the phone selection module is a key for DDPS-LID system. In this section, the proposed DDPS method is introduced in details.

3.1 Phone Lattice Generation

In order to describe the principle of phone selection module, we firstly need to understand the phone lattice generation in a LID system. Actually, there are various definitions of phone lattice in literature. Researches [14, 15] show that there is a common point that the alignment and scores information are correct in the phone lattice and the completeness of such information (i.e. no high scores phones sequences are missing). Followed by the method proposed, in [9], the phone lattice is obtained by Kaldi toolkit.

3.2 Asymmetrical Distance Between Phones and Target Languages

For selecting the phones from the initial phones sequences, we need to determine a language-related distance. A distance metric is proposed to measure the distance between the input sentences and the target languages.

First, the distance $D(ph_a,ph_b)$ between two phones is defined as:

$$D(ph_a, ph_b) = -\log P(ph_a \,|\, ph_b) \tag{6}$$

where ph_a and ph_b are phone a and phone b, respectively. $P(ph_a \,|\, ph_b)$ is a conditional probability.

Second, the asymmetrical distance $D(ph_a,L_j)$ between a phone and a target language is defined as:

$$D(ph_a, L_j) = -\frac{1}{N_{ph(L_j)}} \sum_{k=0}^{N_{ph(L_j)}} \log P(ph_a \,|\, ph_k) \tag{7}$$

where L_j is target language j, $N_{ph(L_j)}$ is the number of phones in target language j.

Third, we can obtain the asymmetrical distance $D(utt_i,L_j)$ between an input sentence and a target language as follows:

$$D(utt_i, L_j) = -\frac{1}{N_{ph(utt_i)} \times N_{ph(L_j)}} \sum_{m=0}^{N_{ph(utt_i)}} \sum_{k=0}^{N_{ph(L_j)}} \log P(ph_m \,|\, ph_k) \tag{8}$$

where utt_i is i-th sentence, $N_{ph(utt_i)}$ is the number of phones in i-th sentence.

Finally, we use a GMM of target language to replace all of the phones in this target language. This way can avoid learning all phones of every target language. Instead, the GMM can better model and represent the phone set. So Eq. (8) is simplified as follows:

$$D(utt_i, L_j) = -\frac{1}{N_{ph(utt_i)}} \sum_{m=0}^{N_{ph(utt_i)}} \log P(ph_m \,|\, GMM(L_j)) \tag{9}$$

where $GMM(L_j)$ represents the GMM of target language j.

3.3 Proposed Data-Driven Phone Selection Method

The proposed DDPS method consists of four steps:

(1) using Eq. (9) to calculate the asymmetrical distances between an input sentence and a target language;
(2) a *weight* factor is proposed to weight the initial phones sequences, thus, *weighted phones* can be obtained for the input sentence;
(3) the important language-related phones are selected from the weighted phones to obtain *selected phones*, which are used to represent the input sentence;
(4) the input sentence is re-scored by the selected phones to obtain *final score* of the input sentence.

Specifically, the *weight* factor and the *final score* of input sentence are computed in Eqs. (10) and (11), respectively.

$$\begin{aligned} W(utt_i, L_j) &= \frac{1}{D(utt_i, L_j)} \\ &= \{w_1(utt_{(i,1)}, L_j), w_2(utt_{(i,2)}, L_j), \ldots, w_k(utt_{(i,k)}, L_j), \ldots, w_{N_{ph(utt_i)}}(utt_{(i,N_{ph(utt_i)})}, L_j)\} \\ &= \{\frac{1}{d_1(utt_{(i,1)}, L_j)}, \frac{1}{d_2(utt_{(i,2)}, L_j)}, \ldots, \frac{1}{d_k(utt_{(i,k)}, L_j)}, \ldots, \frac{1}{d_{N_{ph(utt_i)}}(utt_{(i,N_{ph(utt_i)})}, L_j)}\} \end{aligned} \tag{10}$$

$$Score(utt_i, L_j) = \frac{\sum_{k=0}^{N_{ph(utt_i)}} \alpha w_i(utt_{(i,k)}, L_j) * Ph\_Score(utt_{(i,k)}, L_j)}{\sum_{k=0}^{N_{ph(utt_i)}} \alpha w_i(utt_{(i,k)}, L_j)} \tag{11}$$

where $W(utt_i,L_j)$ is the weight factor of i-th sentence and language j. $Score(utt_i,L_j)$ is the final score of i-th sentence and language j. $utt_{(i,k)}$ is the k-th phone of i-th sentence. $w_i(utt_{(i,k)},L_j)$ is the *weight* factor of k-th phone of i-th sentence and language j, which is the derivative of $d_i(utt_{(i,k)},L_j)$. $d_i(utt_{(i,k)},L_j)$ is the distance of k-th phone of i-th sentence and language j. $Ph\_Score(utt_{(i,k)},L_j)$ is the initial score of k-th phone of i-th sentence and language j, which is extracted from the phone lattice generated by the BLSTM-ASR system of language j. The parameter α is given as:

$$\alpha = \begin{cases} 1, & \text{if } utt_{(i,k)} \in U_{\widetilde{(i)}} \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

where $U_{(\widetilde{i})}$ is the set of the selected phones in i-th sentence. $\alpha = 1$ represents the $utt_{(i,k)}$ is selected, otherwise the $utt_{(i,k)}$ is not selected.

3.4 Proposed DDPS-LID System

As discussed above, the initial phones with higher score associated with language L_j will lead to that the i-th input sentence is classified as language L_j greatly. Therefore, the selection operation is necessary.

Our proposed DDPS-LID system includes following steps:

(1) the initial phones sequence is generated by the BLSTM-ASR of language L_j in the form of phone lattice for the i-th input sentence;
(2) the initial score, $Ph\_Score(utt_{(i,k)},L_j)$, is extracted from the phone lattice;
(3) the scores of the i-th input sentence, $Score(utt_i,L_j)$, are obtained across each language hypothesis by the four steps of the proposed DDPS method described as in Sect. 3.3;
(4) the scores of the i-th input sentence across each language hypothesis are comparable and the highest score is the final LID result for the i-th input sentence.

4 Datasets and Evaluation Metrics

4.1 Datasets

In this study, AP16-OL7 database is used to evaluate our proposed DDPS-LID system, which is provided by SpeechOcean[1] in the AP16-OLR challenge [12]. The multilingual database includes seven oriental languages, 71 h of speech signals in total: Cantonese in China mainland and Hong Kong (ct-cn), Indonesian in Indonesia (id-id), Japanese in Japan (ja-jp), Korean in Korea (ko-kr), Russian in Russia (ru-ru), Vietnamese in Vietnam (vi-vn), Mandarin in China (zh-cn), where original scripts and lexicon are available. 24 speakers are included in each language, 18 of them are selected as training set (~ 8 h) and the other 6 are in test set (~ 2 h), where male and female are balanced. The details are given in Table 1.

Table 1. AP16-OL7 data profile.

Datasets		Training set	Testing set
Code	Description	Hours	Hours
ct-cn	Cantonese in China Mainland and Hong Kong	7.71	2.48
id-id	Indonesian in Indonesia	7.48	3.16
ja-jp	Japanese in Japan	5.82	2.16
ko-kr	Korean in Korea	5.99	1.92
ru-ru	Russian in Russia	9.92	2.99
vi-vn	Vietnamese in Vietnam	8.46	2.94
zh-cn	Mandarin in China	7.66	2.63

[1] www.speechocean.com.

4.2 Evaluation Metrics

In order to assess the performance of LID systems, three different metrics are used. As the main error measure to evaluate the capabilities of one to all language detection, C_{avg} (average cost) is used as defined in the AP16-OLR challenge evaluation plan (as in LRE15). C_{avg} is a measure of the cost of taking bad decisions, and thus, it considers not only discrimination, but also the ability of setting optimal thresholds.

Further, metric Equal Error Rate (EER) is used to evaluate the performance, which considers only scores of each individual language. The EER metrics are not related to the decision result, but the quality of the scoring. Therefore it evaluates the verification system from different angle. Detailed information can be found in LRE'09 evaluation.

In the AP16-OLR challenge, the target languages are known in prior, and the scores are comparable across languages, which means that OLR can be treated as a language identification task, for which the language obtaining the highest score in a trail is regarded as the identification result. For such an identification task, identification recognition results (IDR) is a widely used metric, which treats errors on all languages equally serious.

5 Experimental Settings and Results

5.1 Experimental Settings

We used BLSTM layers with recurrent and non-recurrent projections for ASR as suggested in kaldi. The neural networks were created using alignments from GMM system trained using Kaldi. Following this, a GMM was trained firstly with standard 13-dimensional MFCC features, using Hann windows of 25-ms frames shifted by 10 ms each time for each target language ($GMM(L_i)$, $i \in \{1,2,3,\ldots,7\}$). The forced alignment given by a GMM system was used to create frame-level acoustic targets. At the same time, the GMM is adopted to replace the set of phones in the target language for calculating the distance between an unknown input sentence with a known target language (described in Sects. 3.2, 3.3 and 3.4). Recognition is performed by combining the acoustic probabilities yielded by the network with the state transition probabilities from the HMM and the word transition probabilities from the LM, which can be done efficiently for speech using weighted finite state transducers.

The BLSTM had three hidden layers, with 640 cells in each layer. BLSTM was trained by using stochastic gradient descent (SGD) with one weight update per utterance, and the truncated back propagation through time (BPTT) learning algorithm. Mealwhile, learning rate ranging from 5×10^{-4} to 5×10^{-5} that are exponentially decayed during training was used. At each layer, a constant delay of −3 (or +3) along both directions was adopted.

In phonotactics modeling, each different phonotactics N-gram language model was trained using the phonetic sequences of the challenge training data for each tokenizer. For that purpose, the SRILM toolkit had been used and 4-gram back-off models smoothened using Kneser-Ney discounting are obtained. Finally, for an input speech, each BLSTM model was evaluated and log-likelihood scores were obtained.

5.2 Experimental Results

In our proposed DDPS method, the key points are that how to select phones and how many phones selected from all of the weighted phones in an unknown input sentence. We used an example to explain these two key points, showed in Tables 2 and 3, respectively.

Table 2. An example of the scores of some weighted phones in one input sentence under different language *hypothesis*.

	ik1	*eon2*	*d*	*ing3*	*i1*	*n*
ct-cn	25.7	45.3	13.7	16.4	13.5	32.2
	"e	*"oj*	*dZ*	*"i*	*n*	*"i*
id-id	4.1	6.5	4.3	7.5	2.9	7.5
	ie3	*ong4*	*d*	*shi4*	*n*	*an4*
zh-cn	27.9	27.6	0.3	46.3	13.4	27.4

As an example shown in Table 2, the ground-truth of the input sentence is language "ct-cn", and the scores of initial phones sequences were weighted under different language hypothesis (shows scores of some weighted phones). In fact, the scores between each phone and each language are different. For each language hypothesis, we selected the weighted phones with relatively higher scores. If just selecting one phone, the phone *eon2* could be selected under language "ct-cn" hypothesis. Because that the phone *eon2* has the highest score compared with other weighted phones in the input sentence under language "ct-cn" hypothesis. If selecting two phones, the top two phones *eon2* and *n* could be selected under language "ct-cn" hypothesis.

Table 3. The different scores in sentences level for the input sentence and each language hypothesis by using the data-driven phone selection method

Lan.	Scores of one utterance by choosing various numbers of weighted phones					
	Top 2	Top 4	Top 6	Top 8	Top 10	All
ct-cn	38.8	31.5	26.0	**22.7**	18.2	15.3
id-id	7.1	6.2	5.5	4.9	4.38	4.1
ja-jp	12.0	11.2	10.1	9.12	7.92	6.9
ko-kr	10.7	8.9	8.0	7.1	6.43	5.6
ru-ru	**55.2**	**36.9**	**26.4**	20.8	17.2	14.9
vi-vn	12.7	10.6	9.25	8.4	7.42	6.4
zh-cn	37.2	30.3	26.0	22.6	**18.7**	**15.6**

Table 3 shows the different scores in sentences level for the input sentence and each language hypothesis by using the DDPS method. From Table 3, when selecting the weighted phones with top two weighted scores in all phones, we can see that the highest score of the whole sentence is 55.2 compared with under each language hypothesis. And the recognition result for this utterance is language "ru-ru". However, with increasing the number of the selected weighted phones, we can find the result is changed. Selecting top eight weighted phones obtained "ct-cn" result, where the highest score of the sentence is 22.7 under language "ct-cn" hypothesis, which is the correct result. But selecting top ten weighted phones obtain "zh-cn" result, where the highest score of the sentence is 18.7 under language "zh-cn" hypothesis. Therefore, the selected "degree" of the weighted phones was explored on average performance, showed in Fig. 3.

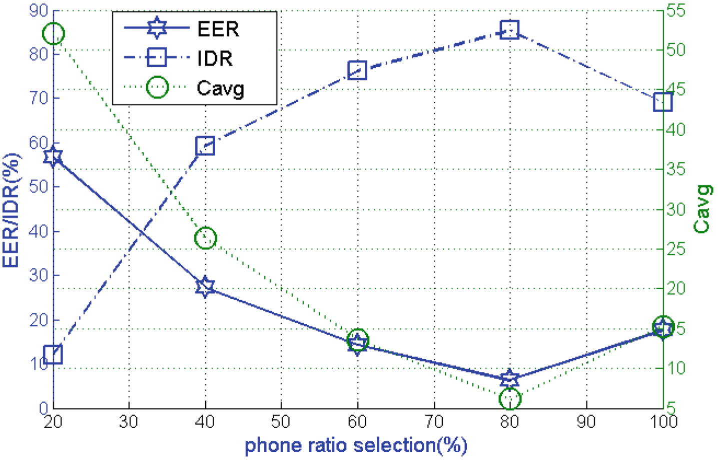

Fig. 3. Performance (EER, IDR and C_{avg} on average) by using the proposed data-driven phone selection method, where selecting various proportion of the weighted phones shows the different results.

Figure 3 shows the average performance on three evaluation metrics EER, IDR and C_{avg} (described in Sect. 4.2). From Fig. 3, we can discover that when choosing the top 20% of the weighted phones in the input sentences, the results are very bad, no matter what the metric is. With the increasing of the proportion of selecting weighted phones, the results are improved. When selecting 70% $\sim$ 85% of the phones, the performance become better. However, selecting more than 85% of phones, especially when using all phones, the result is worse than the use of partial selected weighted phones.

Table 4 summarizes the results of some teams participating in the AP16-OLR challenge and our results obtained in terms of C_{avg}, EER and IDR (on average). For each target language, by using the proposed method, we obtained better result than our submitted in the challenge (our submitted result is obtained by using the initial scores). We can also find that the best result is Haizhou Li team from Singapore. We didn't

Table 4. Methods performance on AP16-OL7

Methods	C_{avg}	EER%	IDR%
Haizhou Li,Singapore	1.13	1.09	97.56
NTUT,Taiwan,China	5.86	5.88	87.02
MMCL_RUC,China	6.06	6.16	86.21
NTU,Singapore	14.72	17.44	71.44
Our-submitted	36.99	40.26	31.91
TLO,China	50.00	53.34	12.37
No phone selection	15.29	17.63	69.06
Phone selection 60%	13.65	14.28	76.22
Phone selection 80%	6.11	6.31	85.39

exceed the state-of-art team, even break through the third team (C_{avg} = 6.06). But, their methods were proprietary and complex [16]. Compared with them, we propose a non-proprietary, simpler solution in LID task.

6 Conclusions

In this paper, we proposed a data-driven phone selection approach for language identification (LID). The bidirectional long short-term memory (BLSTM) in automatic speech recognition (ASR) was designed to obtain the phone lattice for its better speech recognition capability. Next, a data-driven phone selection method was proposed where asymmetrical distance between each phones and target languages was used to weight the phones from lattice. These weighted phones were used to re-score the input sentence which was the final score of the input sentence. Finally, intensive experiments had been conducted to evaluate the proposed method on the AP16-OLR challenge. Our method gave an improvement of 39.96% in terms of C_{avg} with respect to the LID without using phone selection method. In the future, we will further evaluate the quality of the selected phones for improving the performance of LID systems.

Acknowledgements. This work is partially supported by Key Technologies Research & Development Program of Shenzhen (No: JSGG20150512160434776) and Key Technologies Research & Development of Data Retrieval and Monitoring via Multi-layer Network (No: JSG G20160229121006579).

References

1. Torres-Carrasquillo, P.A., Singer, E., Gleason, T., McCree, A., Reynolds, D.A., Richardson, F., Sturim, D.E.: The MITLL NIST LRE 2009 language recognition system. In: Acoustics Speech and Signal Processing (ICASSP) IEEE International Conference on 2010, pp. 4994–4997 (2010)
2. Gonzalez-Dominguez, J., Lopez-Moreno, I., Franco-Pedroso, J., Ramos, D., Toledano, D.T., Gonzalez-Rodriguez, J.: Multilevel and session variability compensated language recognition: ATVS-UAM systems at NIST LRE 2009. IEEE J. Sel. Top. Sig. Proc. **4**(6), 1084–1093 (2010)

3. Ferrer, L., Scheffer, N., Shriberg, E.: A comparison of approaches for modeling prosodic features in speaker recognition. In: International Conference on Acoustics, Speech, and Signal Processing, pp. 4414–4417 (2010)
4. Martinez, D., Lleida, E., Ortega, A., Miguel, A.: Prosodic features and formant modeling for an ivectorbased language recognition system. In: Acoustics, Speech and Signal Processing (ICASSP) IEEE International Conference on 2013, pp. 6847–6851 (2013)
5. Dehak, N., Torres-Carrasquillo, P.A., Reynolds, D.A., Dehak, R.: Language recognition via i-vectors and dimensionality reduction. In: Interspeech ISCA, pp. 857–860 (2011)
6. Martinez, D., Plchot, O., Burget, L., Glembek, O., Matejka, P.: Language recognition in ivectors space. In: Interspeech ISCA, pp. 861–864 (2011)
7. Lopez-Moreno, I., Gonzalez-Dominguez, J., Plchot, O., Martinez, D., Gonzalez-Rodriguez, J., Moreno, P.: Automatic language identification using deep neural networks. In: Acoustics, Speech and Signal Processing (ICASSP) IEEE International Conference on 2014, pp. 5337–5341 (2014)
8. Gonzalez-Dominguez, J., Lopez-Moreno, I., Sak, H., Gonzalez-Rodriguez, J., Moreno, P.J.: Automatic language identification using long short-term memory recurrent neural networks. In: Interspeech, pp. 2155–2159 (2014)
9. Povey, D., Hannemann, M., Boulianne, G., Burget, L., Ghoshal, A., Janda, M., Karafiat, M., Kombrink, S., Motlicek, P., Qian, Y., et al.: Generating exact lattices in the WFST framework. In: Proceedings of ICASSP, pp. 4213–4216 (2012)
10. Irtza, S., Sethu, V., Fernando, S., Ambikairajah, E., Li, H.: Out of set language modelling in hierarchical language identification. In: Interspeech 2016, pp. 3270–3274 (2016)
11. Lopez-Otero, P., Docio-Fernandez, L., Garcia-Mateo, C.: Phonetic unit selection for cross-lingual query-by-example spoken term detection. In: Automatic Speech Recognition and Understanding (ASRU) IEEE Workshop on 2015, pp. 223–229 (2015)
12. Wang, D., Li, L., Tang, D., Chen, Q.: AP16-OL7: a multilingual database for oriental languages and a language recognition baseline, submitted to APSIPA 2016.pdf
13. Graves, A., Mohamed, A., Hinton, G.E.: Speech recognition with deep recurrent neural networks. In: International Conference on Acoustics, Speech, and Signal Processing (2013)
14. Sak, H., Saraclar, M., Güngör, T: On-the-fly lattice rescoring for real-time automatic speech recognition. In: Interspeech, pp. 2450–2453 (2010)
15. Ortmanns, S., Ney, H., Aubert, X.: A word graph algorithm for large vocabulary continuous speech recognition. Comput. Speech Lang. **11**, 43–72 (1997)
16. Irtza, S., Sethu, V., Fernando, S., Ambikairajah, E., Li,H.: Out of set language modelling in hierarchical language identification. In: Interspeech 2016, pp. 3270–3274 (2016)