CSLT TECHNICAL REPORT-20160027 [Saturday 7th January, 2017]

Collaborative Learning for Language and Speaker Recognition

Lantian Li[1,2]
, Zhiyuan Tang[1,2]
, Dong Wang[1,3*]
, Yang Feng[1,2]
and Shiyue Zhang[1,2]

*Correspondence: wang-dong99@mails.tsinghua.edu.cn
[1]Center for Speech and Language Technology, Research Institute of Information Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China
Full list of author information is available at the end of the article

Abstract

This paper presents a unified model to perform language and speaker recognition simultaneously and altogether. The model is based on a multi-task recurrent neural network where the output of one task is fed as the input of the other, leading to a collaborative learning framework that can improve both language and speaker recognition by borrowing information from each other. Our experiments demonstrated that the multi-task model outperforms the task-specific models on both tasks.

Keywords: Language recognition; Speaker recognition; Deep learning; Recurrent neural network

1 Introduction

Language recognition (LRE) [1] and speaker recognition (SRE) [2] are two important tasks in speech processing. Traditionally, the research in these two fields seldom takes account of each other, although some techniques are indeed shared, e.g., SVM [3], the i-vector model [4, 5], and deep neural models [6, 7]. This can be largely attributed to the intuition that speaker characteristics are language independent in SRE, and dealing with speaker variation is regarded as a trivial request in LRE. This independent processing of language identities and speaker traits, however, is not the way we human beings process speech signals: it is easy to imagine that our brain recognizes speaker traits and language identities simultaneously, and the success of identifying languages helps discriminate speakers, and vice versa. In fact, some researchers have noticed that language and speaker are two correlated factors. For example, it has been confirmed that language mismatch indeed leads to serious performance degradation for speaker recognition [8, 9, 10, 11], indicating that language and speaker are correlated and should be modelled, trained, and decoded jointly.

A simple joint learning approach is to pool multilingual data of multiple speakers and train models that cover multiple language and speaker conditions. This pooled training approach is easy to implement and generally effective (e.g., [8, 9]), but it does not consider the interaction between language identities and speaker traits. Another joint learning approach is based on the joint factor analysis (JFA)

framework [12]. This approach treats language and speaker as two dependent random variable and their linear combination explains the distribution of speech data. This approach was employed by Lu et al. [13] to deal with multilingual speaker recognition, and obtain interesting performance gains. A potential issue of the JFA approach is that it is a pure generative model, and therefore is less powerful for discriminative tasks such as SRE and LRE.

This paper presents a novel collaborative learning approach which models correlated factors explicitly as JFA, but trains the model discriminatively. The basic idea is to feed the output of one task as part of the input of the other task, resulting in a multi-task recurrent model. By this way, the two tasks can be learned simultaneously and collaboratively. This collaborative learning approach, which is a special joint learning method, was recently proposed by Tang et al. [14], and has been successfully applied to speech and speaker joint training. In this paper, we apply the collaborative learning approach to SRE and LRE, as illustrated in Fig. 1. Note that collaborative learning is a general framework and the component of each task can be implemented using any model, and we prefer recurrent neural networks (RNN) due to its great potential in various speech processing tasks including SRE [7]. For LRE, although there is no literature to report the performance with RNN, we will show in this paper that it does provide highly competitive performance. In summary, the contributions of this paper include: (1) we demonstrated that SRE and LRE can be jointly learned by collaborative learning; (2) we demonstrated that RNN is highly powerful for SRE and LRE, especially with collaborative learning. Furthermore, when the test speech is extremely short (e.g., one second), the RNN model can deliver impressively better performance than the state-of-the-art i-vector/PLDA approach for SRE and i-vector/SVM approach for LRE.

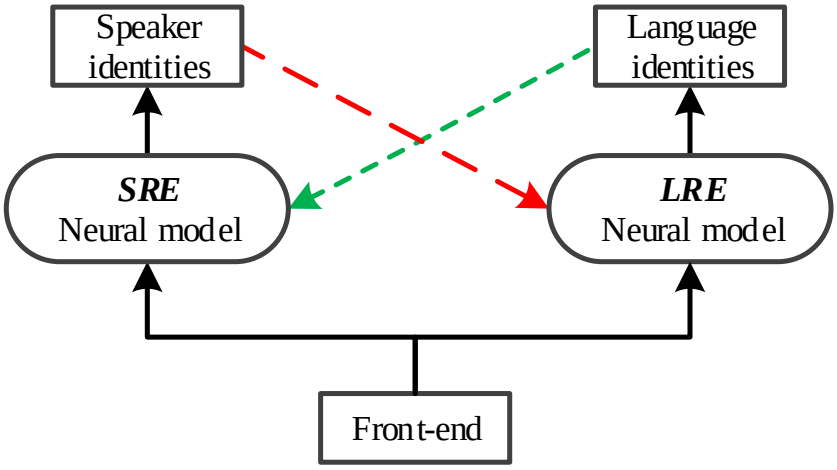

Figure 1 Multi-task recurrent model for language and speaker recognition.

The rest of the paper is organized as follows: Section 2 presents the model architecture, and Section 3 reports the experiments. The conclusions and the future work are presented in Section 4.

2 Multi-task RNN and collaborative learning

This section starts from the neural model structure for single tasks, and then extends to the multi-task recurrent model for collaborative learning.

2.1 Basic single-task model

We choose a particular RNN, the long short-term memory (LSTM) [15], to build the baseline single-task systems for SRE and LRE. LSTM has delivered good performance in SRE [16], and we will show that it is also an effective model for LRE. Particularly, the recurrent LSTM structure proposed in [17] is used, as shown in Fig. 2, and the associated computation is as follows:

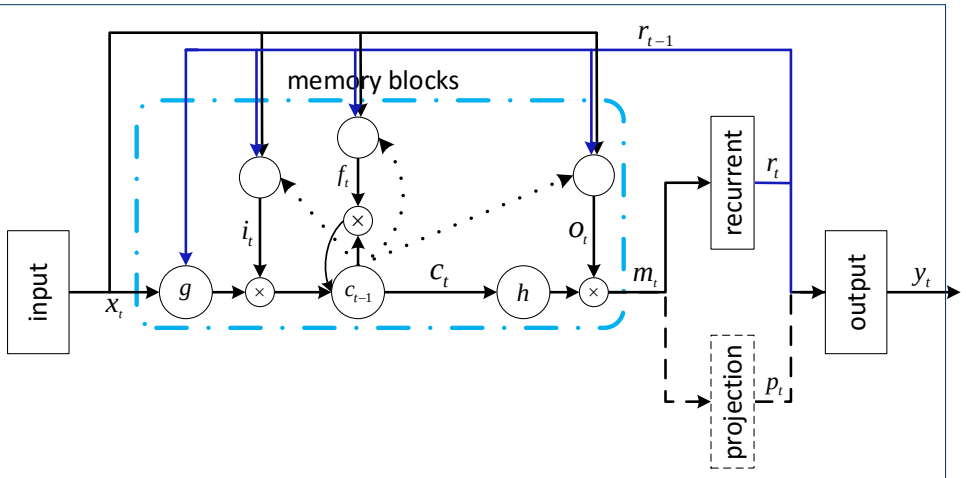

Figure 2 Basic recurrent LSTM model for LRE and SRE single-task baselines.

$$
\begin{aligned}
i_t &= \sigma(W_{ix}x_t + W_{ir}r_{t-1} + W_{ic}c_{t-1} + b_i) \\
f_t &= \sigma(W_{fx}x_t + W_{fr}r_{t-1} + W_{fc}c_{t-1} + b_f) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g(W_{cx}x_t + W_{cr}r_{t-1} + b_c) \\
o_t &= \sigma(W_{ox}x_t + W_{or}r_{t-1} + W_{oc}c_t + b_o) \\
m_t &= o_t \odot h(c_t) \\
r_t &= W_{rm}m_t \\
p_t &= W_{pm}m_t \\
y_t &= W_{yr}r_t + W_{yp}p_t + b_y.
\end{aligned}
$$

In the above equations, the W terms denote weight matrices and the b terms denote bias vectors. x_t and y_t are the input and output vectors; i_t, f_t, o_t respectively represent the input, forget and output gates; c_t is the cell and m_t is the cell output. r_t and p_t are two output components derived from m_t, in which r_t is recurrent and used as input of the next time step, while p_t is not recurrent and contributes to the present output only. $\sigma(\cdot)$ is the logistic sigmoid function, and $g(\cdot)$ and $h(\cdot)$ are non-linear activation functions, often chosen to be hyperbolic. $\odot$ denotes the element-wise multiplication.

2.2 Multi-task recurrent model

The basic idea of the multi-task recurrent model, as shown in Figure 1, is to use the output of one task at the current time step as an auxiliary input of the other task at the next step. In this study, we use the recurrent LSTM model shown in the previous section to build the LRE and SRE components, and then combine them by some inter-task recurrent connections. This results in a multi-task recurrent model by which LRE and SRE can be trained in a collaborative way (collaborative learning). The model structure is shown in Figure 3, where we use the superscript l and s to denote the LRE and SRE task, respectively and the dash lines represent the inter-task recurrent connections.

A multitude of model configurations can be selected. The first question is from where the recurrent feedback should be extracted. For example, it can be extracted from the cell c_t or cell output m_t, or from the output component r_t or p_t, or even from the output y_t. Another question is to where the feedback information should be propagated. It can be the input variable x_t, the input gate i_t, the output gate o_t, the forget gate f_t, or the non-linear function $g(\cdot)$. Note that a weight matrix is introduced for each recurrent feedback.

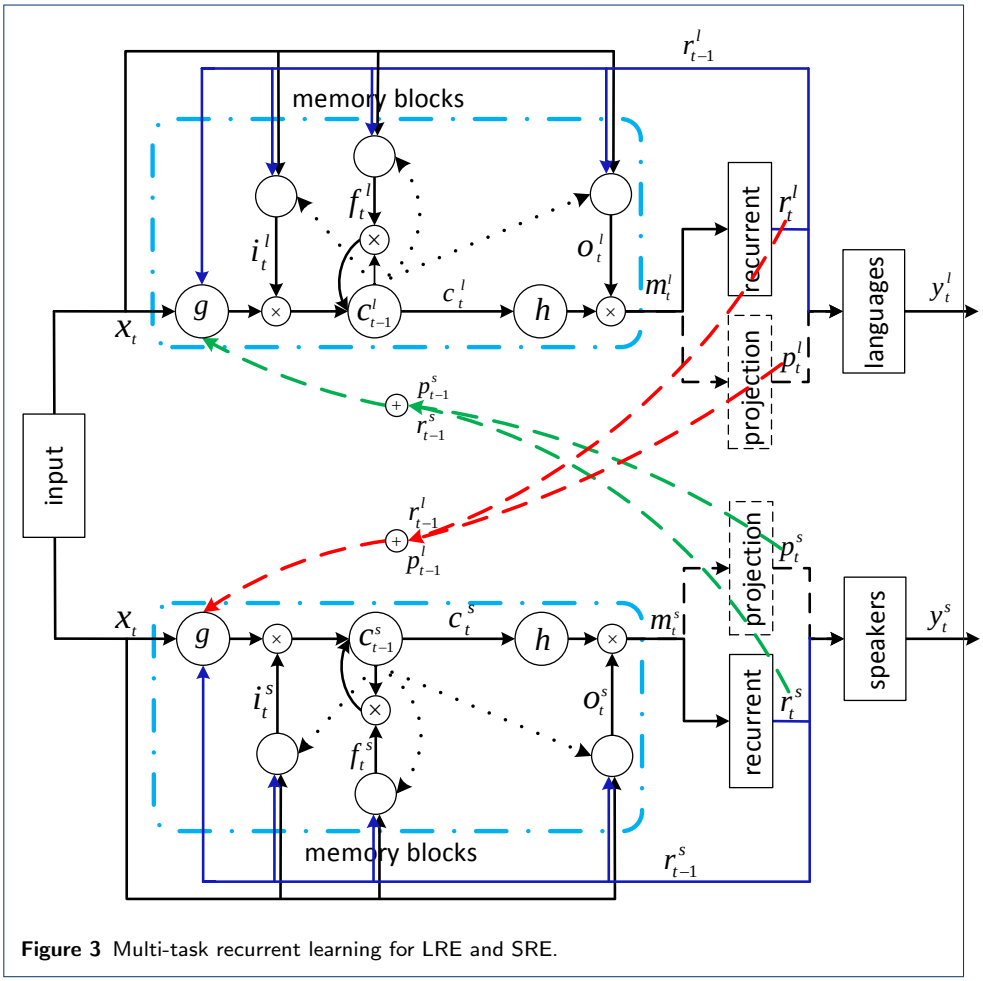

Figure 3 Multi-task recurrent learning for LRE and SRE.

In view of all the above alternatives, the multi-task recurrent model is rather flexible. The structure shown in Fig. 3 is just one simple example, where the recurrent

feedback is extracted from both the recurrent projection r_t and the nonrecurrent projection p_t, and the feedback is propagated to the non-linear function $g(\cdot)$. With the recurrent feedback, the computation for LRE can be expressed as follows:

$$\begin{aligned}
i_t^l &= \sigma(W_{ix}^l x_t + W_{ir}^l r_{t-1}^l + W_{ic}^l c_{t-1}^l + b_i^l) \\
f_t^l &= \sigma(W_{fx}^l x_t + W_{fr}^l r_{t-1}^l + W_{fc}^l c_{t-1}^l + b_f^l) \\
g_t^l &= g(W_{cx}^l x_t^l + W_{cr}^l r_{t-1}^l + b_c^l + \underline{W_{cr}^{ls} r_{t-1}^s + W_{cp}^{ls} p_{t-1}^s}) \\
c_t^l &= f_t^l \odot c_{t-1}^l + i_t^l \odot g_t^l \\
o_t^l &= \sigma(W_{ox}^l x_t^l + W_{or}^l r_{t-1}^l + W_{oc}^l c_t^l + b_o^l) \\
m_t^l &= o_t^l \odot h(c_t^l) \\
r_t^l &= W_{rm}^l m_t^l \\
p_t^l &= W_{pm}^l m_t^l \\
y_t^l &= W_{yr}^l r_t^l + W_{yp}^l p_t^l + b_y^l
\end{aligned}$$

and the computation for SRE is as follows:

$$\begin{aligned}
i_t^s &= \sigma(W_{ix}^s x_t + W_{ir}^s r_{t-1}^s + W_{ic}^s c_{t-1}^s + b_i^s) \\
f_t^s &= \sigma(W_{fx}^s x_t + W_{fr}^s r_{t-1}^s + W_{fc}^s c_{t-1}^s + b_f^s) \\
g_t^s &= g(W_{cx}^s x_t^s + W_{cr}^s r_{t-1}^s + b_c^s + \underline{W_{cr}^{sl} r_{t-1}^l + W_{cp}^{sl} p_{t-1}^l}) \\
c_t^s &= f_t^s \odot c_{t-1}^s + i_t^s \odot g_t^s \\
o_t^s &= \sigma(W_{ox}^s x_t^s + W_{or}^s r_{t-1}^s + W_{oc}^s c_t^s + b_o^s) \\
m_t^s &= o_t^s \odot h(c_t^s) \\
r_t^s &= W_{rm}^s m_t^s \\
p_t^s &= W_{pm}^s m_t^s \\
y_t^s &= W_{yr}^s r_t^s + W_{yp}^s p_t^s + b_y^s
\end{aligned}$$

3 Experiments

In this section, we first describe the data profile, and then present the baseline systems. Finally, experimental results of the collaborative learning approach are given.

3.1 Data

Two databases were used to perform the experiment: the WSJ database in English and the CSLT-C300 database in Chinese. All the utterances in the two databases were labelled with both language and speaker identities. The development set involves two subsets: WSJ-E200 that contains 200 speakers (24,031 utterances) selected from WSJ, and CSLT-C200 that contains 200 speakers (20,000 utterances) selected from the CSLT-C300 database. The development set was used to train SVM, the i-vector model and the multi-task recurrent model.

The evaluation set involves an English subset WSJ-E110 that contains 110 speakers selected from WSJ, and a Chinese subset CSLT-C100 that contains 100 speakers

selected from the CSLT-C300 database. For each speaker in each subset, 10 utterances were used to enroll its speaker and language identity, and the rest 13,236 utterances in English and 9,000 utterances in Chinese are used to perform test. For SRE, the test is pair-wised, leading to 13,236 target trails and 1,442,724 imposter trails in English, plus 9,000 target trails and 891,000 imposter trials in Chinese. For LRE, the number of test trails is the same as the number of test utterances, which is 13,236 for English trails and 9,000 for Chinese trails.

3.2 LRE and SRE baselines

We first present the LRE and SRE baselines. All the experiments were conducted with the Kaldi toolkit [18]. Two baseline systems were constructed, one is based on i-vectors, and the other is based on LSTM.

3.2.1 i-vector baseline

For the i-vector baseline, the acoustic feature was 39-dimensional MFCCs. The number of Gaussian components of the UBM was 1,024, and the dimension of i-vectors was 200. The produced i-vectors were used to conduct both SRE and LRE with different scoring methods. For SRE, we consider the simple Cosine distance as well as the popular discriminative models LDA and PLDA; for LRE, we consider Cosine distance and SVM. All the discriminant models were trained on the development set.

The results of the SRE baseline are reported in Table 1, in terms of equal error rate (EER). We tested two scenarios, one is a full-length test which uses the entire enrollment and test utterance; the other is a short-length test which involves only 1 second of speech (sampled from the original data after voice activity detection applied). In both scenarios, the language of each test is assumed to be known in prior, i.e., the test on English and Chinese datasets are independent.

Table 1 SRE baseline results.

Test	System	Dataset	EER(%)		
			Cosine	LDA	PLDA
Full	i-vector	English	0.88	0.70	**0.62**
		Chinese	1.28	0.97	**0.84**
	r-vector	English	1.25	1.38	3.57
		Chinese	1.70	1.61	4.93
Short	i-vector	English	7.00	4.01	3.47
		Chinese	9.12	6.16	5.69
	r-vector	English	3.27	**2.70**	7.88
		Chinese	4.77	**3.99**	8.21

For LRE, the purpose is to discriminate two languages (English and Chinese). Hence, it is an identification task. We use identification error rate (IDR) [19] to measure performance of LRE systems, which is the fraction of the identification mistakes in the total identification trials. For a more thorough comparison, the number of identification errors (IDE) is also reported. The results of the i-vector/SVM baseline system are reported in Table 2.

3.2.2 r-vector baseline

The r-vector baseline is built based on the recurrent LSTM structure shown in Fig. 2. The SRE and LRE systems use the same configurations: the dimensionality of the

Table 2 LRE baseline results.

Test	System	IDR(%)	IDE
Full	i-vector/Cosine	3.43	763
	i-vector/SVM	**0.01**	2
	r-vector/Cosine	0.11	25
	r-vector/SVM	0.21	47
	r-vector/Softmax	0.13	29
Short	i-vector/Cosine	10.21	2270
	i-vector/SVM	1.40	311
	r-vector/Cosine	0.98	218
	r-vector/SVM	0.63	139
	r-vector/Softmax	**0.58**	129

cell was set to $1,024$, and the dimensionality of both the recurrent and nonrecurrent projections was set to 100. For the SRE system, the output corresponds to the 400 speakers in the training set; For the LRE system, the output corresponds to the two languages to identify: English and Chinese.

The output of the recurrent and nonrecurrent projections were concatenated and averaged over all the frames of an utterance, resulting in an 'r-vector' for that utterance. The r-vector derived from the SRE system represents speaker characters, and the r-vector derived from the LRE system represents the language identity. As in the i-vector baseline, decisions were made based on distance between r-vectors, measured by either the Cosine distance or some discriminative models. The same discriminative models as in the i-vector baseline were used, except that in the LRE system, the softmax outputs of the task-specific LSTMs can be directly used to identify language. The results are shown in Table 1 and Table 2 for SRE and LRE, respectively.

The results in Table 1 show that for SRE, the i-vector system with PLDA performs better than the r-vector system in the Full-length test. However, in the Short-length test, the r-vector system is clearly better. This is understandable as the i-vector model is generative and relies on sufficient data to estimate the data distribution; the LSTM model, in contrast, is discriminative and the speaker information can be extracted even with a single frame. The discrepancy on model type also explains the observation that the discriminative models are more effective for the i-vector system than for the r-vector system, as the former is discriminative already. A pair-wised t-test confirms that the performance advantage of the r-vector/LDA system over the i-vector/PLDA system is statistically significant ($p <$ 1e-5).

The results in Table 2 show a similar trend, that the i-vector system (with SVM) works well in the Full-length test, but in the Short-length test, the r-vector system shows much better performance, even with the simple Cosine distance. Again, this can be explained by the fact that the i-vector model is generative, while the r-vector model is discriminative. To the authors' best knowledge, we are among the earliest to report the impressive success of RNN on short-utterance LRE.

3.3 Multi-task recurrent model with collaborative learning

The multi-task recurrent LSTM system, as shown in Fig. 3, were constructed by combining the LRE and SRE r-vector systems, with inter-task recurrent connections augmented. Following the experience in [14], we employ the recurrent projection output of the SRE (LRE) system as the feedback, and tested the results when this

Table 3 SRE results with collaborative learning.

Feedback Input				EER(%)			
				Full		Short	
i	f	o	g	Eng.	Chs.	Eng.	Chs.
r-vector Baseline				1.38	1.61	2.70	3.99
√				1.27	1.43	2.50	3.61
	√			1.38	1.38	2.55	3.52
		√		**1.19**	**1.31**	**2.48**	3.66
			√	1.37	1.48	2.67	**3.52**
√	√	√	√	1.32	1.31	2.52	3.69

Table 4 LRE results with collaborative learning.

Feedback Input				IDE					
				Full			Short		
i	f	o	g	Cosine	SVM	Softmax	Cosine	SVM	Softmax
r-vector Baseline				25	47	29	218	139	129
√				5	2	0	11	6	2
	√			1	0	0	3	1	1
		√		11	2	0	21	8	3
			√	0	0	1	2	2	1
√	√	√	√	6	2	0	17	10	2

feedback is propagated to different components of the LRE (SRE) system. The results are reported in Table 3 and Table 4 for SRE and LRE, respectively.

The results show clearly that the collaborative learning provides consistent performance improvement on both SRE and LRE, despite which component the feedback is applied to. Experimental performance suggests that the output gate is an appropriate component for SRE to receive the feedback, whereas for LRE, the forget gate seems a more suitable choice. However, these observations are based on the relatively small databases. More experiments on large data are required to confirm and understand the observations. We finally highlight that the collaborative training provides very impressive performance gains for LRE: it significantly improves the single-task r-vector baseline, and beats the i-vector baseline even on the Full-length task. This strongly support our conjecture that correlated tasks should be learned jointly, as our brain does every day.

4 Conclusions

We report a novel collaborative learning approach that performs speaker and language recognition as a unified process, based on a multi-task recurrent neural network. Primary results demonstrated that the proposed approach can deliver consistent performance improvement over the single-task baselines, particularly for the LRE task. Future work involves experimenting with large databases and analyzing the properties of the collaborative mechanism, e.g., trainability, stability and extensibility.

Author details

[1]Center for Speech and Language Technology, Research Institute of Information Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China. [2]Center for Speech and Language Technologies, Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology, ROOM 1-303, BLDG FIT, 100084 Beijing, China. [3]Department of Computer Science and Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China.

References

1. Jiri Navratil, "Spoken language recognition-a step toward multilinguality in speech processing," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 6, pp. 678–685, 2001.
2. Frédéric Bimbot, Jean-François Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meignier, Teva Merlin, Javier Ortega-García, Dijana Petrovska-Delacrétaz, and Douglas A Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 2004, pp. 430–451, 2004.
3. William M Campbell, Joseph P Campbell, Douglas A Reynolds, Elliot Singer, and Pedro A Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, vol. 20, no. 2, pp. 210–229, 2006.

4. Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.
5. Najim Dehak, A.Torres-Carrasquillo Pedro, Douglas Reynolds, and Reda Dehak, "Language recognition via ivectors and dimensionality reduction," in *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2011, pp. 857–860.
6. Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan 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.
7. Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer, "End-to-end text-dependent speaker verification," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5115–5119.
8. Bin Ma and Hellen Meng, "English-Chinese bilingual text-independent speaker verification," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2004, pp. V–293.
9. Roland Auckenthaler, Michael J. Carey, and John.S.D. Mason, "Language dependency in text-independent speaker verification," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2001, pp. 441–444.
10. Abhinav Misra and John H. L. Hansen, "Spoken language mismatch in speaker verification: An investigation with nist-sre and crss bi-ling corpora," in *IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 372–377.
11. Rozi Askar, Dong Wang, Fanhu Bie, and Thomas Fang Zheng, "Cross-lingual speaker verification based on linear transform," in *IEEE China Summit and International Conference on Signal and Information Processing*. IEEE, 2015, pp. 519–523.
12. Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.
13. Liang Lu, Yuan Dong, Zhao Xianyu, Liu Jiqing, and Wang Haila, "The effect of language factors for robust speaker recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2009, pp. 4217–4220.
14. Zhiyuan Tang, Lantian Li, and Dong Wang, "Multi-task recurrent model for speech and speaker recognition," *arXiv preprint arXiv:1603.09643*, 2016.
15. Sepp Hochreiter and Schmidhuber Jürgen, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
16. Lantian Li, Yiye Lin, Zhiyong Zhang, and Dong Wang, "Improved deep speaker feature learning for text-dependent speaker recognition," in *Proceedings of APSIPA Annual Summit and Conference*. APSIPA, pp. 426–429.
17. Hasim Sak, Andrew W Senior, and Françoise Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2014, pp. 338–342.
18. Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.
19. Bo Yin, Ambikairajah Eliathamby, and Chen Fang, "Hierarchical language identification based on automatic language clustering," in *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2007, pp. 178–181.