The Speaker and Language Recognition Workshop (Odyssey 2018)
26-29 June 2018, Les Sables d'Olonne, France

An analysis of transfer learning for domain mismatched text-independent speaker verification

Chunlei Zhang, Shivesh Ranjan, John H.L. Hansen

Center for Robust Speech Systems (CRSS), University of Texas at Dallas, Richardson, Texas, U.S.A.
{chunlei.zhang,john.hansen}@utdallas.edu

Abstract

In this paper, we present transfer learning for deep neural network based text-independent speaker verification, in the presence of a severe mismatch between the enrollment and the test data. Given a pre-trained speaker embedding network developed with out-of-domain data, we explore and analyze how this pre-trained model can benefit for the in-domain speaker verification task. Two alternative strategies are investigated to perform transfer learning, i.e., vanilla transfer learning (V-TL) and curriculum learning based transfer learning (CL-TL). The proposed methods are validated on UT-SCOPE-physical speech corpus, where we create a setup to introduce mismatched evaluation conditions with the neutral and the physical task stressed speech. Experimental results confirm the effectiveness of both V-TL and CL-TL techniques. Employing transfer learning based on the pre-trained model, we are able to achieve a +47.7% relative improvement over a conventional i-vector/PLDA system and a +30.6% relative improvement over a recent proposed end-to-end system, respectively.

Index Terms:Transfer learning, speaker embedding, i-vector, domain adaptation, convolutional neural networks

1. Introduction

In speaker recognition, i-Vector/PLDA systems achieve very impressive performances in the presence of noise and channel variabilities [1, 2, 3, 4]. However, these systems often rely on a large collection of in-domain and well-annotated data, e.g., transcriptions for ASR DNN acoustic modeling and speaker labels for PLDA training [5, 6]. Studies have shown a significant performance gap between in-domain and out-of-domain systems [7, 8, 9, 10]. Also, it is expensive to collect a large amount of labeled data for every new domain. All these factors make domain mismatched speaker recognition a challenging task.

In the domain adaptation challenge (DAC), the challenge setup was primarily focused on developing techniques which utilize information from unlabeled in-domain data. Researchers were able to retrieve most of the performance drop with unsupervised speaker clustering and PLDA adaptation in the i-Vector space [7, 8]. However, it is noted that the DAC corpus is mostly English speech. Consequently, we find that similar techniques are not so effective in the NIST SRE16 setup, where a more severe domain mismatch (i.e., language mismatch between training data and enrollment/test data) is designed to encourage effective domain adaptation methods [9, 10]. This observation implies that the domain mismatch could also exist in front-end feature representations (i.e., i-Vectors in the i-Vector/PLDA systems). In the i-Vector estimation process, an acoustic model (either unsupervised GMMs or supervised ASR DNNs) is employed to generate frame-level soft alignments [11, 5, 12]. Adapting an acoustic model to new domain low resource data remains an interesting yet difficult topic in both speech recognition and speaker recognition [13, 14]. Especially in the case of DNN based acoustic model, satisfactory amount of in-domain data are always essential for systems to retain good performance. This formulates the bottleneck of speaker recognition systems which rely on acoustic partitioning paradigm.

Recently, end-to-end systems using a single network have drawn sufficient attention in speaker recognition [15, 16, 17]. With an objective function that directly separates different-speaker utterances and aggregates same-speaker utterances, the end-to-end systems achieve state-of-the-art performance in various tasks while simplifying the speaker recognition problem in the meanwhile. Under this new framework, it is much easier to adapt the pre-trained model to new domain using transfer learning, by utilizing the speaker discriminative ability from out-of-domain data [18].

We examine transfer learning for neural network based text-independent speaker verification. A pre-trained Inception-resnet-v1 network is utilized as the out-of-domain model [15]. UT-SCOPE-physical speech corpus is employed as the in-domain data [19]. Instead of focusing on duration, noise or channel variabilities like the conventional NIST SREs [3, 20], the corpus addresses more on the variability introduced by speakers themselves (i.e., $intrinsic$ neutral/physical stressed mismatch). To create the mismatched conditions, neutral-read speech, stressed-read speech and stressed-spontaneous speech are collected from each speaker. The corpus, therefore, provides a useful setup to study different transfer learning strategies based on the degree of speaker verification difficulty. To conduct speaker verification experiments, we apply both the vanilla transfer learning (V-TL) and a curriculum learning based transfer learning (CL-TL) in this study [21]. Both methods are proved to be effective compared with an i-Vector/PLDA baseline and two end-to-end baselines with just in-domain data. The CL-TL paradigm outperforms V-TL as CL-TL initials the training process with easier samples (neutral data) and successively adds more difficult samples (stressed data) later.

The remainder of the paper is organized as follows. Speaker verification systems including a i-Vector/PLDA baseline and an end-to-end baseline are introduced in Sec.2. Transfer learning basics and CL-TL are described in Sec.3. A corpus description together with trials on different conditions can be found in Sec.4. Sec.5 details the experimental results and analysis. We

This project was funded in part by AFRL under contract FA8750-15-1-0205 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen.

10.21437/Odyssey.2018-26

conclude our work in Sec.6.

2. Speaker verification baselines

2.1. Baseline1:i-Vector/PLDA system

The i-Vector system is based on the Kaldi SRE10 V1 [22]. The front-end features consist of 20 MFCCs with a frame length of 30ms that are mean-normalized over a sliding window of up to 3 seconds. Delta and acceleration are appended to create 60 dimension feature vectors. Nonspeech parts of the utterances are removed with energy based voice activity detection. The UBM is a 1024 component full-covariance GMM. The system uses a 400 dimension i-vector extractor. Prior to PLDA scoring, i-Vectors are centered and length normalized. The dataset employed for the UBM, T-Matrix and PLDA training is described in Sec.4.

2.2. Baseline2: end-to-end system with fixed length input [15]

A simplified block diagram of the end-to-end baseline is depicted in Fig.1. For the speaker discriminative network training, a $triplet\ sampling$ module samples a batch of triplets so that each triplet consists of an $anchor$ x^a as a reference utterance, a $positive$ utterance x^p which shares the same speaker ID as $anchor$ sample, and a $negative$ utterance x^n from a different speaker. Assuming a deep neural network f_θ that maps acoustic features x into the fixed length embeddings $f_\theta(x)$. The objective of the network training is to minimize the $distance$ between the embeddings of $anchor$ and $positive$ samples, and maximize the $distance$ between the embeddings of $anchor$ and $negative$ samples. Instead of identifying speaker ID directly, the proposed framework is forced to minimize within-speaker variability and maximize between-speaker variability by $triplet\ loss$ in a speaker embedding space [23, 15, 16]. L_2 normalization is applied before triplet loss calculation, and the network parameter θ is updated after each batch of triplets.

2.2.1. Triplet loss

For speaker verification, an utterance triplet (x_i^a, x_i^p, x_i^n) is mapped into an speaker embedding triplet $(f(x_i^a), f(x_i^p), f(x_i^n))$, the network training wants embeddings to follow:

$$\|f(x_i^a) - f(x_i^p)\|_2^2 < \|f(x_i^a) - f(x_i^n)\|_2^2, \quad (1)$$

$$\forall (x_i^a, x_i^p, x_i^n) \in \mathcal{T}$$

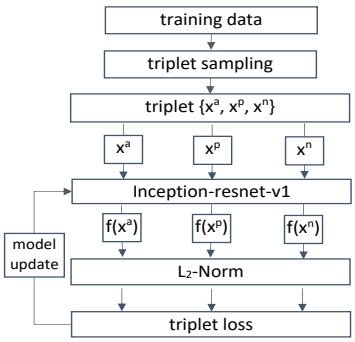

Figure 1: A simplified speaker embedding training system with triplet loss

where $\mathcal{T}$ is the set of triplets. A margin α is empirically defined such that enough distance is enforced between positive and negative pairs with network mapping. We employ $Euclidean\ distance$ as the similarity criteria. The triplet loss is formulated as Eq. (3) with the objective to minimize this loss over the whole set $\mathcal{T}$:

$$\Delta_i = \|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha, \quad (2)$$

$$L = \sum_{i \in \mathcal{T}} max(0, \Delta_i), \quad (3)$$

where L is the triplet loss over a mini-batch, the gradient w.r.t the "anchor" input f_θ^a, "positive" input f_θ^p, and "negative input f_θ^n:

$$\frac{\partial L}{\partial f_\theta^a} = \sum_{i=1}^{N} \begin{cases} 2(f(x_i^n) - f(x_i^p)), & if \Delta_i \geq 0 \\ 0, & otherwise \end{cases} \quad (4)$$

$$\frac{\partial L}{\partial f_\theta^p} = \sum_{i=1}^{N} \begin{cases} 2(f(x_i^p) - f(x_i^a)), & if \Delta_i \geq 0 \\ 0, & otherwise \end{cases} \quad (5)$$

$$\frac{\partial L}{\partial f_\theta^n} = \sum_{i=1}^{N} \begin{cases} 2(f(x_i^a) - f(x_i^n)), & if \Delta_i \geq 0 \\ 0, & otherwise \end{cases} \quad (6)$$

with a hinge loss like design in Eq. (3), triplet samples which are already well separated (corresponding gradient is 0) will not contribute to the gradient calculation for the batch-wise network update according to Eq.(4), Eq.(5) and Eq.(6), which speeds up the learning process during training.

2.2.2. Network architecture and the pre-trained model

We continue to employ Inception-resnet-v1 network architecture to map acoustic features into speaker embeddings, which is the same model developed in [15]. 4s fixed-length input version is adopted as the pre-trained model in our study. The model is trained with 1673 speakers and approximately 600 hours speech, which means sufficient speaker and context variabilities have been incorporated in the process[1].

2.2.3. Input features

Same feature format is utilized to maintain consistency with the pre-trained model, i.e., 4s spectrogram with cropping or padding strategy [24]. More specifically, the height and length of a spectrogram is based on 16 kHz sample-rate, 512 point STFT and a 0.5 skip rate. In the frequency domain, we reserve 0-5K range, and make height to be 160. With 4s in time axis, we can create a 160×250 2-d spectrogram from one utterance.

2.2.4. Two similarity scoring metrics

The embedding can be considered as a speaker representation and therefore used to measure the similarity between speakers. In this study, the first likelihood scoring metric which we use is the negative Euclidean distance between pairs.

$$S(x_{enroll}, x_{test}) = -\|f(x_{enroll}) - f(x_{test})\|_2^2, \quad (7)$$

where $\|\cdots\|_2$ is the 2-norm operation of a vector. For an L_2-normalized vector $\mathbf{x}$, $\mathbf{y}$ (i.e., $\|\mathbf{x}\|_2 = 1$, $\|\mathbf{y}\|_2 = 1$), we prove

[1] http://kingline.speechocean.com/exchange.php?id=1191&act=view

that squared Euclidean distance is proportional to the cosine distance:

$$\begin{aligned} \|\mathbf{x}-\mathbf{y}\|_2^2 &= (\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y}) \\ &= \mathbf{x}^T\mathbf{x} - 2\mathbf{x}^T\mathbf{y} + \mathbf{y}^T\mathbf{y} \\ &= 2 - 2\cos\angle(\mathbf{x},\mathbf{y}) \end{aligned} \tag{8}$$

Eq.(8) indicates that triplet speaker embedding with cosine distance scoring (CDS) is actually the end-to-end system with negative Euclidean distance scoring.

Motivated by its effectiveness of separating speaker information from other sources of undesired variability (channel, noise etc.), we employ PLDA as the second back-end scoring method. In the experimental section, we show PLDA brings additional performance gain in the triplet speaker embedding space.

3. Transfer learning

Transfer learning is machine learning with an additional source of information apart from the standard training data: knowledge from one or more related tasks. The goal of transfer learning is to improve learning in the target task by leveraging knowledge from the source task. In our case, we want to transfer speaker discriminative capability from source domain into target domain which leads to improved learning of three common measures: 1) better initial performance; 2) faster speed of convergence; 3) better final performance.

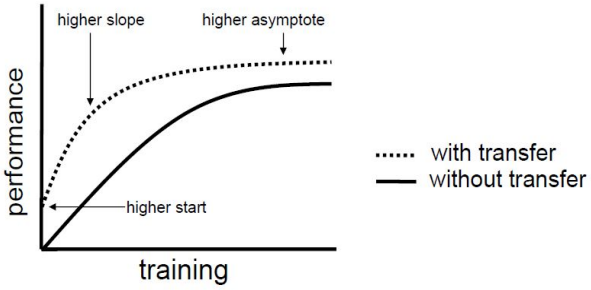

Figure 2: Three measures in which transfer might improve learning.

3.1. Transfer learning scenarios for speaker verification

In practice, there are several ways of performing transfer learning for neural network based speaker verification. There are two major transfer learning scenarios which are suitable for our task.

3.1.1. Inception-resnet-v1 as fixed feature extractor

Take the pre-trained Inception-resnet-v1 model, then treat the whole network as a fixed feature extractor for the domain-of-interest dataset (in-domain dataset). Once feature extraction process is done for all in-domain utterances, apply/train a back-end classifier (e.g., CDS or PLDA) for the in-domain dataset.

3.1.2. Fine-tuning the pre-trained Inception-resnet-v1

The second strategy is to fine-tune the weights of the pre-trained network by continuing the backpropagation. It is possible to fine-tune all the layers of the Inception-resnet-v1, or it is possible to keep some of the earlier layers fixed and only fine-tune some higher-level portion of the network. This is motivated by the observation that the earlier layers of a convolutional neural

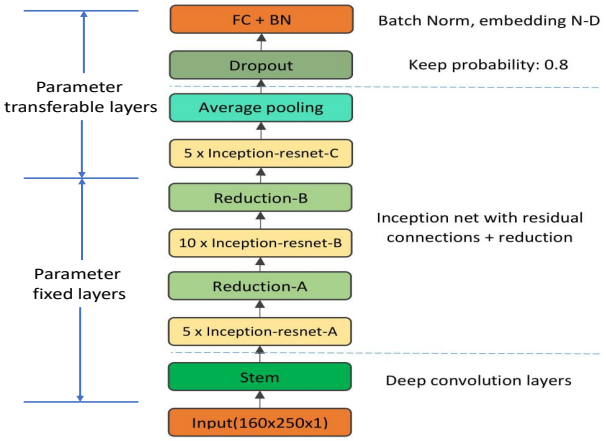

Figure 3: An illustration of V-TL strategy based on Inception-resnet-v1.

network contain more generic features that should be useful to many tasks, but later layers of the convolutional neural network becomes more specific to the details of the classes contained in the original dataset [25]. Since in this study, the neural network that we are investigateing is a very deep CNN architecture, fine-tune the whole network parameters with limited in-domain data is not feasable. We split the Ineption-resnet-v1 into fixed layers and transferable layers. As illustrated in Fig. 3, the higher layers (i.e., an Inception-resnet-C block and a final fully connected layer) which are more related to the triplet loss objective is adapted with the in-domain dataset. For convenience, we note this as Vanilla Transfer Learning (V-TL).

3.2. Curriculum learning based transfer learning (CL-TL)

Inspired by Curriculum Learning based Probabilistic Linear Discriminant Analysis (CL-PLDA) [21], we argue that the neural network fine-tuning based on transfer learning can also be conducted progressively with curriculum learning criteria.

Transfer learning is done iteratively via fine-tuning back-propagation. Like the CL-PLDA model, CL-TL also requires to start learning process with easy training samples, and progressively add more difficult training samples. We hypothesize that this human learning inspired approach could lead to better local minima for non-convex functions (i.e., Inception-resnet-v1), which also leads to improved speaker verification performance.

Similar to the method that uses SNR as the measure of difficulty to guide CL-PLDA training for noise robust speaker verification [21], we also need to find such a learning difficulty metric for CL-TL. In UT-SCOPE-physical speech corpus, neutral-read, stressed-read and stressed-spontaneous speech is collected for each speaker. This creates an intuitive way of dataset partition: neutral-read subset stands for an easy entry task; The learning process becomes more difficult when the stressed-read subset is added into the neutral-read subset; finally, the speaker verification becomes the most difficult when neutral-read, stressed-read and stressed-spontaneous subsets are mixed as a whole dataset, where a sever mismatch is formulated (e.g., a neutral-read utterance and a stressed-spontaneous utterance is grouped in one triplet). The details of corpus statistics and datasets separation are presented in the following section.

4. UT-SCOPE-physical speech corpus

UT-SCOPE-physical speech corpus includes 119 speakers, among which 33 speaker speakers are male and 86 speakers are female. The dataset was collected by the Center for Robust Speech Systems (CRSS) at University of Texas at Dallas [26]. Physical stress is induced by requiring subjects to maintain a 10-mph pace based on visual speed display on an elliptical stair-stepping machine. For each speech type (i.e., neutral and physical task stressed), speech was collected using 35 sentence prompts, prompted through headphones, and a 3 min spontaneous speech segment involving a conversation between the experimenter and the subject. It is noted that the spontaneous speech was collected while the subject was continuing in stair-stepper mode, therefore the stress load should be equal or larger than that in stressed-read speech.

4.1. Corpus statistics

Table 1 is the statistics that can better describe the corpus. We split the dataset into train, validation and test subsets by speaker so that the validation/test speakers are unseen in the training set. It is noted that only the training set is employed in UBM, T-Matrix, PLDA and transfer learning process. There are 10 overlapped speakers between validation set and test set, since UT-SCOPE-physical speech corpus is a small-scale corpus. For each subset, we always keep the male/female ratio close to 33:86.

Table 1: Corpus statistics, utterance count in each subset. NR, SR and SS stands for neutral-read, stressed-read and stressed-spontaneous speech, respectively.

	NR	SR	SS	total	mean/s
training	2755	2799	1223	6777	3.52
validation	457	456	183	1096	3.28
test	1890	1809	741	4440	3.57

4.2. Evaluation trials on mismatched conditions

Gender dependent trials are generated from test set in the experimental analysis. With this setup, 12 evaluation conditions are created from NR, SR and SS subsets as described in Table 2. Of all the 10316 random generated trials, 26.3% of them are target trials.

Table 2: Trials in different evaluation conditions.

eval-cond	NR-NR	NR-SR	NR-SS	SR-SR	SR-SS	SS-SS
female	1212	1444	1406	1748	1406	1186
male	312	300	300	400	300	300

5. Experimental results

In this section, the results of two baselines are firstly presented to give a sense of the speaker verification task on UT-SCOPE-physical speech corpus. Then, we show how V-TL and CL-TL based model adaptation improves the system performance. Finally, an analysis of relationship between stress level and speaker verification performance is illustrated in the following section.

5.1. Baseline performances: i-Vector VS speaker embedding

Table 3 details three baseline results (i.e., i-Vector, speaker embedding with pre-trained model (*Emb-p*) and speaker embedding trained from scratch (*Emb-s*)) on every evaluation condition. In terms of overall performance, the observation *Emb-s*>*Emb-p*>*iVec* indicates:

1) neural network based speaker embedding systems are more effective that i-Vector system on UT-SCOPE-physical speech corpus;

2) pre-trained (out-of-domain) model is able to obtain a reasonable initial performance for in-domain dataset, especially when in-domain data is small;

3) Transfer in-domain information to adapt pre-trained model is essential to improve the performance for domain-mismatched speaker verification.

Table 3: EERs (%) on different conditions for the i-Vector and two speaker embedding baselines with the same PLDA backend. The entry noted as $Cond$ covers the evaluation trial conditions depends on gender (M and F) and physial stress load.

	Cond	NR-NR	NR-SR	NR-SS	SR-SR	SR-SS	SS-SS	overall
iVec	M	**14.4**	15.0	18.0	**12.0**	17.0	16.0	18.7
	F	**10.2**	16.8	27.0	**11.7**	22.6	20.9	
Emb-p	M	6.7	8.0	21.0	8.0	19.0	19.0	15.6
	F	8.2	13.5	23.3	10.5	20.9	21.6	
Emb-s	M	5.8	7.0	21.0	6.0	18.0	**24.0**	14.1
	F	7.9	10.2	21.9	8.4	19.3	**20.9**	

Another interesting observation from two baselines with only in-domain data (*iVec & Emb-s*) is that: there are some conditions that violate a common observation that female trials are usually more challenging than male trials in speaker verification. We argue that significantly more female speakers in the dataset can be one possible reason which leads to better female acoustic/speaker modeling. While for *Emb-p* system, we do not see such a violation because the pre-trained model is based on a gender balanced dataset [15], thus the speaker verification performance follows the common expectation with respect to gender.

5.2. V-TL for the speaker embedding system

Given all the in-domain training data and a pre-trained model, it is intuitive to perform V-TL first. As shown in Fig.4, the V-TL adapted speaker embedding system with PLDA achieves +31.9%, +38.5% and +48.7% relative improvement over the *Emb-s*, *Emb-p*, and *iVec* baseline. PLDA brings additional +27.3% relative improvement compared with CDS metric, which conforms the observation made earlier in the Switchboard experiments [27].

5.3. CL-TL for the speaker embedding system

Data selection for transfer learning is actively studied recently in many fields as it might also lead to negative transfer [25, 28]. We apply CL criteria to transfer learning and observe learning process with a three stage CL-TL. NR samples are firstly employed to fine-tune the pre-trained Inception-resnet-v1, followed by NR and SR samples. Finally, all the training data with the most challenging SS samples are introduced to optimize the network training.

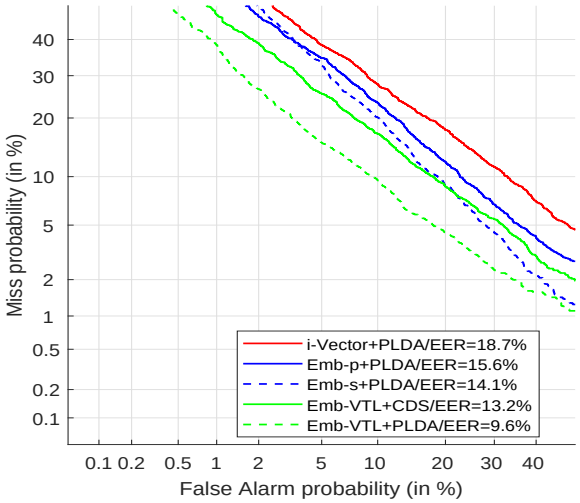

Figure 4: DET curves of V-TL adapted speaker embedding and baseline systems. Only the overall performance is reported for simplicity.

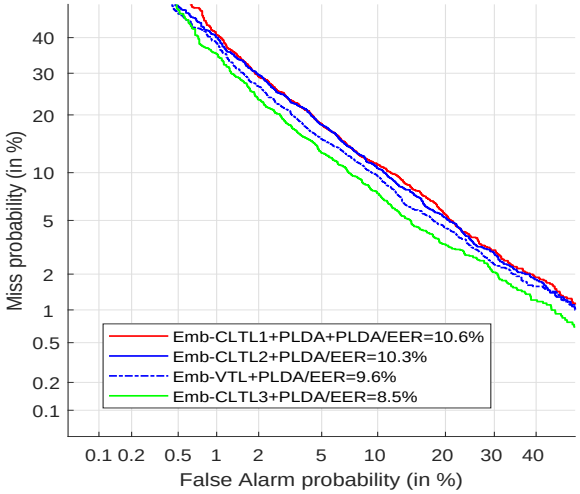

Figure 5: DET curves of CL-TL adapted speaker embedding systems.

Fig.5 illustrates the DET curves of this progressively learning strategy. Transfer learning with NR in-domain data produces the largest improvement compared with adding other subsets, mainly because of domain-mismatch between UT-SCOPE-physical speech corpus and corpus in [15]. In the mean while, +11.4% relative improvement over V-TL system proves the effectiveness of CT-TL learning strategy.

5.4. Performance on subsets with different stress level

The EERs on each stress level subset are also examined in our study. To do so, we utilize the data from all three subsets (i.e., NR, SR and SS) as enrollment data, and test EER on single subset accordingly. 4 systems are reported in Table 4, including the best *Emb-CLTL3* and three baselines. As shown in Table 4, all-SR condition achieves the best performance for all the systems, mainly because SR has the least mismatch between NR and SS enrollment data. Compared with NR data, sharing the prompts contrained context varibility, which helps to improve the speaker verification performance on NR-SR trials. While for SR-SS trials, the varibility introduced by physical stress is learnt and compensated through network training or i-Vector/PLDA modeling. This observation suggests us that it is a possible solution to adapt two sever mismatched conditions into a common sapce for either speaker verification or general domain adaptation tasks.

Table 4: Analysis on EERs (%) for different stress level subsets.

system	all-NR	all-SR	all-SS	system	all-NR	all-SR	all-SS
iVec	19.4	16.7	23.6	Emb-p	14.8	13.8	21.8
	/	-2.7	+4.2		/	-1	+7
Emb-s	12.7	11.9	20.7	Emb-CLTL3	7.3	7.0	13.4
	/	-0.8	+8		/	-0.3	+6.1

From the statistics of Table 4, we are able to cut more than 50% of error rate, and also reduce the mismatch between different stress conditions, with the introduction of the neural network based speaker embedding system and the proposed v-TL and CL-TL transfer learning strategies.

6. Conclusions

In this study, we investigated transfer learning for a neural network based speaker verification task with domain mismatch. Compared with the conventional i-Vector based method, it is easier and more flexible to apply transfer learning with neural networks. Two alternative strategies (i.e., V-TL and CL-TL) are explored for data selection in transfer learning. Both methods are proved to be effective, CL-TL achieved the best performance, which improved EER from 18.7% to 8.5% compared to the i-Vector/PLDA baseline. We provide experimental studies that better initial performance, faster speed of convergence, better final performance can be achieved with transfer leaning in the neural network based speaker verification framework.

7. References

[1] G. Liu and J. H. L. Hansen, "An investigation into back-end advancements for speaker recognition in multi-session and noisy enrollment scenarios," *IEEE/ACM Trans. on Audi., Spee. and Lang. Proc.*, vol. 22, no. 12, pp. 1978–1992, 2014.

[2] P. Matějka, O. Glembek, O. Novotnỳ, O. Plchot, F. Grézl, L. Burget, and J. H. Cernockỳ, "Analysis of DNN approaches to speaker identification," in *IEEE ICASSP*, 2016.

[3] S.O. Sadjadi, J. Pelecanos, and S. Ganapathy, "The IBM speaker recognition system: Recent advances and error analysis," in *ISCA INTERSPEECH*, 2016.

[4] C. Zhang, S. Ranjan, M. K. Nandwana, Q. Zhang, A. Misra, G. Liu, F. Kelly, and J. H. L Hansen, "Joint information from nonlinear and linear features for spoofing detection: an i-vector/dnn based approach," in *IEEE ICASSP*, 2016, pp. 5035–5039.

[5] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *IEEE ICASSP*, 2014.

[6] Patrick Kenny, "Bayesian speaker verification with heavy-tailed priors.," in *ISCA Odyssey*, 2010, p. 14.

[7] S. Shum, D. A Reynolds, D. Garcia-Romero, and A. McCree, "Unsupervised clustering approaches for domain adaptation in speaker recognition systems," in *ISCA Odyssey*, 2014.

[8] D. Garcia-Romero, A. McCree, S. Shum, N. Brummer, and C. Vaquero, "Unsupervised domain adaptation for i-vector speaker recognition," in *ISCA Odyssey*, 2014.

[9] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. Greenberg, D. Reynolds, E. Singer, L. Mason, and J. Hernandez-Cordero, "The 2016 NIST speaker recognition evaluation," in *ISCA INTERSPEECH*, 2017.

[10] C. Zhang, F. Bahmaninezhad, S. Ranjan, C. Yu, N. Shokouhi, and J. H. L. Hansen, "UTD-CRSS systems for 2016 NIST speaker recognition evaluation," in *ISCA INTERSPEECH17*, 2017.

[11] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Audi., Spee., and Lang. Proc.*, vol. 19, no. 4, pp. 788–798, 2011.

[12] Q. Zhang and J. H. L. Hansen, "Language/dialect recognition based on unsupervised deep learning," *IEEE/ACM Trans. on Audi., Spee., and Lang. Proc.*, vol. 26, no. 5, pp. 873–882, 2018.

[13] G. Hinton, L. Deng, D. Yu, G. E Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Sign. Proc. Mage.*, vol. 29, no. 6, pp. 82–97, 2012.

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

[15] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances," in *ISCA INTERSPEECH*, 2017.

[16] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with flexibility in utterance duration," in *IEEE ASRU*, 2017.

[17] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network based speaker embedding for end-to-end speaker verification," in *IEEE SLT*, 2016.

[18] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *IEEE CVPR*, 2014.

[19] C. Zhang, G. Liu, C. Yu, and J. H. L. Hansen, "I-vector based physical task stress detection with different fusion strategies," in *ISCA INTERSPEECH*, 2015.

[20] T. Hasan, S. O. Sadjadi, G. Liu, N. Shokouhi, H. Bořil, and J. H. L. Hansen, "CRSS systems for 2012 NIST speaker recognition evaluation," in *IEEE ICASSP*, 2013.

[21] S. Ranjan, A. Misra, and J. H. L. Hansen, "Curriculum learning based probabilistic linear discriminant analysis for noise robust speaker recognition," in *ISCA INTERSPEECH*, 2017.

[22] 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 ASRU*, 2011.

[23] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE CVPR*, 2015.

[24] C. Zhang, C. Yu, and J. H.L. Hansen, "An investigation of deep-learning frameworks for speaker verification antispoofing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 684–694, 2017.

[25] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Advances in neural information processing systems*, 2014.

[26] V. Varadarajan, J. H. L. Hansen, and I. Ayako, "Ut-scope–a corpus for speech under cognitive/physical task stress and emotion," in *Proc. of LREC Workshop en Corpora for Research on Emotion and Affect, Genoa*, 2006.

[27] C. Zhang, K. Koishida, and J. H. L. Hansen, "Text-independent speaker verification based on triplet convolutional neural network embeddings," *To appear in IEEE/ACM Trans. on Audi., Spee. and Lang. Proc.*, 2018.

[28] S. Ruder and B. Plank, "Learning to select data for transfer learning with bayesian optimization," in *Proc. EMNLP*, 2017.