DOMAIN ADVERSARIAL TRAINING FOR IMPROVING KEYWORD SPOTTING PERFORMANCE OF ESL SPEECH

Jingyong Hou[1], *Pengcheng Guo*[1], *Sining Sun*[1], *Frank K. Soong*[2], *Wenping Hu*[2], *Lei Xie*[1*]

[1]School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, China
[2]Microsoft Research Asia, Beijing, China

{jyhou, pcguo, snsun, lxie}@nwpu-aslp.org, {wenh, frankkps}@microsoft.com

ABSTRACT

A second language (L2) learner usually cannot speak L2 well in both pronunciations and forming-of-words. Hence his/her L2 speech cannot be well recognized by a recognizer trained with native data. Domain adversarial training (DAT), capable of reducing the acoustic mismatch between training and testing, can be useful for improving speech recognition of L2 learners. To get around the ungrammatical L2 speech in scenario-based conversation training, keyword spotting (KWS) is an effective solution by relaxing the language model constraint in decoding. On the acoustic pronunciation side, DAT is investigated in this study for training a neural net-based acoustic model. DAT model is trained with both native and English as second language (ESL) learners' speech to extract more invariant features from native to ESL speech by equalizing their intrinsic difference. The model is jointly optimized for improved senone classification in training. Testing on ESL learners' speech and native English, the DAT model improves recognition performance which is comparable to jointly trained multi-condition model but significantly improves the performance of native speech recognition. In KWS, DAT shows a consistent better performance than the multi-condition training. The improved performance of proposed model is also obtained without increasing its computation complexity or the model size.

Index Terms— Domain adversarial training, CALL, ESL, ASR, Keyword spotting

1. INTRODUCTION

Computer assisted language learning (CALL) system has become a convenient and effective tool for learning a new language. For better user experience of CALL, it is highly desirable to customize the learning program for each learner, based upon his/her progress and the data accumulated over the learning period. In a scenario-based conversation practice, a second language (L2) learner can benefit more from the interactive conversations. However, it is difficult to continue the conversation if a decent speech recognizer is not available. To recognize an L2 learner's speech well is therefore crucial to the success of a CALL system in L2 learning.

However, in CALL applications, good speech recognition performance is not guaranteed in conventional speech recognition system, as the acoustic model (AM) is trained with speech data of native speakers and the n-gram language model (LM) is trained with the common usage of the words and word relation statistics. Both do not match well with an L2 learner's sentence. Keyword spotting (KWS) is a convenient choice to relax the LM in decoding, e.g. a lattice-based (or phoneme/word graph-based) KWS [1, 2, 3] is the choice of this study. The lattice can be generated with a large vocabulary continuous speech recognition (LVCSR) system first, followed with a scoring process. Word posteriors [4] are used to evaluate whether a keyword is included in an input utterance.

Unlike ordinary KWS, recognizing L2 learners' speech and detecting keywords from it have many challenges due to possibly his/her pronunciation deficiency and adverse acoustic environments at sound pickup. Consequently, the L2 learners' speech can be highly different from the speaker independent, continuous speech databases recorded by native speakers. There are ways to reduce the acoustic mismatches between them. The most direct ones are multi-condition training [5] and transfer learning [6, 7]. In multi-condition training, data from different distributions are mixed to train a new model which usually works well in many scenarios. Transfer learning can adjust a well-trained model with respect to a different distribution of a new dataset. However, the performance on data with the original distribution will be degraded in general for transfer learning trained model.

Recently, Ganin et al. [8, 9] proposed an unsupervised domain adversarial training (DAT) to tackle the data mismatch problem by learning domain-invariant features. DAT has achieved the state-of-the-art results for a few unsupervised domain adaptation tasks in computer visions [10, 11]. Sun et al. [12] and Wang et al. [13] have successfully applied it to

This work is supported by the National Natural Science Foundation of China (Grant No.61571363). We also would like to thank mTutor team from Microsoft Research Asia for collecting the original Xiaoying database used in this paper.

*Lei Xie is the corresponding author

robust speech and speaker recognition. DAT is also applied to supervised learning, e.g. face recognition [14, 15], noisy speech recognition [16] and accented speech recognition [17].

In this paper, we apply supervised DAT to acoustic model (AM) training with data collected from both English as second language (ESL) learners and native English speakers. Speech recognition and keyword spotting are used to evaluate the proposed approach. We have investigated DAT, transfer learning and multi-condition training in training time-delay neural network (TDNN) [18] AMs with native English speech and/or ESL speech. The experimental results show that multi-condition and DAT are better than other methods in dealing with speech with different pronunciation deficiencies. Testing on ESL learners speech with different proficiency levels and native English, the DAT model improves recognition performance which is comparable to jointly trained multi-condition model but significantly improves the performance of native speech recognition with an 18.5% relative WER reduction. For keyword spotting, DAT model yields consistent better performances than the multi-condition model for different lattice beam width settings on all test sets.

2. ACOUSTIC MODELING FOR ESL LEARNERS' SPEECH USING DAT

2.1. Domain-invariant feature representation

We have two datasets from two different domains: S_1 from native English speech domain and S_2 from first language (L1) Chinese non-native English speech domain. $S_1 = \{x_i, y_i\}_{i=1}^{|S_1|}$ and $S_2 = \{x_i, y_i\}_{i=1}^{|S_2|}$, where $x_i \in X$ is speech feature and $y_i \in Y$ is the corresponding HMM senone label. Normally, an AM learns the distribution $D(x, y)$ on $X \otimes Y$. However, the distribution $D(x, y)$ may be different between two different domains of data sets. Assume there are two distributions $D_1(x, y)$ and $D_2(x, y)$ corresponding to two different domains, these two distributions are different mostly because of the different marginal distribution of $D_1(x)$ and $D_2(x)$. The basic idea of DAT is to learn a feature mapping, $f = F(x)$ to map the x to a domain-invariant space V. In space V, the mismatch between the two domains is reduced. Then the AM only needs to learn a more uniform distribution $Q(f, y)$ on $V \otimes Y$ and the AM could recognize them better.

As shown in Fig. 1, the DAT network consists of three sub networks: $M_1(x, \theta_1)$, $M_2(f, \theta_2)$ and $M_3(f, \theta_3)$. The $M_1(x, \theta_1)$ here is called a feature extractor, which takes speech x from different domains as input and outputs the domain-invariant feature f, where θ_1 is the parameters of M_1. $M_2(f, \theta_2)$ is a senone classification network whose input is f and its parameters are represented by θ_2. $M_3(f, \theta_3)$ is a domain classification network with f as input and θ_3 as parameters.

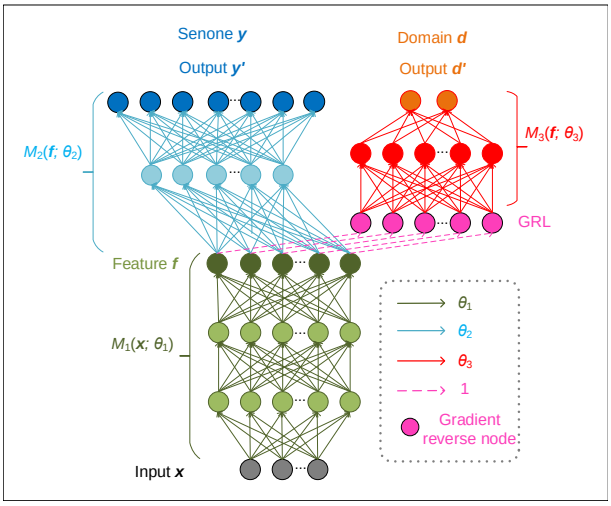

Fig. 1. *Domain adversarial training*

2.2. DAT via back propagation

Assuming there are N training examples in each mini-batch, the object function is defined as follows:

$$E(\theta_1, \theta_2, \theta_3) = \sum_{i=1}^{N} L_1(M_2(M_1(x_i; \theta_1); \theta_2), y_i) - \lambda \sum_{i=1}^{N} L_2(M_3(M_1(x_i; \theta_1); \theta_3), d_i) \quad (1)$$

Here, $L_1(.,.)$ is the cross-entropy loss function for senone classification, d_i is the domain label of i-th training sample, $L_2(.,.)$ is the cross-entropy loss function for domain classification. Unlike [17], non-speech parts in the data are also used for training, since in our task, there is not only accent mismatch but also channel mismatch between the two domains. λ is a hyper parameter of DAT and is positive.

Then DAT can be viewed as two optimization problems:

$$(\theta_1, \theta_2) = \arg\min_{(\theta_1, \theta_2)} E(\theta_1, \theta_2, \theta_3) \quad (2)$$

$$\theta_3 = \arg\max_{\theta_3} E(\theta_1, \theta_2, \theta_3) \quad (3)$$

Above equation 2, 3 mean that we want to optimize our networks M_1 and M_2 so that the network can discriminate different senones while making the network generated features f that can not be determined by M_3 which domain it is from. At the same time, the M_3 is optimized to distinguish which domain the data comes from. The model parameters can be updated with back propagation as stochastic gradient descent:

$$\theta_1 \leftarrow \theta_1 - \alpha \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\partial L_1(y_i', y_i)}{\partial \theta_1} - \lambda \frac{\partial L_2(d_i', d_i)}{\partial \theta_1} \right) \quad (4)$$

$$\theta_2 \leftarrow \theta_2 - \alpha \frac{1}{N} \sum_{i=1}^{N} \frac{\partial L_1(y_i^{'}, y_i)}{\partial \theta_2} \quad (5)$$

$$\theta_3 \leftarrow \theta_3 - \alpha \frac{1}{N} \sum_{i=1}^{N} \frac{\partial L_1(d_i^{'}, d_i)}{\partial \theta_3} \quad (6)$$

where α is the learning rate, $y_i^{'}$ and $d_i^{'}$ are the predicted senone and domain labels, respectively. In practice, to achieve this, we have implemented a gradient reverse layer (GRL) [8] as shown in Fig. 1. This layer will scale the gradient by $-\lambda$ when the gradient propagates backward while do nothing during forward propagation.

3. EXPERIMENTAL SETUP

3.1. Database

Same as our previous work [19], two databases are used to train the AMs. One is a native speakers' speech database, i.e. switchboard (SWBD). The other one is an ESL speech database, denoted as 'Xiaoying'. They have 282 hrs and 160 hrs respectively.

We collect an extra Xiaoying data set for testing different AMs, 2691 utterances in total and by non-native English learners with different proficiencies. There is no overlapping of sentence transcription and speaker between the training and testing Xiaoying databases. To test the sensitivity of our ASR and KWS systems to users' English proficiency, we divide the above testing database into three groups, i.e., G1, G2 and G3 [19], according to its utterance level pronunciation score evaluated by the pronunciation scoring algorithm described in [20]. Different from [19], in this work, 500 native speech utterances are also collected for testing, denoted as G4. G4, collected from native speakers (read speech with very standard pronunciation), contains all the sentences in above three test sets. The detail information of four test groups is listed in Table 1. To better distinguish different search data sets, there are discontinuities in the score ranges for the four groups, e.g., the utterances with scores between 31-39 and 56-64 are discarded. 60 keywords, identical to [19], are selected for the KWS task and each word contains at least 2 syllables or 5 phonemes. To show the difficulty of searching keywords in each test group, the priors are calculated as the averaged occurrences of each keyword in each group of testing set (last column in Table 1). More detailed information about the Xiaoying data and selected keywords can be found in section 3.1 of [19].

Table 1. *Details of four search data sets*

	score range	speaker type	#utt	avg length (s)	prior (%)
G1	15-30	non-native	866	4.7	0.64
G2	40-55	non-native	953	5.1	0.81
G3	65-80	non-native	872	4.8	0.74
G4		native	500	4.9	0.90

3.2. Acoustic models

We have tried five different methods for AM training:

- **SWBD**: only native data SWBD used;
- **Xiaoying**: only non-native data Xiaoying used;
- **Transfer learning**: trained with SWBD, fine tuned with Xiaoying;
- **Multi-condition**: trained with SWBD and Xiaoying mix;
- **DAT**: similar to multi-condition training, SWBD and Xiaoying are assigned different domain labels used for DAT.

For the first four methods, we train TDNN AMs with 6 hidden layers, each with 1024 nodes and ReLU activation function. From the first hidden layer to the output layer, we splice frames at offsets of {-2,-1,0,1,2}, {-1,2}, {-3,3}, {-7,2}, $\{0\}$ and $\{0\}$. 36-dimensional filter bank features are used as network input. The output consisted of 8,816 softmax nodes, corresponding to the # senones of the AM. Kaldi [21] nnet3 training method is used to train our network. For the DAT network, the acoustic part (M_1+M_2) is exactly the same as the above network. The difference is that we take the output of the 5-th hidden layer of the network as input to M_3 (a hyper-parameter set to 5, by following [12, 17]), which is a fully connected network with two hidden layers, each with 512 ReLU activation nodes. The G_3 has two softmax output nodes, for the two different domains. The hyper-parameter λ in section 2.2 is set to 0.5 (we have tried different settings, from 0.1 to 0.8, and 0.5 got the best results). It is important to note that the same two datasets were used for training the multi-condition model and the DAT model. We disconnect M_3 from the DAT network in inference (testing), hence the computational complexities remain the same. A 3-gram language model trained with the SWBD transcriptions is used for decoding.

3.3. Alignment of Xiaoying data

We first aligned the SWBD data with the GMM-HMM model and trained a TDNN AM with the SWBD data. Then we aligned the Xiaoying data with the SWBD TDNN model. Finally, we used above aligned SWBD and Xiaoying data to train a multi-condition TDNN model and used this model to realign all training data. All models mentioned in Sec. 3.2 are trained with the realigned data.

WER is used to evaluate the performance of different ASR systems. To evaluate the performance of different KWS systems, two metrics are used, i.e., the Mean Average Precision (MAP) and Mean Precision at N (MP@N), where the AP and P@N are defined as:

- **AP**: Averaged precision at the true hit utterance position over all ranked utterances;

Table 2. *WERs (%) of different LVCSR systems, numbers in the parentheses (last column) are relative improvements (%) compared to the multi-condition method.*

	SWBD	Xiaoying	Transfer learning	Multi-condition	DAT ($\lambda = 0.5$)
G1	99.99	56.29	54.30	**52.91**	53.96 (-1.98)
G2	89.39	38.25	36.41	**35.70**	35.83 (-0.04)
G3	70.02	28.68	26.59	25.57	**25.01** (2.19)
G4	14.27	32.11	25.17	15.77	**12.85** (**18.52**)

- **P@N**:Top N precision of the returned ranked list; N is the number of utterances that contain the keyword.

4. EXPERIMENTAL RESULTS

4.1. Speech recognition results of different AMs

Table 2 shows speech recognition results of different AMs on all test sets. As the level of pronunciation (rated by the corresponding pronunciation score) improves from G1 to G4, WER decreases on all ASR systems. When the model was trained with the SWBD data only, the performance of G4, the native test set is good. However, poorer WERs are obtained on the dataset from G1 to G3, due to the corresponding mismatch level with the native AM model. For the model trained by Xiaoying data, there is a considerable improvement on G1 to G3 over the SWBD model but deteriorates the G4 performance significantly since Xiaoying data contains mostly mismatched ESL speech. Thirdly, the transfer learning model, multi-condition model and DAT model outperformed the Xiaoying model on G1 to G3 datasets, and they were comparable to each other on the three test sets. Compared with the multi-condition model, DAT model achieved 2.19% and 18.52% relative WER improvements on G3 and G4 while minimally degraded performance on G1 and G2 (relatively 1.98% and 0.04% increasing of WER).

4.2. Keyword search results of different acoustic models

Similarly, we carried out KWS experiments with different ASR systems on the four test sets, and the results in table 3 were similar to the recognition experiments (systems with high recognition accuracy performed well on KWS tasks, too). We focus on the result differences between the multi-condition model and the DAT model since they are the two best performed models in speech recognition. Experimental results showed that the DAT model performed better than multi-condition model in KWS on all test sets. Especially, in the MP@N metric, we observed more improvement on all test sets.

In the last experiment, we compared DAT model and multi-condition model with different lattice beam settings. Lattice beam plays a key role in the scoring process of KWS by controlling the size of the data to store and transmit, hence affects the response latency of keyword spotting, a crucial indicator in real-time applications. From Fig. 2, we can see that with the increase of lattice beam width, the performance of DAT model and multi-condition model on each dataset is improved, and the results saturate at a beam width of 9. Moreover, under all tested beam width, the DAT model performs consistently better than the multi-condition model. We also observed that the gap between them are larger at smaller lattice beam width, indicating that DAT tended to outperform with a small lattice beam setting.

Table 3. *Performance comparison of different KWS systems*

	SWBD	Xiaoying	Transfer learning	Multi-condition	DAT ($\lambda = 0.5$)
			MAPs		
G1	0.129	0.749	0.785	0.796	**0.802**
G2	0.291	0.865	0.870	**0.902**	**0.903**
G3	0.631	0.937	**0.956**	**0.953**	**0.958**
G4	0.990	0.915	0.944	**0.988**	**0.989**
			MP@Ns		
G1	0.128	0.711	0.752	0.746	**0.760**
G2	0.329	0.826	0.851	0.868	**0.877**
G3	0.605	0.915	**0.931**	**0.927**	**0.930**
G4	0.985	0.893	0.926	0.979	**0.985**

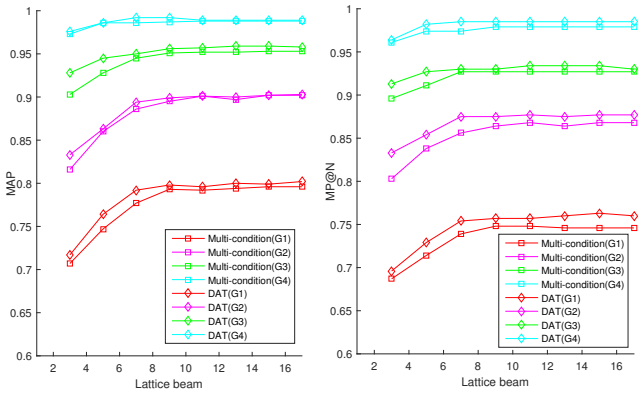

Fig. 2. *Performance comparison of DAT and multi-condition based KWS systems with different lattice beam*

5. CONCLUSIONS

We propose DAT to recognize L2 learners' speech in both speech recognition and keyword spotting applications. DAT adjusts the model parameters of three modules jointly in AM training, including: a feature extraction module for finding domain-invariant features; a discriminative senone classification module for improved phonetic recognition; and a domain classification module to classify which domain of the input data is from. DAT uses mixed-domain data, i.e., native speakers' speech and ESL speech from Microsoft mTutor English learning on-line service. Compared with other baseline systems trained on the same data sets, DAT model yields a better performance by learning the domain-invariant features for highly discriminative phonetic recognition.

6. REFERENCES

[1] Kenney Ng and Victor W Zue, "Subword-based approaches for spoken document retrieval," *Speech Communication*, vol. 32, no. 3, pp. 157–186, 2000.

[2] Murat Saraclar, "Lattice-based search for spoken utterance retrieval," in *Proc. HLT-NAACL*, 2004, pp. 129–136.

[3] Ciprian Chelba and Alex Acero, "Position specific posterior lattices for indexing speech," in *Proc. ACL*, 2005, pp. 443–450.

[4] Hui Jiang, "Confidence measures for speech recognition: A survey," *Speech communication*, vol. 45, no. 4, pp. 455–470, 2005.

[5] Michael L Seltzer, Dong Yu, and Yongqiang Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. ICASSP*, 2013, pp. 7398–7402.

[6] Rich Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[7] Georg Heigold, Vincent Vanhoucke, Alan Senior, Patrick Nguyen, MarcAurelio Ranzato, Matthieu Devin, and Jeffrey Dean, "Multilingual acoustic models using distributed deep neural networks," in *Proc. ICASSP*, 2013, pp. 8619–8623.

[8] Yaroslav Ganin and Victor Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. ICML*, 2015, pp. 1180–1189.

[9] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2017.

[10] Han Zhao, Shanghang Zhang, Guanhang Wu, Joo P Costeira, Jos M. F Moura, and Geoffrey J Gordon, "Multiple source domain adaptation with adversarial training of neural networks," *arXiv preprint arXiv:1705.09684*, 2017.

[11] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell, "Adversarial discriminative domain adaptation," in *Proc. CVPR*, 2017, pp. 7167–7176.

[12] Sining Sun, Binbin Zhang, Lei Xie, and Yanning Zhang, "An unsupervised deep domain adaptation approach for robust speech recognition," *Neurocomputing*, vol. 257, pp. 79–87, 2017.

[13] Qing Wang, Wei Rao, Sining Sun, Lei Xie, Eng Siong Chng, and Haizhou Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," in *Proc. ICASSP*, 2018, pp. 4889–4893.

[14] Yujia Li, Kevin Swersky, and Richard Zemel, "Learning unbiased features," *arXiv preprint arXiv:1412.5244*, 2014.

[15] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel, "The variational fair autoencoder," *arXiv preprint arXiv:1511.00830*, 2015.

[16] Yusuke Shinohara, "Adversarial multi-task learning of deep neural networks for robust speech recognition.," in *Proc. INTERSPEECH*, 2016, pp. 2369–2372.

[17] Sining Sun, Ching-Feng Yeh, Mei-Yuh Hwang, Mari Ostendorf, and Lei Xie, "Domain adversarial training for accented speech recognition," in *Proc. ICASSP*, 2018, pp. 4854–4858.

[18] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. INTERSPEECH*, 2015, pp. 3214–3218.

[19] Jingyong Hou, Wenping Hu, Frank K. Soong, and Lei Xie, "A refined query-by-example approach to spoken term detection on ESL learners' speech," in *Proc. ISCSLP*, 2018.

[20] Wenping Hu, Yao Qian, and Frank K Soong, "A new DNN-based high quality pronunciation evaluation for Computer-Aided Language Learning (CALL).," in *Proc. INTERSPEECH*, 2013, pp. 1886–1890.

[21] 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 *Proc. ASRU*, 2011, number EPFL-CONF-192584.