SPREADER-INVARIENT TRAINING VIA ADVERSARIAL LEARNING

Zhong Meng, Jinyu Li, Zhuo Chen, Yong Zhao, Vadim Mazalov, Yifan Gong, Biing-Hwang (Fred) Juang

1 Microsoft AI and Research, Redmond, WA, USA
2 Georgia Institute of Technology, Atlanta, GA, USA

ABSTRACT

We propose a novel adversarial multi-task learning scheme, aiming at actively curtailing the inter-talker feature variability while maximizing its senone discriminability so as to enhance the performance of a deep neural network (DNN) based ASR system. We call the scheme speaker-invariant training (SIT). In SIT, a DNN acoustic model and a speaker classifier network are jointly optimized to minimize the senone (tied triphone state) classification loss, and simultaneously mini-maximize the speaker classification loss. A speaker-invariant and senone-discriminative deep feature is learned through this adversarial multi-task learning. With SIT, a canonical DNN acoustic model with significantly reduced variance in its output probabilities is learned with no explicit speaker-independent (SI) transformations or speaker-specific representations used in training or testing. Evaluated on the CHiME-3 dataset, the SIT achieves 4.99% relative word error rate (WER) improvement over the conventional SI acoustic model. With additional unsupervised speaker adaptation, the speaker-adapted (SA) SIT model achieves 4.86% relative WER gain over the SA SI acoustic model.

Index Terms— speaker-invariant training, adversarial learning, speech recognition, deep neural networks

1. INTRODUCTION

The deep neural network (DNN) based acoustic models have been widely used in automatic speech recognition (ASR) and have achieved extraordinary performance improvement[1][2]. However, the performance of a speaker-independent (SI) acoustic model trained with speech data from a large number of speakers is still affected by the spectral variations in each speech unit caused by the inter-speaker variability. Therefore, speaker adaptation methods are widely used to boost the recognition system performance [3][4][5][6][7][8][9][10][11][12][13].

Recently, adversarial learning has captured great attention of deep learning community given its remarkable success in estimating generative models [14]. In speech, it has been applied to noise-robust [15][16][17][18][19] and conversational ASR [20] using gradient reversal layer [21] or domain separation network [22]. Inspired by this, we propose speaker-invariant training (SIT) via adversarial learning to reduce the effect of speaker variability in acoustic modeling. In SIT, a DNN acoustic model and a DNN speaker classifier are jointly trained to simultaneously optimize the primary task of minimizing the senone classification loss and the secondary task of mini-maximizing the speaker classification loss. Through this adversarial multi-task learning procedure, a feature extractor is learned as the bottom layers of the DNN acoustic model that maps the input speech frames from different speakers into speaker-invariant and senone-discriminative deep hidden features, so that further senone classification is based on representations with the speaker factor already normalized out. The DNN acoustic model with SIT can be directly used to generate word transcription for unseen test speakers through one-pass online decoding. On top of the SIT DNN, further adaptation can be performed to adjust the model towards the test speakers, achieving even higher ASR accuracy.

We evaluate SIT with ASR experiments on CHiME-3 dataset, the SIT DNN acoustic model achieves 4.99% relative WER improvement over the baseline SI DNN. Further, SIT achieves 4.86% relative WER gain over the SI DNN when the same unsupervised speaker adaptation process is performed on both models. With t-distributed stochastic neighbor embedding (t-SNE) [23] visualization, we show that, after SIT, the deep feature distributions of different speakers are well aligned with each other, which demonstrates the strong capability of SIT in reducing speaker-variability.

2. RELATED WORK

Speaker-adaptive training (SAT) is proposed to generate canonical acoustic models coupled with speaker adaptation. For Gaussian mixture model (GMM)-hidden Markov model (HMM) acoustic model, SAT applies unconstrained [24] or constrained [25] model-space linear transformations that separately model the speaker-specific characteristics and are jointly estimated with the GMM-HMM parameters to maximize the likelihood of the training data. Cluster-adaptive training (CAT) [26] is then proposed to use a linear interpolation of all the cluster means as the mean of the particular speaker instead of a single cluster as representative of a particular speaker. However, SAT of GMM-HMM needs to have two sets of models, the SI model and canonical model. During testing, the SI model is used to generate the first pass decoding transcription, and the canonical model is combined with speaker-specific transformation to adapt to the new speaker.

For DNN-HMM acoustic model, CAT [27] and multi-basis adaptive neural networks [28] are proposed to represent the weight and/or the bias of the speaker-dependent (SD) affine transformation in each hidden layer of a DNN acoustic model as a linear combination of SI bases, where the combination weights are low-dimensional SD speaker representations. The canonical SI bases with reduced variances are jointly optimized with the SD speaker representations during the SAT to minimize the cross-entropy loss. During unsupervised adaptation, the test speaker representations are re-estimated using alignments from the first-pass decoding of the test data with SI DNN as the supervisions and are used in the second-pass decoding to generate the transcription. Factorized hidden layer [13] is
similar to [12][7], but includes SI DNN weights as part of the linear combination. In [5], SD speaker codes are transformed by a set of SI matrices and then directly added to the biases of the hidden-layer affine transformations. The speaker codes and SI transformations are jointly estimated during SAT. For these methods, two passes of decoding are required to generate the final transcription in unsupervised adaptation setup, which increases the computational complexity of the system.

In [6][3], an SI adaptation network is learned to derive speaker-normalized features from i-vectors to train the canonical DNN acoustic model. The i-vectors for the test speakers are then estimated and used for decoding after going through the SI adaptation network. In [20], a reconstruction network is trained to predict the input i-vector given the speech feature and its corresponding i-vector are at the input of the acoustic model. The mean-squared error loss of the i-vector reconstruction and the cross-entropy loss where $\theta$ is the set of all senones modeled by the acoustic model.

To perform SIT, we need a sequence of speech frames $X = \{x_1, \ldots, x_N\}$, a sequence of senones labels $Y = \{y_1, \ldots, y_N\}$ aligned with $X$ and a sequence of speaker labels $S = \{s_1, \ldots, s_N\}$ aligned with $X$. The goal of SIT is to reduce the variances of hidden and output units distributions of the DNN acoustic model that are caused by the inherent inter-speaker variability in the speech signal. To achieve speaker-robustness, we learn a speaker-invariant and senone-discriminative deep hidden feature in the DNN acoustic model through adversarial multi-task learning and make senone posterior predictions based on the learned deep feature. In order to do so, we view the first few layers of the acoustic model as a feature extractor network $M_f$ with parameters $\theta_f$ that maps $X$ from different speakers to deep hidden features $F = \{f_1, \ldots, f_N\}$ (see Fig. 1) and the upper layers of the acoustic model as a senone classifier $M_s$ with parameters $\theta_s$ that maps the intermediate features $F$ to the senone posterior $p_s(a|f; \theta_s)$, $a \in Q$ as follows:

$$M_f(f_i) = M_f(M_f(x_i)) = p_y(q|x; \theta_f, \theta_y)$$  \hspace{1cm} (1)

where $Q$ is the set of all senones modeled by the acoustic model.

We further introduce a speaker classifier network $M_s$ which maps the deep features $F$ to the speaker posteriors $p_s(a|f; \theta_s)$, $a \in A$ as follows:

$$M_s(M_f(x_i)) = p_s(a|x_i; \theta_s, \theta_f)$$ \hspace{1cm} (2)

This minimax competition will first increase the discriminativity of $M_s$ and the speaker-variability of the features generated by $M_f$, and will eventually converge to the point where $M_f$ generates extremely confusing features that $M_s$ is unable to distinguish.

At the same time, we want to make the deep features senone-discriminative by minimizing the cross-entropy loss between the predicted senone posteriors and the senone labels as follows:

$$L_{\text{senone}}(\theta_f, \theta_y) = - \sum_{i=1}^{N} \log p_y(y_i|x_i; \theta_f, \theta_y)$$

$$= - \sum_{i=1}^{N} \sum_{q \in Q} \mathbb{1}[q = y_i] \log M_y(M_y(x_i))$$  \hspace{1cm} (4)

In SIT, the acoustic model network and the condition classifier network are trained to jointly optimize the primary task of senone classification and the secondary task of speaker classification with an adversarial objective function. Therefore, the total loss is constructed as

$$L_{\text{total}}(\theta_f, \theta_y, \theta_s) = L_{\text{senone}}(\theta_f, \theta_y) - \lambda L_{\text{speaker}}(\theta_s, \theta_f)$$  \hspace{1cm} (5)
where $\lambda$ controls the trade-off between the senone loss and the speaker classification loss in Eq. (4) and Eq. (3) respectively.

We need to find the optimal parameters $\theta_y$, $\theta_f$ and $\theta_s$ such that

$$\hat{\theta}_f, \hat{\theta}_y = \min_{\theta_f, \theta_y} \ell_{\text{total}}(\hat{\theta}_f, \hat{\theta}_y, \hat{\theta}_s)$$

$$\hat{\theta}_s = \max_{\theta_s} \ell_{\text{total}}(\hat{\theta}_f, \hat{\theta}_y, \hat{\theta}_s)$$

The parameters are updated as follows via back propagation with stochastic gradient descent (SGD):

$$\theta_f \leftarrow \theta_f - \mu \left[ \frac{\partial \ell_{\text{senone}}}{\partial \theta_f} - \lambda \frac{\partial \ell_{\text{speaker}}}{\partial \theta_f} \right]$$

$$\theta_s \leftarrow \theta_s - \mu \frac{\partial \ell_{\text{speaker}}}{\partial \theta_s}$$

$$\theta_y \leftarrow \theta_y - \mu \frac{\partial \ell_{\text{senone}}}{\partial \theta_y}$$

where $\mu$ is the learning rate.

Note that the negative coefficient $-\lambda$ in Eq. (8) induces reversed gradient that maximizes $\ell_{\text{speaker}}(\theta_f, \theta_s)$ in Eq. (7) and makes the deep feature speaker-invariant. For easy implementation, gradient reversal layer is introduced in [21], which acts as an identity transform in the forward propagation and multiplies the gradient by $-\lambda$ during the backward propagation.

The optimized network consisting of $M_f$ and $M_s$ is used as the SIT acoustic model for ASR on test data.

### 4. EXPERIMENTS

In this work, we perform SIT on a DNN-hidden Markov model (HMM) acoustic model for ASR on CHiME-3 dataset.

#### 4.1. CHiME-3 Dataset

The CHiME-3 dataset is released with the 3rd CHiME speech Separation and Recognition Challenge [27], which incorporates the Wall Street Journal corpus sentences spoken in challenging noisy environments, recorded using a 6-channel tablet based microphone array.

CHiME-3 dataset consists of both real and simulated data. The real speech data was recorded in five real noisy environments (on buses (BUS), in cafés (CAF), in pedestrian areas (PED), at street junctions (STR) and in booth (BTH)). To generate the simulated data, the clean speech is first convolved with the estimated impulse response of the environment and then mixed with the background noise separately recorded in that environment [28]. The noisy training data consists of 1999 real noisy utterances from 4 speakers, and 7138 simulated noisy utterances from 83 speakers in the WSJ0 SI-84 training set recorded in 4 noisy environments. There are 3280 utterances in the development set including 410 real and 410 simulated utterances for each of the 4 environments. There are 2640 utterances in the test set including 330 real and 330 simulated utterances for each of the 4 environments. The speakers in training set, development set and the test set are mutually different (i.e., 12 different speakers in the CHiME-3 dataset). The training, development and test data sets are all recorded in 6 different channels.

In the experiments, we use 9137 noisy training utterances in the CHiME-3 dataset as the training data. The real and simulated development data in CHiME-3 are used as the test data. Both the training and test data are far-field speech from the 5th microphone channel.

#### 4.4. Visualization of Deep Features

We randomly select two male speakers and two female speakers from the noisy training set and extract speech frames aligned with the phoneme “ah” for each of the four speakers. In Figs. 2 and 3, we visualize the deep features $F$ generated by the SI and SIT DNN acoustic models when the “ah” frames of the four speakers are given as the input using t-SNE. In Fig. 2, the deep feature distributions in the SI model for the male (in red and green) and female speakers (in back and blue) are far away from each other and even the distributions for the speakers of the same gender are separated from each other. While after SIT, the deep feature distributions for all the
male and female speakers are well aligned with each other as shown in Fig. 3. The significant increase in the overlap among distributions of different speakers justifies that the SIT remarkably enhances the speaker-invariance of the deep features $F$. The adversarial optimization of the speaker classification loss does not just serve as a regularization term to achieve better generalization on the test data.

In our experiment, we adapt the SI and SIT DNNs to each of the 4 speakers in the test set in an unsupervised fashion. The constrained re-training (CRT) [30] method is used for adaptation, where we re-estimate the DNN parameters of only a subset of layers while holding the remaining parameters fixed during cross-entropy training. The adaptation target (1-best alignment) is obtained through the first-pass decoding of the test data, and the second-pass decoding is performed using the SA SI and SI DNNs.

The WER results for unsupervised speaker adaptation is shown in Table 2, in which only the bottom 2 layers of the SI and SIT DNNs are adapted during CRT. The speaker-adapted (SA) SIT DNN achieves 15.46% WER which is 4.86% relatively higher than the SA SI DNN. The CRT adaptation provides 8.91% and 8.79% relative WER gains over the unadapted SI and SIT models respectively. The lower WER after speaker adaptation indicates that SIT has effectively reduced the high variance and overlap in an SI acoustic model caused by the inter-speaker variability.

In the future, we will evaluate the performance of the i-vector based speaker-adversarial multi-task learning [20] on CHiME-3 dataset and compare it with the proposed SIT. Moreover, we will perform SIT on thousands of hours of data to verify the its scalability to large dataset.
6. REFERENCES

[1] G. Hinton, L. Deng, D. Yu, et al., “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82–97, 2012.

[2] Dong Yu and Jinyu Li, “Recent progresses in deep learning based acoustic models,” IEEE/CAA Journal of Automatica Sinica, vol. 4, no. 3, pp. 396–407, 2017.

[3] G. Saon, H. SoLtau, D. Nahamoo, and M. Picheny, “Speaker adaptation of neural network acoustic models using i-vectors,” in 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Dec 2013, pp. 55–59.

[4] J. Xue, J. Li, D. Yu, M. Seltzer, and Y. Gong, “Singular value decomposition based low-footprint speaker adaptation and personalization for deep neural network,” in ICASSP, 2014, pp. 6359–6363.

[5] S. Xue, O. Abdel-Hamid, H. Jiang, L. Dai, and Q. Liu, “Fast adaptation of deep neural network based on discriminant codes for speech recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 12, pp. 1713–1725, Dec 2014.

[6] Y. Miao, H. Zhang, and F. Metze, “Speaker adaptive training of deep neural network acoustic models using i-vectors,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 11, pp. 1938–1949, Nov 2015.

[7] C. Wu and M. J. F. Gales, “Multi-basis adaptive neural network for rapid adaptation in speech recognition,” in Proc. ICASSP, April 2015, pp. 4315–4319.

[8] Y. Zhao, J. Li, and Y. Gong, “Low-rank plus diagonal adaptation for deep neural networks,” in Proc. ICASSP, 2016, pp. 5005–5009.

[9] Z. Huang, S. Siniscalchi, I. Chen, et al., “Maximum a posteriori adaptation of network parameters in deep models,” in Proc. Interspeech, 2015.

[10] Z. Huang, J. Li, S. Siniscalchi, et al., “Rapid adaptation for deep neural networks through multi-task learning,” in Interspeech, 2015.

[11] P. Swietojanski, J. Li, and S. Renals, “Learning hidden unit contributions for unsupervised acoustic model adaptation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 8, pp. 1450–1463, Aug 2016.

[12] T. Tan, Y. Qian, and K. Yu, “Cluster adaptive training for deep neural network based acoustic model,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 3, pp. 459–468, March 2016.

[13] L. Samarakoon and K. C. Sim, “Factorized hidden layer adaptation for deep neural network based acoustic modeling,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 12, pp. 2241–2250, Dec 2016.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, “Generative adversarial nets,” in NIPS, pp. 2672–2680, 2014.

[15] Yusuke Shinohara, “Adversarial multi-task learning of deep neural networks for robust speech recognition,” in INTERSPEECH, 2016, pp. 2369–2372.

[16] Dmitriy Serdyuk, KartiK Audhkhasi, Philélon Brakel, Bhuvana Ramabhadran, Samuel Thomas, and Yoshua Bengio, “Invariant representations for noisy speech recognition,” in NIPS Workshop, 2016.

[17] 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, Machine Learning and Signal Processing for Big Multimedia Analysis.

[18] Z. Meng, Z. Chen, V. Mazalov, J. Li, and Y. Gong, “Unsupervised adaptation with domain separation networks for robust speech recognition,” in Proceeding of ASRU, Dec 2017.

[19] Zhong Meng, Jinyu Li, Yifan Gong, and Biing-Hwang (Fred) Juang, “Adversarial teacher-student learning for unsupervised domain adaptation,” in Proc. ICASSP, IEEE, 2018.

[20] George Saon, Gakuto Kurata, Tom Sercu, et al., “English conversational telephone speech recognition by humans and machines,” Proc. Interspeech, 2017.

[21] Yaraslav Ganin and Victor Lempitsky, “Unsupervised domain adaptation by backpropagation,” in Proc. ICMIL, Lille, France, 2015, vol. 37, pp. 1180–1189, PMLR.

[22] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan, “Domain separation networks,” in Proc. NIPS, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds., pp. 343–351. Curran Associates, Inc., 2016.

[23] Laurens van der Maaten and Geoffrey Hinton, “Visualizing data using t-sne,” Journal of Machine Learning Research, vol. 9, no. Nov, pp. 2579–2605, 2008.

[24] Tasos Anastasakos, John McDonough, Richard Schwartz, and John Makhoul, “A compact model for speaker-adaptive training,” in Spoken Language. 1996. ICSLP 96. Proceedings., Fourth International Conference on. IEEE, 1996, vol. 2, pp. 1137–1140.

[25] M. J. F. Gales, “Maximum likelihood linear transformations for hmm-based speech recognition,” Computer Speech & Language, vol. 12, no. 2, pp. 75 – 98, 1998.

[26] M. J. F. Gales, “Cluster adaptive training of hidden markov models,” IEEE Transactions on Speech and Audio Processing, vol. 8, no. 4, pp. 417–428, Jul 2000.

[27] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third chime speech separation and recognition challenge: Dataset, task and baselines,” in Proc.ASRU, Dec 2015, pp. 504–511.

[28] T. Hori, Z. Chen, H. Erdogan, et al., “The mert/sri system for the 3rd chime challenge using beamforming, robust feature extraction, and advanced speech recognition,” in Proc.ASRU, Dec 2015, pp. 475–481.

[29] J. Li, D. Yu, J.-T. Huang, and Y. Gong, “Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM,” in Proc. SLT. IEEE, 2012, pp. 131–136.

[30] Hakan Erdogan, Tomoki Hayashi, John R Hershey, Takaaki Hori, Chiori Hori, Wei-Ning Hsu, Suyoun Kim, Jonathan Le Roux, Zhong Meng, and Shinji Watanabe, “Multi-channel speech recognition: Lstms all the way through,” in CHiME-4 workshop, 2016.