ADVERSARIAL SPEAKER ADAPTATION

Zhong Meng, Jinyu Li, Yifan Gong

Microsoft Corporation, Redmond, WA, USA
{zhme, jinyl, ygong}@microsoft.com

ABSTRACT

We propose a novel adversarial speaker adaptation (ASA) scheme, in which adversarial learning is applied to regularize the distribution of deep hidden features in a speaker-dependent (SD) deep neural network (DNN) acoustic model to be close to that of a fixed speaker-independent (SI) DNN acoustic model during adaptation. An additional discriminator network is introduced to distinguish the deep features generated by the SD model from those produced by the SI model. In ASA, with a fixed SI model as the reference, an SD model is jointly optimized with the discriminator network to minimize the senone classification loss, and simultaneously to mini-maximize the SI/SD discrimination loss on the adaptation data. With ASA, a senone-discriminative deep feature is learned in the SD model with a similar distribution to that of the SI model. With such a regularized and adapted deep feature, the SD model can perform improved automatic speech recognition on the target speaker’s speech. Evaluated on the Microsoft short message dictation dataset, ASA achieves 14.4% and 7.9% relative word error rate improvements for supervised and unsupervised adaptation, respectively, over an SI model trained from 2600 hours data, with 200 adaptation utterances per speaker.

Index Terms—adversarial learning, speaker adaptation, neural network, automatic speech recognition

1. INTRODUCTION

With the advent of deep learning, the performance of automatic speech recognition (ASR) has greatly improved [1, 2]. However, the ASR performance is not optimal when acoustic mismatch exists between training and testing [3]. Acoustic model adaptation is a natural solution to compensate for this mismatch. For speaker adaptation, we are given a speaker-independent (SI) acoustic model that performs reasonably well on the speech of almost all speakers in general. Our goal is to learn a personalized speaker-dependent (SD) acoustic model for each target speaker that achieves optimal ASR performance on his/her own speech. This is achieved by adapting the SI model to the speech of each target speaker.

The speaker adaptation task is more challenging than the other types of domain adaptation tasks in that it has only access to very limited adaptation data from the target speaker and has no access to the source domain data, e.g., speech from other general speakers. Moreover, a deep neural network (DNN) based SI model, usually with a large number of parameters, can easily get overfitted to the limited adaptation data. To address this issue, transformation-based approaches are introduced in [4, 5] to reduce the number of learnable parameters by inserting a linear network to the input, output or hidden layers of the SI model. In [6, 7], the trainable parameters are further reduced by singular value decomposition (SVD) of weight matrices of a neural network and perform adaptation on an inserted square matrix between the two low-rank matrices. Moreover, i-vector [8] and speaker-code [9, 10] are widely used as auxiliary features to a neural network for speaker adaptation. Further, regularization-based approaches are proposed in [11, 12, 13, 14] to regularize the neuron output distributions or the model parameters of the SD model such that it does not stray too far away from the SI model.

In this work, we propose a novel regularization-based approach for speaker adaptation, in which we use adversarial multi-task learning (MTL) to regularize the distribution of the deep features (i.e., hidden representations) in an SD DNN acoustic model such that it does not deviate too much from the deep feature distribution in the SI DNN acoustic model. We call this method adversarial speaker adaptation (ASA). Recently, adversarial training has achieved great success in learning generative models [15]. In speech area, it has been applied to acoustic model adaptation [16, 17], noise-robust [18, 19, 20], speaker-invariant [21, 22, 23], ASR, speech enhancement [24, 25, 26] and speaker verification [27, 28] using gradient reversal layer [29] or domain separation network [30]. In these works, adversarial MTL assists in learning a deep intermediate feature that is both senone-discriminative and domain-invariant.

In ASA, we introduce an auxiliary discriminator network to classify whether an input deep feature is generated by an SD or SI acoustic model. By using a fixed SI acoustic model as the reference, the discriminator network is jointly trained with the SD acoustic model to simultaneously optimize the primary task of minimizing the senone classification loss and the secondary task of mini-maximizing the SD/SI discrimination loss on the adaptation data. Through this adversarial MTL, senone-discriminative deep features are learned in the SD model with a distribution that is similar to that of the SI model. With such a regularized and adapted deep feature, the SD model is expected to achieve improved ASR performance on the test speech from the target speaker. As an extension, ASA can also be performed on the senone posteriors (ASA-SP) to regularize the output distribution of the SD model.

We perform speaker adaptation experiments on Microsoft short message (SMD) dictation dataset with 2600 hours of live US English data for training. ASA achieves up to 14.4% and 7.9% relative word error rate (WER) improvements for supervised and unsupervised adaptation, respectively, over an SI model trained on 2600 hours of speech.

2. ADVERSARIAL SPEAKER ADAPTATION

In speaker adaptation task, for a target speaker, we are given a sequence of adaptation speech frames $X = \{x_1, \ldots, x_T\}, x_t \in \mathbb{R}^{d_x}, t = 1, \ldots, T$ from the target speaker and a sequence of senone labels $Y = \{y_1, \ldots, y_T\}, y_t \in \mathbb{P}$ aligned with $X$. For supervised adaptation, $Y$ is generated by aligning the adaptation data against the transcription using SI acoustic model while for unsupervised...
adaptation, the adaptation data is first decoded using the SI acoustic model and the one-best path of the decoding lattice is used as $Y$.

As shown in Fig. 1 we view the first few layers of a well-trained SI DNN acoustic model as an SI feature extractor network $M^f_{SD}$ with parameters $\theta^f_{SD}$ and the the upper layers of the SI model as an SI senone classifier network $M^y_{SI}$ with parameters $\theta^y_{SI}$. $M^f_{SI}$ maps input adaptation speech frames $X$ to intermediate SI deep hidden features $F^y_{SI} = \{f^y_{SI}, \ldots, f^y_T\}$, $f^y_t \in \mathbb{R}^{r_f}$, i.e.,

$$t^y_f = M^f_{SI}(x_t)$$

(1)

and $M^y_{SI}$ with parameters $\theta^y_{SI}$ maps $F^y_{SI}$ to the posteriors $p(s|f^y_{SI}; \theta^y_{SI})$ of a set of senones in $S$ as follows:

$$M^y_{SI}(t^y_f) = p(s|x_t; \theta^y_{SI}).$$

(2)

where $\mathbb{I}[\cdot]$ is the indicator function which equals to 1 if the condition in the squared bracket is satisfied and 0 otherwise.

In KLD adaptation, the adaptation data $X$ is usually very limited for the target speaker and the SI model with a large number of parameters can easily get overfitted to the adaptation data. Therefore, we need to force the distribution of deep hidden features $F^y_{SD}$ in the SD model to be close to that of the deep features $F^y_{SI}$ in SI model while minimizing $L_{senone}$ as follows

$$\begin{align}
\min_{\theta^f_{SD}, \theta^y_{SD}} L_{senone}(\theta^f_{SD}, \theta^y_{SD}),
\end{align}$$

(5)

In KLD adaptation, the senone distribution estimated from the SD model is forced to be close to that estimated from an SI model by adding KLD regularization to the adaptation criterion. However, KLD is a distribution-wise asymmetric measure which does not serve as a perfect distance metric between distributions [31]. For example, in [11], the minimization of $KL(P^S||P^D)$ does not guarantee $KL(P^D||P^S)$ is also minimized. In some cases, $KL(P^S||P^D)$ even increases as $KL(P^D||P^S)$ becomes smaller [32, 33]. In ASA, we use adversarial MTL instead to push the distribution of $F^y_{SD}$ towards that of $F^y_{SI}$ while being adapted to the target speech since the adversarial learning can guarantee that the global optimum is achieved and only if $F^y_{SD}$ and $F^y_{SI}$ share exactly the same distribution [13].

To achieve Eq. (5), we introduce an additional discriminator network $D_d$ with parameters $\theta_d$ which takes $F^y_{SD}$ and $F^y_{SI}$ as the input and outputs the posterior probability that an input deep feature is generated by the SD model, i.e.,

$$\begin{align}
M_d(t^y_f) &= p(t^y_f \in D_{SD}|x_t; \theta^y_{SD}, \theta_d),
M_d(t^y_f) &= 1 - p(t^y_f \in D_{SI}|x_t; \theta^y_{SI}, \theta_d),
\end{align}$$

(6)

where $D_{SD}$ and $D_{SI}$ denote the sets of SD and SI deep features, respectively. The discrimination loss $L_{disc}(\theta_f, \theta_d)$ for $M_d$ is formulated below using cross-entropy:

$$L_{disc}(\theta^y_{SD}, \theta^y_{SI}, \theta_d) = -\frac{1}{T} \sum_{t=1}^{T} \left[ \log p(t^y_f \in D_{SD}|x_t; \theta^y_{SD}, \theta_d) + \log p(t^y_f \in D_{SI}|x_t; \theta^y_{SI}, \theta_d) \right]$$

(7)

$$= -\frac{1}{T} \sum_{t=1}^{T} \left[ \log M_d(M^y_{SD}(x_t)) + \log \left[ 1 - M_d(M^y_{SD}(x_t)) \right] \right]$$

(8)

To make the distribution of $F^y_{SD}$ similar to that of $F^y_{SI}$, we perform adversarial training of $M^f_{SD}$ and $M_d$, i.e, we minimize $L_{disc}$ with respect to $\theta_d$ and maximize $L_{disc}$ with respect to $\theta^y_{SD}$. This minimax competition will first increase the capability of $M^f_{SD}$ to generate $F^y_{SD}$ with a distribution similar to that of $F^y_{SI}$ and increase the discrimination capability of $M_d$. It will eventually converge to the point where $M^f_{SD}$ generates extremely confusing $F^y_{SD}$ that $M_d$ is unable to distinguish whether it is generated by $M^f_{SD}$ or $M^f_{SI}$. At this point, we have successfully regularized the SD model such that it does not deviate too much from the SI model and generalizes well to the test speech from target speaker.

With ASA, we want to learn a senone-discriminative SD deep feature with a similar distribution to the SI deep features as in Eq. (5) and (6). To achieve this, we perform adversarial MTL, in which the SD model and $M_d$ are trained to jointly optimize the primary task of

Fig. 1. The framework of ASA. Only the optimized SD acoustic model consisting of $M^f_{SD}$ and $M^y_{SD}$ are used for ASR on test data. $M^f_{SI}$ is fixed during ASA. $M^f_{SI}$ and $M_d$ are discarded after ASA.
\[
(\hat{\theta}^f, \hat{\theta}^{\text{SD}}) = \arg\min_{\theta^f, \theta^{\text{SD}}} L_{\text{senone}}(\theta^f, \theta^{\text{SD}}) - \lambda L_{\text{disc}}(\theta^f, \theta^{\text{SD}}, \hat{\theta}_d), \quad (10)
\]

\[
(\hat{\theta}_d) = \arg\min_{\theta_d} L_{\text{disc}}(\theta^f, \theta^{\text{SD}}, \theta_d), \quad (11)
\]

where \( \lambda \) controls the trade-off between \( L_{\text{senone}} \) and \( L_{\text{disc}} \), and \( \hat{\theta}^f, \hat{\theta}^{\text{SD}} \) and \( \hat{\theta}_d \) are the optimized parameters. Note that the SI model serves only as a reference in ASA and its parameters \( \theta_y^{\text{SI}}, \theta_f^{\text{SI}} \) are fixed throughout the optimization procedure.

The parameters are updated via back propagation with stochastic gradient descent:

\[
\theta^f \leftarrow \theta^f - \mu \left[ \frac{\partial L_{\text{senone}}}{\partial \theta^f} - \lambda \frac{\partial L_{\text{disc}}}{\partial \theta^f} \right], \quad (12)
\]

\[
\theta_d \leftarrow \theta_d - \mu \frac{\partial L_{\text{disc}}}{\partial \theta_d}, \quad (13)
\]

\[
\theta^{\text{SD}} \leftarrow \theta^{\text{SD}} - \mu \left[ \frac{\partial L_{\text{senone}}}{\partial \theta^{\text{SD}}} - \lambda \frac{\partial L_{\text{disc}}}{\partial \theta^{\text{SD}}} \right], \quad (14)
\]

where \( \mu \) is the learning rate. For easy implementation, gradient reversal layer is introduced in [29], which acts as an identity transform in the forward propagation and multiplies the gradient by \(-\lambda\) during the backward propagation. Note that only the optimized SD DNN acoustic model consisting of \( M^{\text{SD}}_f \) and \( M^{\text{SD}}_y \) is used for ASR on test data. \( M_d \) and SI model are discarded after ASA.

The procedure of ASA can be summarized in the steps below:

1. Divide a well-trained and fixed SI model into a feature extractor \( M^{\text{SI}}_f \) followed by a senone classifier \( M^{\text{SI}}_y \).
2. Initialize the SD model with the SI model, i.e., clone \( M^{\text{SD}}_f \) and \( M^{\text{SD}}_y \) from \( M^{\text{SI}}_f \) and \( M^{\text{SI}}_y \), respectively.
3. Add an auxiliary discriminator network \( M_d \) taking SD and SI deep features, \( F^{\text{SD}} \) and \( F^{\text{SI}} \), as the input and predict the posterior that the input is generated by \( M^{\text{SD}}_d \).
4. Jointly optimize \( M^{\text{SD}}_f \), \( M^{\text{SD}}_y \) and \( M_d \) with adaptation data of a target speaker via adversarial MTL as in Eq. (10) to (14).
5. Use the optimized SD model consisting of \( M^{\text{SD}}_f \), \( M^{\text{SD}}_y \) for ASR decoding on test data of this target speaker.

### 3. ADVERSARIAL SPEAKER ADAPTATION ON SENONE POSTERIORS

As shown in Fig. 2 in ASA-SP, adversarial learning is applied to regularize the vectors of senone posteriors \( y^{\text{SD}}_t = [p(s|x_t; \theta^{\text{SD}}_{\text{AM}})] \) predicted by the SD model to be close to that of a well-trained and fixed SI model, i.e., \( y^{\text{SI}}_t = [p(s|x_t; \theta^{\text{SI}}_{\text{AM}})] \) while simultaneously minimizing the senone loss \( L_{\text{senone}}(\theta^{\text{SD}}_{\text{AM}}) \), where \( \theta^{\text{SD}}_{\text{AM}} = \{\theta^{\text{SD}}_y, \theta^{\text{SD}}_f\} \) and \( \theta^{\text{SI}}_{\text{AM}} = \{\theta^{\text{SI}}_y, \theta^{\text{SI}}_f\} \) are SI and SD model parameters, respectively.

In this case, the discriminator \( M_d \) takes \( p(s|x_t; \theta^{\text{SD}}_{\text{AM}}) \) and \( p(s|x_t; \theta^{\text{SI}}_{\text{AM}}) \) as the input and predicts the posterior that the input is generated by the SD model. The discrimination loss \( L_{\text{disc}}(\theta_f, \theta_d) \) for \( M_d \) is formulated below using cross-entropy:

\[
L_{\text{disc}}(\theta^{\text{SD}}_{\text{AM}}, \theta_d) = -\frac{1}{T} \sum_{t=1}^{T} \left[ \log p(y^{\text{SD}}_t \in \mathbb{E}^{\text{SD}}|x_t; \theta^{\text{SD}}_{\text{AM}}, \theta_d) + \log p(y^{\text{SI}}_t \in \mathbb{E}^{\text{SI}}|x_t; \theta^{\text{SI}}_{\text{AM}}, \theta_d) \right], \quad (15)
\]

![Fig. 2. The framework of ASA-SP. The SI acoustic model is fixed during ASA-SP. Only the SD acoustic model is used in ASR.](image)

where \( \mathbb{E}^{\text{SD}} \) and \( \mathbb{E}^{\text{SI}} \) denote the sets of senone posterior vectors generated by SD and SI models, respectively. Similar adversarial MTL is performed to make the distributions of \( Y^{\text{SD}} = \{y^{\text{SD}}_1, \ldots, y^{\text{SD}}_T\} \) similar to that of \( Y^{\text{SI}} = \{y^{\text{SI}}_1, \ldots, y^{\text{SI}}_T\} \):

\[
(\hat{\theta}^{\text{AM}}_{\text{SD}}) = \arg\min_{\theta^{\text{SD}}_{\text{AM}}} L_{\text{senone}}(\theta^{\text{SD}}_{\text{AM}}) - \lambda L_{\text{disc}}(\theta^{\text{SD}}_{\text{AM}}, \theta^{\text{SI}}_{\text{AM}}, \theta_d), \quad (16)
\]

\[
(\hat{\theta}_d) = \arg\min_{\theta_d} L_{\text{disc}}(\theta^{\text{SD}}_{\text{AM}}, \theta^{\text{SI}}_{\text{AM}}, \theta_d). \quad (17)
\]

\( \theta^{\text{SI}}_{\text{AM}} \) is optimized by back propagation below, \( \theta_d \) is optimized via Eq. (13) and \( \theta^{\text{SD}}_{\text{AM}} \) remain unchanged during the optimization.

\[
\theta^{\text{SD}}_{\text{AM}} \leftarrow \theta^{\text{SD}}_{\text{AM}} - \mu \left[ \frac{\partial L_{\text{senone}}}{\partial \theta^{\text{SD}}_{\text{AM}}} - \lambda \frac{\partial L_{\text{disc}}}{\partial \theta^{\text{SD}}_{\text{AM}}} \right]. \quad (18)
\]

ASA-SP is an extension of ASA where the deep hidden feature moves up to the output layer and becomes the senone posteriors vector. In this case, the senone classifier disappears and the feature extractor becomes the entire acoustic model.

### 4. EXPERIMENTS

We perform speaker adaptation on a Microsoft Windows Phone SMD task. The training data consists of 2600 hours of Microsoft internal live US English data collected through a number of deployed speech services including voice search and SMD. The test set consists of 7 speakers with a total number of 20,203 words. Four adaptation sets of 20, 50, 100 and 200 utterances per speaker are used for acoustic model adaptation, respectively, to explore the impact of adaptation data duration. Any adaptation set with smaller number of utterances is a subset of a larger one.

#### 4.1. Baseline System

We train an SI long short-term memory (LSTM)-hidden Markov model (HMM) acoustic model with 2600 hours of training data. This SI model has 4 hidden layers with 1024 units in each layer and the output size of each hidden layer is reduced to 512 by a linear projection. 80-dimensional log Mel filterbank features are extracted from training, adaptation and test data. The output layer has a dimension of 5980. The LSTM is trained to minimize the frame-level cross-entropy criterion. There is no frame stacking, and the...
output HMM state label is delayed by 5 frames. A trigram LM is used for decoding with around 8M n-grams. This SI LSTM acoustic model achieves 13.95% WER on the SMD test set.

We further perform KLD speaker adaptation [11] with different regularization weights $\rho$. In Tables 1 and 2 KLD with $\rho = 0.5$ achieves 12.54% - 13.24% and 13.55% - 13.85% WERs for supervised and unsupervised adaptation, respectively with 20 - 200 adaptation utterances.

<table>
<thead>
<tr>
<th>System</th>
<th>Number of Adaptation Utterances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20</td>
</tr>
<tr>
<td>SI</td>
<td>13.95</td>
</tr>
<tr>
<td>KLD ($\rho = 0.0$)</td>
<td>13.68</td>
</tr>
<tr>
<td>KLD ($\rho = 0.2$)</td>
<td>13.20</td>
</tr>
<tr>
<td>KLD ($\rho = 0.5$)</td>
<td>13.24</td>
</tr>
<tr>
<td>KLD ($\rho = 0.8$)</td>
<td>13.55</td>
</tr>
<tr>
<td>ASA ($\lambda = 1.0$)</td>
<td>12.99</td>
</tr>
<tr>
<td>ASA ($\lambda = 3.0$)</td>
<td>13.03</td>
</tr>
<tr>
<td>ASA ($\lambda = 5.0$)</td>
<td>13.14</td>
</tr>
<tr>
<td>ASA-SP ($\lambda = 1.0$)</td>
<td>13.04</td>
</tr>
<tr>
<td>ASA-SP ($\lambda = 3.0$)</td>
<td>13.05</td>
</tr>
<tr>
<td>ASA-SP ($\lambda = 5.0$)</td>
<td>13.05</td>
</tr>
</tbody>
</table>

Table 1. The WER (%) performance of supervised speaker adaptation using KLD and ASA with different $\lambda$ on Microsoft SMD task.

For supervised ASA, the same alignment is used as in KLD. In Table 1 the best ASA setups achieve 12.99%, 12.71%, 12.35% and 11.94% WERs for 20, 50, 100, 200 adaptation utterances which improve the WERs by 6.9%, 8.9%, 11.5%, 14.4% relatively over the SI LSTM, respectively. Supervised ASA ($\lambda = 3.0$) also achieves up to 5.3% relative WER reduction over the best KLD setup ($\rho = 0.2$).

For unsupervised ASA, the same decoded senone labels are used as in KLD. In Table 2 the best ASA setups achieve 13.66%, 13.61%, 13.09% and 12.85% WERs for 20, 50, 100, 200 adaptation utterances which improves the WERs by 2.1%, 2.4%, 6.2%, 7.9% relatively over the SI LSTM, respectively. Unsupervised ASA ($\lambda = 3.0$) also achieves up to and 5.2% relatively WER gains over the best KLD setup ($\rho = 0.5$). Compared with supervised ASA, the unsupervised one decreases the relative WER gain over the SI LSTM by about half on the same number of adaptation utterances.

For both supervised and unsupervised ASA, the WER first decreases as $\lambda$ grows larger and then increases when $\lambda$ becomes too large. ASA performs consistently better than SI LSTM and KLD with different number of adaptation utterances for both supervised and unsupervised adaptation. The relative gain increases as the number of adaptation utterance grows.

### 4.3. Adversarial Speaker Adaptation on Senone Posteriors

We perform standard ASA as described in Section 4.2. The SD acoustic model is cloned from the SI LSTM as the initialization. $M_d$ shares the same architecture as the one in Section 4.2. In Table 1 for supervised adaptation ASA-SP ($\lambda = 1.0$) achieves 6.5%, 7.7%, 9.2%, 12.8% relative WER gain over the SI LSTM, respectively and up to 3.5% relative WER reduction over the best KLD setup ($\rho = 0.2$). In Table 2 for unsupervised adaptation ASA-SP ($\lambda = 1.0$) achieves 0.9%, 1.6%, 3.4%, 5.7%, 2.9% relative WER gain over the SI LSTM, respectively and up to 4.8% relative WER reduction over the best KLD setup ($\rho = 0.5$).

Although ASA-SP consistently improves over KLD on different number of adaptation utterances for both supervised and unsupervised adaptation, it performs worse than standard ASA where the regularization from SI model is performed at the hidden layers. The reason is that the senone posterior vectors $y^{st}$, $y^{sd}$ lie in a much higher-dimensional space than the deep features $f^{st}$, $f^{sd}$ so that the discriminator is much harder to learn given much sparser-distributed samples. We also notice that ASA-SP performance is much less sensitive to the variation of $\lambda$ compared with standard ASA.

### 5. CONCLUSION

In this work, a novel adversarial speaker adaptation method is proposed, in which the deep hidden features (ASA) or the output senone posteriors (ASA-SP) of an SD DNN acoustic model are forced by the adversarial MTL to conform to a similar distribution as those of a fixed reference SI DNN acoustic model while being trained to be senone-discriminative with the limited adaptation data.

We evaluate ASA on Microsoft SMD task with 2600 hours of training data. ASA achieves up to 14.4% and 7.9% relative WER gain for supervised and unsupervised adaptation, respectively, over the SI LSTM acoustic model. ASA also improves consistently over the KLD regularization method. The relative gain grows as the number of adaptation utterances increases. ASA-SP performs consistently better than KLD but worse than the standard ASA.

---

3It has been shown in [17][16] that the ASR performance increases with the growth of $N_h$ for adversarial domain adaptation. The same trend is observed for ASA experiments.
6. REFERENCES


