Neurocomputing 000 (2017) 1–9

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

An unsupervised deep domain adaptation approach for robust speech recognition

Sining Sun, Binbin Zhang, Lei Xie*, Yanning Zhang

Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, School of Computer Science, Northwestern Polytechnical University, Xi'an, China

ARTICLE INFO

Article history:
Received 15 June 2016
Revised 13 November 2016
Accepted 29 November 2016
Available online xxx

Keywords:
Domain adaptation
Robust speech recognition
Deep neural network
Deep learning

ABSTRACT

This paper addresses the robust speech recognition problem as a domain adaptation task. Specifically, we introduce an unsupervised deep domain adaptation (DDA) approach to acoustic modeling in order to eliminate the training–testing mismatch that is common in real-world use of speech recognition. Under a multi-task learning framework, the approach jointly learns two discriminative classifiers using one deep neural network (DNN). As the main task, a label predictor predicts phoneme labels and is used during training and at test time. As the second task, a domain classifier discriminates between the source and the target domains during training. The network is optimized by minimizing the loss of the label classifier and to maximize the loss of the domain classifier at the same time. The proposed approach is easy to implement by modifying a common feed-forward network. Moreover, this unsupervised approach only needs labeled training data from the source domain and some unlabeled raw data of the new domain. Speech recognition experiments on noise/channel distortion and domain shift confirm the effectiveness of the proposed approach. For instance, on the Aurora-4 corpus, compared with the acoustic model trained only using clean data, the DDA approach achieves relative 37.8% word error rate (WER) reduction.

1. Introduction

The increasing availability of multimedia big data, including various genres of speech, is fostering a new wave of multimedia analytics that aim to effectively access the content and pull meaning from the data. Automatic speech recognition (ASR), which transcribes speech into text, serves as a necessary preprocessing step for multimedia analytics. With the help of big data, supercomputing infrastructure and deep learning [1], the speech recognition accuracy has been dramatically lifted during the past years [2]. Besides the Gaussian mixture model–hidden Markov model (GMM–HMM) architecture that dominates the acoustic modeling in speech recognition for many years, artificial neural networks have been historically used as an alternative model but with limited success [3]. Only recently, neural network has re-emerged as an effective tool for acoustic modeling because of the power of big data and effective learning method [4,51]. The DNN–HMM architecture has come to the central stage in speech recognition [2,5], replacing the GMM–HMM architecture. We have witnessed the success of various types of (deep) neural networks (DNNs) not only in speech recognition, but also in visual data processing, data mining and other areas [6–10].

Speech is a typical big data: not just in volume, but also noisy and heterogeneous. In practice, we desire a *robust* speech recognizer that is able to handle noisy data. For many machine learning tasks, including ASR, we usually assume that the training data and the testing data have the same probability distributions. However, real-world applications often fail to meet this hypothesis [5,11]. In speech recognition, both the GMM–HMM and DNN–HMM systems are Bayesian classifiers by nature. Theoretical investigation has shown that the training–testing mismatch notoriously leads to increase of errors in Bayesian classification [11]. There are many reasons that lead to the mismatch such as environmental noises, channel distortions [12] and room reverberations [13, [14]. To improve the environmental robustness of a speech recognizer, a common and efficient approach is multi-condition training [15] that uses the contaminated noisy data, together with the clean data, in the acoustic modeling training. But it is impossible to cover all kinds of real-world conditions and the mismatch still exists. Therefore, environment robustness is still a big challenge remain unsolved. On the other hand, real-world speech data is heterogeneous. Speech in different domains, e.g., broadcast news, lectures, meeting recordings and conversations, has different char-

* Corresponding author.
E-mail addresses: snsun@nwpu-aslp.org (S. Sun), xielei21st@gmail.com (L. Xie).

http://dx.doi.org/10.1016/j.neucom.2016.11.063
0925-2312/

acteristics. This causes another mismatch that apparently decrease the speech recognition performance [14].

In order to eliminate the training–testing mismatch, a large number of robust speech recognition methods have been proposed, which in general fall into two categories: feature-space approaches and model-space approaches [16,17]. Most approaches need some prior knowledge about the mismatch. For example, noise characteristics have to be known beforehand or clean-noisy speech pairs[1] are needed [18]. Model adaptation is a typical model-space approach that is quite useful in noise robustness. The acoustic model, e.g., GMM-HMMs, is adapted using the new data either in a supervised manner [19] or an unsupervised manner [20]. For feature-space approaches, it is common to combine information about speaker, environment and noise, such as using *i*-vector [21], to acoustic features.

In this paper, we regard the robust speech recognition problem as a domain adaptation (DA) task [22]. Learning a discriminative classifier in the presence of the mismatch between training and testing distributions is known as *domain adaptation*. The essence of the domain adaptation and robust speech recognition is identical, that is, to eliminate the mismatch between the training data and the test data. We find that the speech features yield to different distributions if they come from different domains (such as clean and noisy speech conditions [16], and data sets with different genres). Specifically, if we train a DNN acoustic model using clean speech, we discovered that the feature distributions of clean and noisy speech yielded from this acoustic model are significantly different. Hence we would like to embed the domain information during the acoustic model training in order to obtain a "domain-invariant feature extractor".

Our work is inspired by a recent DNN based unsupervised domain adaptation approach for image classification [23]. This *deep domain adaptation* (DDA) approach combines domain adaptation and deep feature learning within a single training process. Specifically, under a multi-task learning framework [52], the approach jointly learns one feature extractor and two discriminative classifiers using one single DNN: the feature extractor is trained to extract domain-invariant and classification-discriminative features; the label predictor predicts class labels and is used both during training and testing; a *domain predictor* discriminates between the source and the target domains during training. In order to obtain domain-invariant and classification-discriminative features, the feature extractor sub-network is optimized by minimizing the loss of the label predictor and maximizing the loss of the domain predictor at the same time, which is achieved by a special objective function we defined later. The parameters of two predictor sub-networks are optimized in order to minimize their losses on the training set. Compared with other unsupervised adaptation approaches, the DDA approach is easy to implement by simply augmenting a common feed-forward network with few standard layers and a simple new gradient reversal layer. Moreover, this approach only needs the labeled training data from the source domain and some unlabeled raw data of the new domain. Experiments show that the DDA approach outperforms previous state-of-the-art image classification approaches on several popular datasets [23].

In this study, we introduce the DDA approach to robust speech recognition. Applying DDA to speech recognition is not trivial. This is because speech recognition is a more challenging task as compared with image classification. We elaborate some of the major challenges as follows.

- *The large number of labels*: In the typical image classification task in [23], the number of classes are only dozens. In contrast, in speech recognition, the class labels are thousands of senones (i.e., phoneme states). The effectiveness of DDA on a large scale classification task like speech recognition desires an intensive study.
- *Decoding*: As compared with image classification, speech recognition is a rather complicated task with frame-level classification (classify each speech frame into senone labels) and decoding (Viterbi search from a large graph based on the classified frame labels). The accuracy gain in frame-level classification may not ensure consistent accuracy gain at the word level [24].
- *Deeper networks*: The neural networks in speech recognition usually have many hidden layers in order to learn highly nonlinear and discriminative features which are robust to irrelevant variabilities.

To bridge the gap, in this paper, we study how to integrate DDA into acoustic modeling and present a systematic analysis of the performance of DDA in robust speech recognition. Our study shows that the DDA approach can significantly boost the speech recognition performance in both noisy/channel distortion and domain-shift conditions.

The rest of this paper is structured as follows. Section 2 surveys the related work. Section 3 presents the framework of deep domain adaptation and studies how to use it in the speech recognition task. Experimental settings and results are discussed in Sections 4–6 and finally conclusions are drawn in Section 7.

2. Related work

As we just mentioned, robust speech recognition methods can be classified into two categories: feature-space approaches and model-space approaches [16,17]. Compared with model-space approaches, feature-space approaches do not need to modify or retrain the acoustic model. Instead, various operations can be performed in the acoustic features to improve the noise (or other distortions) robustness of the features. As for the model-space approaches, rather than focusing on the modification of features, the acoustic model parameters are adjusted to match the testing data.

2.1. Traditional methods

In the feature space, feature normalization is the most straightforward strategy to eliminate the training–testing mismatch. Popular strategies include cepstral mean subtraction (CMS) [25], cepstral mean variance normalization (CMVN) [26] and histogram equalization (HEQ) [27]. Obviously, speech enhancement methods [28,29] can be adopted to remove the noise before speech recognition. But the unavoidable distortions in the enhanced speech may cause another new mismatch problem.

Rather than updating the features, the acoustic model parameters can be compensated to match the testing conditions. A simple example of updating the models is to re-train them with the new data; or more popular, adding a variety of noise samples to clean training data, known as multi-style or multi-condition training [15,17]. However, due to the unpredictable nature of real-world noise, it is impossible to account for all noise conditions that may be encountered. Thus adaptive and predictive methods are proposed in the model-space. The adaptive methods update the model parameters when sufficient corrupted speech data are available. Popular methods include maximum a posteriori re-estimation (MAP) [30] and maximum likelihood linear regression (MLLR) [31]. In the predictive methods, a noise model is combined with the clean speech models to provide a corrupted speech acoustic model using some model of the acoustic environment. Parallel model combination (PMC) [32] and vector Taylor series (VTS) [33] fall into this category.

[1] Noisy speech may be generated manually by adding noises into clean speech.

Please cite this article as: S. Sun et al., An unsupervised deep domain adaptation approach for robust speech recognition, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2016.11.063

2.2. DNN based methods

Compared with GMMs, DNNs have an outstanding non-linear learning ability, which makes DNN a more robust acoustic model. Hence the DNN–HMM architecture is inherently noise robust to some extent as compared with GMM–HMM [17]. However, it is not enough to solve the mismatch problem merely relying on the non-linear learning ability. Recently, many methods in the feature and model spaces have been proposed to make DNN–HMM more robust to the mismatched test data. In order to account for the mismatch, many useful auxiliary features, reflecting environmental noise and speaker information [17,18], are combined with acoustic features as the DNN input. Neural networks can be used as a speech enhancement tool. In [34,35], a denoising autoencoder (DAE) is adopted to reconstruct clean speech features from noisy ones. This kind of method needs stereo data, i.e., clean speech and corresponding noisy speech, to train the denoising DNN. DNN feature enhancement and DNN acoustic model can be trained jointly [36,37]. Multi-task training is another popular strategy to improve the robustness of the acoustic model [38]. By adding one or more auxiliary output layers in the DNN and optimizing several tasks (e.g., main task: prediction of senone labels, side task: denoising) at the same time, the network gains more robustness [38].

3. Deep domain adaptation for robust ASR

3.1. The model

We treat the training–testing mismatch problem as a domain adaptation task, bridging the target (testing) and the source (training) domains. The main purpose of deep domain adaptation (DDA) [23] is to embed the domain information into the process of learning representation, so that the final classification decisions are made based on features that are both *discriminative* and *invariant* to the changes of domains. This means the representation learned by the DNN classifier has the same or very similar distributions in the source and the target domains.

Assume that the neural network model works with input samples $\mathbf{x} \in X$ and certain labels $\mathbf{y} \in Y$ where X and Y are input space and output space, respectively. Here in speech recognition, $\mathbf{x}$ and $\mathbf{y}$ are framewise acoustic features and senones (phoneme states), respectively. There are two distributions $S(\mathbf{x}, \mathbf{y})$ and $T(\mathbf{x}, \mathbf{y})$ on $X \otimes Y$, which are referred to as the source distribution (for training) and the target distribution (for testing) and both the two distributions are assumed complicated and unknown. Due to domain shift, S and T are similar but different.

In the training–testing mismatch scenario, we train the model with $S(\mathbf{x}, \mathbf{y})$, but we test the model with the data yields to distribution $T(\mathbf{x}, \mathbf{y})$. However, we can access to many training samples $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ from source domain and target domain according to the marginal distributions $S(\mathbf{x})$ and $T(\mathbf{x})$. Denote with $\mathbf{d}_i$([0, 1] or [1, 0]) the (*domain label*) for the ith sample, which indicates whether $\mathbf{x}_i$ comes from the source domain ($\mathbf{x}_i \sim S(\mathbf{x})$ if $\mathbf{d}_i = [1, 0]$) or from the target domain ($\mathbf{x}_i \sim T(\mathbf{x})$ if $\mathbf{d}_i = [0, 1]$).

The unsupervised deep domain adaptation architecture [23] is depicted in Fig. 1. The architecture is simply based on a feed-forward neural network. But different from a common one, this network has two output layers, which are the main class label $\mathbf{y} \in Y$ and the domain label $\mathbf{d} \in \{[0, 1], [1, 0]\}$. Specifically, this model is decomposed into three parts to perform different mappings: a feature extractor G_f, a label predictor G_y and a domain predictor G_d.

More formally, the mapping functions are:

$$\mathbf{f} = G_f(\mathbf{x}; \Theta_f); \tag{1}$$

$$\mathbf{y} = G_y(\mathbf{f}; \Theta_y); \tag{2}$$

$$d = G_d(\mathbf{f}; \Theta_d); \tag{3}$$

where Θ_f, Θ_y, Θ_d are the parameters of the network (in Fig. 1) and $\mathbf{f}$ is a D-dimension feature vector.

Our aim is to jointly train G_f, G_y and G_d. Specifically, we want to seek Θ_f to minimize the label prediction loss and to maximize the domain classification loss at the same time. The maximization of the domain classification loss is actually to make the two feature domain distributions as similar as possible. Meanwhile, in order to assure the domain classification, the Θ_d has to make the mapping G_d perform well in domain classification. This leads to the loss function of this network:

$$\begin{aligned} E(\Theta_f, \Theta_y, \Theta_d) &= \sum_{\substack{i=1,\ldots N \\ \mathbf{d}_i=[1,0]}} L_y(G_y(G_f(\mathbf{x}_i; \Theta_f); \Theta_y), \mathbf{y}_i) \\ &\quad - \lambda \sum_{i=1,\ldots N} L_d(G_d(G_f(\mathbf{x}_i; \Theta_f); \Theta_d), \mathbf{d}_i) \\ &= \sum_{\substack{i=1,\ldots N \\ \mathbf{d}_i=[1,0]}} L_y^i(\Theta_f, \Theta_y) - \lambda \sum_{i=1,\ldots N} L_d^i(\Theta_f, \Theta_d) \end{aligned} \tag{4}$$

where $L_y(.,.)$ and $L_d(.,.)$ are loss functions for label and domain predictors respectively, while $L_y^i(.,.)$ and $L_d^i(.,.)$ denote the loss of the ith training sample. Loss functions can be cross entropy or mean square error function depends on the tasks. λ is a positive hyper parameter used to trade off two losses in practice. Frankly, the similar loss functions are common used in many other machine learning task [39–41].

3.2. Optimization

According to the loss function derived from Section 3.1, we can optimize the DDA network using an approach similar to stochastic gradient decent (SGD) [42]. The aim of the optimization is to seek the optimized parameters that:

$$(\hat{\Theta}_f, \hat{\Theta}_y) = \arg\min_{\Theta_f, \Theta_y} E(\Theta_f, \Theta_d, \Theta_y), \tag{5}$$

$$\hat{\Theta}_d = \arg\max_{\Theta_d} E(\Theta_f, \Theta_d, \Theta_y). \tag{6}$$

Although Θ_d is optimized by maximizing Eq. (4), it equals to minimize the second item of Eq. (4). So Θ_d will make sure the performance of domain predictor. Θ_f is optimized by minimizing the first item and maximizing the second item (because of the minus symbol). This training strategy will keep the feature extracted from the neural network domain-invariant and classification-discriminative. Under the multi-task learning framework, the following equations are used to update the parameters:

$$\Theta_f \leftarrow \Theta_f - \mu\left(\frac{\partial L_y^i}{\partial \Theta_f} - \lambda \frac{\partial L_d^i}{\Theta_f}\right) \tag{7}$$

$$\Theta_d \leftarrow \Theta_d - \mu \frac{\partial L_d^i}{\partial \Theta_d} \tag{8}$$

$$\Theta_y \leftarrow \Theta_y - \mu \frac{\partial L_y^i}{\partial \Theta_y} \tag{9}$$

where μ is step size.

3.3. Applying DDA to speech recognition

State-of-the-art ASR systems are Bayesian classifiers by nature. A typical speech recognition system can be formulated as a simple equation:

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W} \in \mathcal{L}} P(\mathbf{X}|\mathbf{W}) P(\mathbf{W}) \tag{10}$$

Please cite this article as: S. Sun et al., An unsupervised deep domain adaptation approach for robust speech recognition, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2016.11.063

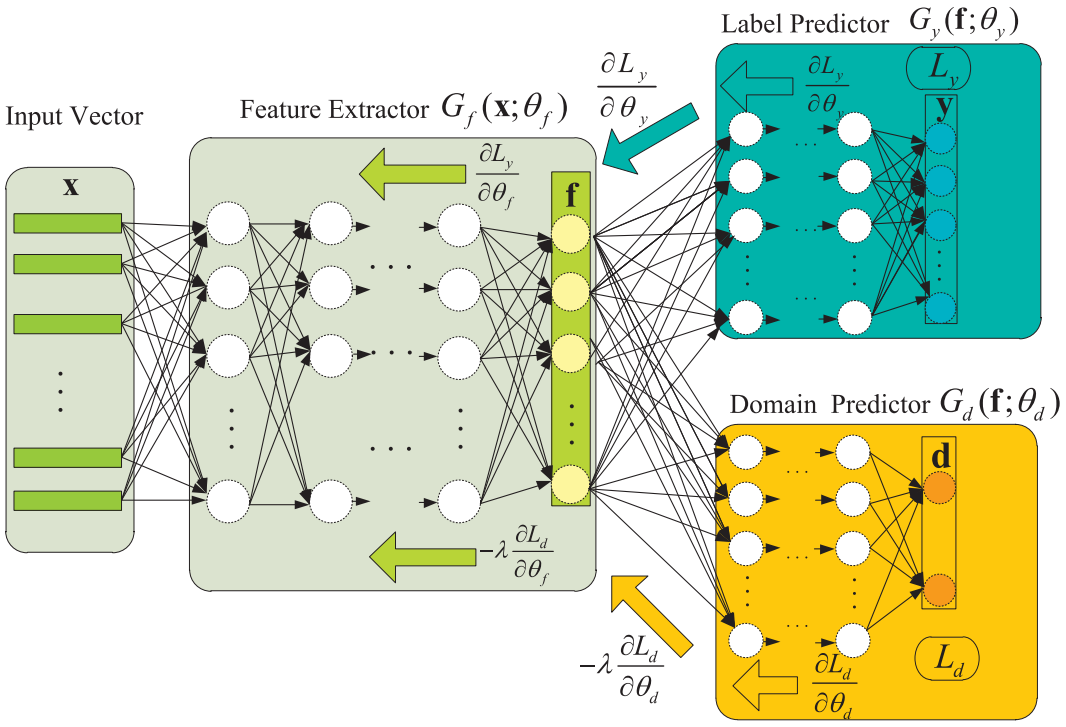


Fig. 1. Unsupervised deep domain adaptation architecture.

where $\mathbf{W} = \{w_1, w_2, \ldots\}$ is a possible word sequence in language $\mathcal{L}$, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$ is the *observation* sequence with frame-level acoustic feature $\mathbf{x}$, $P(\mathbf{X}|\mathbf{W})$ is the acoustic model and $P(\mathbf{W})$ is the language model. Therefore speech recognition (or decoding) is to find out the optimal word sequence $\hat{\mathbf{W}}$ that maximizes the joint acoustic and language probabilities.

As for the language model, word level *N*-gram model [43], trained from a large set of textual data, is usually used. The acoustic model is often built at fine-grained phoneme (subword) level, trained from labeled speech data with transcripts. The distribution of speech data is complex and the speech production is apparently a dynamic process. Traditionally, hidden Markov models (HMMs) are used to model this dynamic process in a phoneme through state transitions, while Gaussian mixture models (GMMs) are used to depict the distribution of speech data at HMM state level (subphoneme or so-called senone). This is the so-called GMM–HMM architecture. In practice, context-dependent models, e.g., triphones, are used to model the important coarticulation phenomenon in speech production. Recently, neural networks have re-emerged as a powerful acoustic modeling tool with superior performance [2,5], replacing GMMs to depict the distribution of speech data, namely the DNN–HMM architecture. Either GMM–HMM or DNN–GMM, if the distributions of the training data and the test data have some differences, the error of the Bayesian classifier will be increased [11]. Hence in this study, we use the unsupervised deep domain adaptation (DDA) strategy to adjust the acoustic model during the training time. Our purpose is to let the DNN acoustic model learn similar distributions both in the training data and the test data, which may increase the robustness of the Bayesian classifier.

Fig. 2 shows how to use the DDA strategy in speech recognition. A speech recognition system is composed of an acoustic model training stage[2] and a testing stage. In the acoustic model training stage, the first step is to extract acoustic features (represented by input vector $\mathbf{x}$ in Fig. 2), such as MFCC or FBank, for the training speech samples. Then the acoustic feature sequences are used to train triphone GMM–HMM acoustic models (so-called senones). The GMM–HMM models are just used to perform forced alignment to the training samples, obtaining the labeled training samples (speech frame and its corresponding senone label). Within the pairwise frame-label data, a DNN acoustic model is thus learned that classifies the input frame-level acoustic vector into senone label. In this process, we can use the DDA approach to learn the senone label classifier and the domain classifier at the same time using the labeled training data and some of the unlabeled raw data from the testing domain. At the test stage, the domain predictor is discarded and we only use the senone predictor as the acoustic model.

Given the predicted senone label scores, a speech recognizer still needs a *decoder* to obtain the best word sequence. As we mentioned in the beginning of this section, decoding involves not only an acoustic model, but also a language model. The acoustic score and the language score are combined in the decoding process for the decision of the final word sequence. Here we use the weighted finite-state transducers (WFST) [24] based static decoder to do the combination. In order to compose the decoding WFST, apart from the acoustic model and the language model, a lexicon and the context are also needed [24,44]. Using the compose operation in WFST, the different level representations are integrated in just one WFST graph, which maps the HMM states to words. For efficiency reasons, token passing [45] and beam search algorithms are often applied in the decoding process.

4. Experiments for noise/channel robustness

We evaluate the noise robustness of DDA on Aurora-4 [15], a popular corpus for robust ASR research. Aurora-4 is designed to verify the effectiveness of robust ASR methods on a medium vocabulary continuous speech recognition task. There are two different training conditions: (1) clean training condition, which

[2] A language model is also needed, but its training is out of the scope of this paper.

Please cite this article as: S. Sun et al., An unsupervised deep domain adaptation approach for robust speech recognition, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2016.11.063

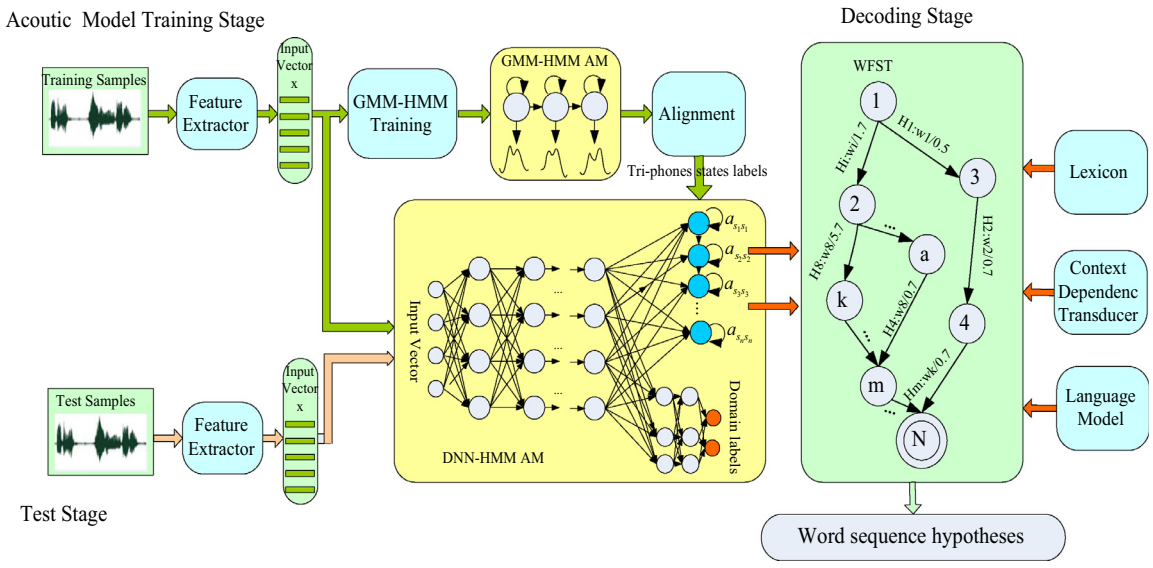


Fig. 2. The DDA approach used for robust ASR.

includes 7138 utterances recorded with the primary microphone without any added noise or distortions; and (2) multi-condition training condition, including the same 7138 utterances, but with one half of the data was recorded by the primary microphone and the other half recorded using the second microphone; all are contaminated with six types of added noises at 10–20 dB SNR. In order to investigate different noise/channel distortion conditions, the Aurora-4 test set is composed of four subsets.

- Subset A (Clean): 330 clean utterances without any noises or distortions, recorded with the primary microphone;
- Subset B (Noise): 330 × 6 utterances, by corrupting Subset A with six different noises;
- Subset C (Channel distortion): 330 utterances, same as Subset A, but recorded with the second microphone, without any added noises.
- Subset D (Noise + Channel distortion): 330 × 6 utterances, by corrupting Subset C with six different noises

All the speech files are sampled at 16KHz, quantified by 16 bits.

4.1. Clean condition training with multi-condition testing

This experiment is designed to evaluate the robustness of the DDA approach in mismatched training–testing condition: acoustic model is trained using clean speech while tested in multiple conditions with contaminated speech. Specifically, we use the clean-condition training set of Aurora-4, which includes 7138 utterances, to train a triphone GMM–HMM acoustic model. The acoustic feature is 39-dim MFCC. Then the GMM–HMM acoustic model is used to align the training data to obtain the triphone state (senones) labels.

After that, two different DNN–HMM acoustic models are trained: the conventional DNN–HMM model trained with a standard feed-forward network and the new DNN–HMM model trained using the DDA approach in Fig. 1. For clarity, they are named as Clean-DNN–HMM and DDA–DNN–HMM, respectively. The Clean-DNN–HMM model is trained using all the 7138 clean-condition training utterances, as a baseline model. The training data of DDA–DNN–HMM consists of two parts: 7138 clean-condition utterances with senone labels and 3000 multi-condition utterances without senone labels. The clean-condition utterances are used to train the whole network (G_f, G_y, G_d) while the multi-condition utterances are used to train the feature extractor and the domain classifier (G_f, G_d). Because the data from the target domain does not have senone labels, we randomly generate senone labels for the target domain data in order to train the model in a uniform framework. Specifically, we use a binary flag to control if the errors of the current frame is used to optimize the feature extractor and the senone labels predictor or not. If the current frame comes from the target domain, the senone predictor errors are thus discarded. As for the domain predictor, we also have two domain labels to predict. Although there are various kinds of noises in our training data, we do not distinguish them because we do not want to use too much priori knowledge of the data. Hence for simplicity, there are just two class labels to predict (clean and noise).

For the two DNN–HMM systems, the input layer is a context window of 11 frames of 40-dim FBANK with delta and acceleration coefficients (40 × 3 × 11). The G_f part of the network has 6 hidden layers with 1024 units in each layer. We also compare our approach with a state-of-the-art approach – DNN-PP [35]. Two DNNs are used in this approach [35]: speech enhancement DNN and acoustic model DNN. The first DNN, as a pre-processor for denoising, trained with clean-noisy speech pairs. All the training data, including clean and noisy samples, go through the first DNN and then used for DNN acoustic model (the second DNN) training. Apart from these experiments, we also experiment with the semi-supervised method for comparison. For the target domain data, we do not have senone labels. Hence we first decode the unlabeled target data using the Clean-DNN–HMM model and get the senone labels. Please note that the resultant senone labels do have inevitable errors. The adapted model, namely Semi-Ada-DNN–HMM, is then obtained by fine-tuning the Clean-DNN–HMM acoustic model using these labels. The Semi-Ada-DNN–HMM model is used to test the target domain test data.

Table 1 shows the experimental results. From the results, we notice that the Clean-DNN–HMM model, which is trained using clean data, performs badly under noisy and channel mismatch conditions. The word error rate sharply increases from 3.36% to 50.73% when the system encounters both noise and channel distortions. Meanwhile, we clearly observe that the DDA–DNN–HMM model consistently reduces the word error rates for all testing subsets. Especially for the most challenging condition, i.e., Subset D (with both noise and channel distortion), the WER is significantly dropped from 50.73% to 34.55%. In average, DDA–DNN–HMM

Please cite this article as: S. Sun et al., An unsupervised deep domain adaptation approach for robust speech recognition, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2016.11.063

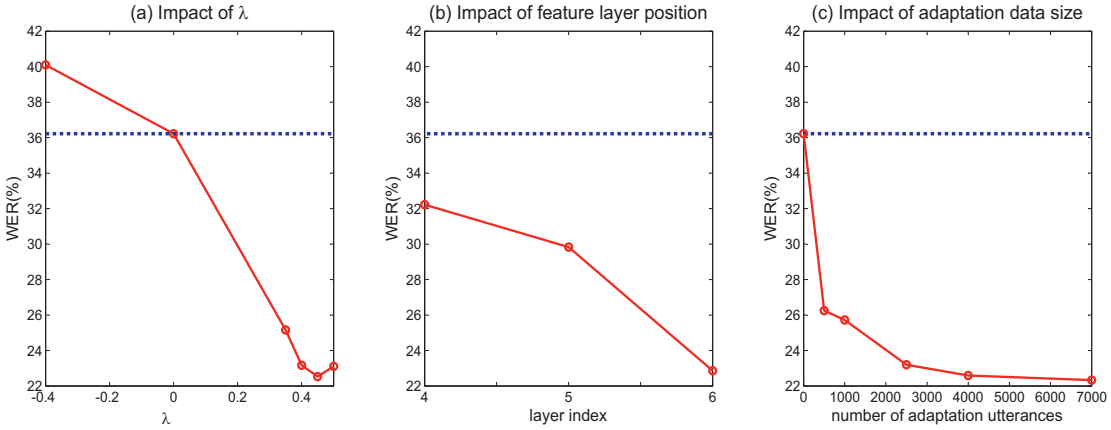


Fig. 3. Relationship between WER and (a) hyper-parameter λ, (b) position of feature representation layer and (c) the amount of adaptation data. For comparison, the blue dotted line represents the WER of Clean-DNN–HMM.

Table 1
Experimental results for clean condition training with multi-condition test on Aurora-4 in terms of WER (Word Error Rate). The hyper-parameter $\lambda = 0.45$ for DDA–DNN–HMM.

Model	A	B	C	D	Avg.
Clean-DNN–HMM	3.36	29.74	21.02	50.73	36.22
DDA–DNN–HMM	3.24	14.52	17.82	34.55	22.53
Semi-Ada-DNN–HMM	4.13	17.55	15.67	37.73	25.11
DNN-PP [35]	5.1	12.0	10.5	29.0	18.7

achieves relative 37.8% WER reduction (from 36.22% to 22.53%). Our approach is even better than the Semi-Ada-DNN–HMM model. This is because of the inevitablely wrong senone labels used for model fine-tuning in the semi-supervised approach. The average WER of DDA–DNN–HMM is even close to DNN-PP [35], a method that needs pairwise clean-noisy data for front-end speech enhancement.

4.2. Impact of hyper-parameters

We also investigate the impacts of hyper-parameters λ, the position of feature representation layer **f** and the amount of adaptation data. Their impacts are depicted in Fig. 3. Fig. 3(a) shows how λ affects the average WER. When $\lambda = 0$, the DDA–DNN–HMM model becomes the Clean-DNN–HMM model, in which the domain predictor is not working. We can see that WER goes down with the increase of λ and the lowest WER is achieved when $\lambda = 0.45$. On the contrary, when we set λ a value below zero, WER increases. This is because the domain difference is enlarged when λ is set to a negative value, as seen in Eq. (6). Another factor which may affect the DDA–DNN–HMM acoustic model is the position where we put the feature layer **f**. If we regard the G_f and G_y as an whole network and change the position of feature representation layer from top (near to softmax layer of G_y) to down (near to the input of G_f), we find that WER increases as shown in Fig. 3(b). Fig. 3(c) shows the relationship between WER and the amount of adaptation data. We find that the performance improves with the increase of adaptation data. But beyond 4000 adaptation utterances, the performance gain becomes very small.

4.3. Multi-condition training with surprise noise testing

As we pointed out in Section 2, multi-condition training is an effective approach to improve the robustness of an ASR system. This is achieved by training the acoustic model using contaminated speech. Hence the distributions of the training data and test data become identical or similar. However, in real-world, multi-condition training cannot cover all types of contamination (noise or channel distortion). We carry out an experiment to check if the DDA approach still works when the multi-condition trained ASR system encounters some surprise types of noise. In the experiment, test data is derived by adding three kinds of new noise to the clean test data with 5–10 dB SNR[3]. The multi-condition DNN–HMM, denoted as MultiCon-DNN–HMM, is trained only using the multi-condition training data from Aurora-4. The DDA–DNN–HMM is trained using the multi-condition training and 3000 noisy utterances corrupted by the three new noises. The network is the same with that in Section 4.1. Results are summarized in Table 2. We notice that multi-condition training is quite effective and the WER of MultiCon-DNN–HMM is significantly decreased as compared with the Clean-DNN–HMM in Table 2. But with the DDA approach, the WER is further reduced from 8.22% to 7.45% and relative WER reduction of 9.36% is thus achieved.

Table 2
Experimental results for multi-condition training with surprise noise testing on Aurora-4.

Model	WER (%)
MultiCon-DNN–HMM	8.22
DDA–DNN–HMM	7.45

5. Experiments for domain shift

As we discussed in Section 1, real-world speech is heterogeneous with different genres. We test the proposed DDA approach to see if it shows robustness when the speech recognizer is used in another domain.

In this experiment, we regard the WSJ [46] and Librispeech [47] corpus as data from different "domains". The WSJ0 and WSJ1 corpus[4] consist primarily of read speech with texts drawn from a machine-readable corpus of Wall Street Journal news text. WSJ0 includes a 5000-word text while WSJ1 includes a 20,000-word text. Each utterance was recorded in two channels: a high-quality "primary" microphone (a head-mounted, noise-canceling Sennheiser HMD410), and an additional microphone (desk-mounted Crown or other). The total duration of WSJ0 and WSJ1 are about 80 h. LibriSpeech is a 1000-h corpus derived

[3] These three types of noise are from another noise dataset and they are totally different with the noises in Aurora-4.
[4] The WSJ corpus contains WSJ0 and WSJ1.

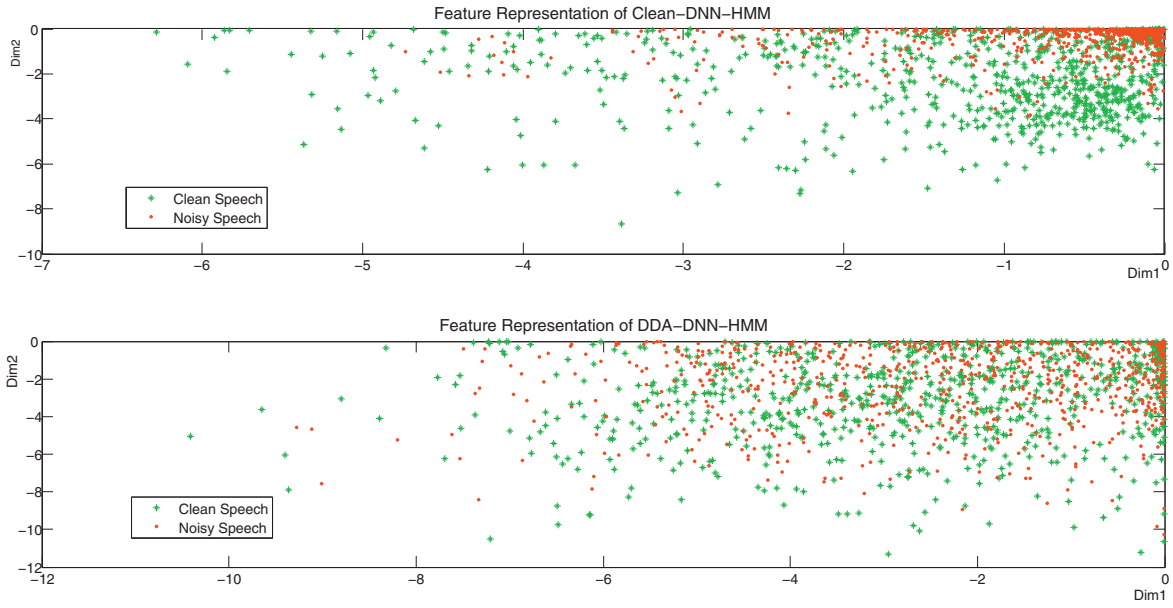


Fig. 4. Comparison of learned feature representations of Clean-DNN–HMM and DDA–DNN–HMM. The top figure is obtained by feeding the clean and corresponding noisy speech to the Clean-DNN–HMM acoustic model described in Section 4.1. The bottom figure is obtained by feeding the same clean and noisy speech to the DDA–DNN–HMM acoustic model. We only visualize two dimensions for clarity. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

Table 3
Experimental results for domain shift. The DDA–DNN–HMM acoustic model is trained using 80 h WSJ labeled data and 30 h LibriSpeech unlabeled data.

Model	WER (%)
Baseline	31.19
DDA–DNN–HMM	29.40

from audiobooks that are part of the LibriVox Project. The WSJ and LibriSpeech corpus can be used to train large vocabulary continuous speech recognition (LVCSR) acoustic models.

We first train a GMM–HMM acoustic model according to the configuration in [47], resulting in 3414 senones. Then, we train a DNN–HMM acoustic model using the 80-h WSJ data as a baseline system. After that, we train a DDA–DNN–HMM acoustic model using 80-h WSJ data (with senone labels) and 40-h adaptation data from Librispeech (without senone labels) out of 500-h "train-other-500" subset. The DNN has the same topology with the one used in Section 4. We use the 5.4-h Librispeech "test-other" set for testing. Table 3 shows the results on this test set. We can see that about 6.9% relative WER reduction is achieved when the DDA approach is used. This confirms that the proposed approach shows robustness to domain shift.

6. Analysis

As we mentioned in Section 3.1, our purpose is to learn domain-invariant feature representations which have the same or similar distributions in the source and the target domains. In our experiments, we regard the training data as the target domain and the test data as the target domain. The learned feature representations, denoted as **f** in Fig. 1, can be visualized for analysis. The dimension of this representation is 1024 in our model and we randomly choose two dimensions to visualize. To this end, we feed some clean speech frames and corresponding noisy speech frames to Clean-DNN–HMM and DDA–DNN–HMM models, respectively, discussed in Section 4.1 and the two feature dimensions are plotted in Fig. 4. From the top figure in Fig. 4, it is obvious that the representations of clean speech (denoted as red points) and noisy speech (denoted as green points) obtained by Clean-DNN–HMM acoustic model have very different distributions, which shows the mismatch between the training and test data. In contrast, this difference in distributions clearly becomes smaller for DDA–DNN–HMM, in which the deep domain adaptation approach effectively narrows the training–testing mismatch.

7. Conclusion

In this paper, we has addressed the training–testing mismatch problem in speech recognition using an unsupervised deep domain adaptation approach. Through a multi-task learning framework, a deep neural network feature extractor is learned by minimizing the loss of the phoneme classifier (main task) and to maximize the loss of the domain classifier (second task) at the same time. Specifically, during the acoustic model training, the domain classifier tries to eliminate the differences of data distribution between the source and the target domains. This approach significantly improves the performance of DNN acoustic model using some unlabeled data from the new domain. When evaluated in the "clean condition training and multi-condition testing" scenario on Aurora-4 corpus, the proposed approach decreases the word error rate from 36.22% to 22.53%, with 37.8% relative error reduction. In the domain shift experiment, the approach achieves 6.9% relative word error rate reduction. Analysis shows that the performance gain comes from the elimination of the mismatches between the distributions of the training and testing data. In the future work, we plan to implement the domain adaptation approach in convolutional neural network (CNN) [48] and recurrent neural networks (RNN) [49] that have shown superior performances in speech recognition. We also want to investigate the performances if treating different types of noises as different domains in the DDA framework. We notice that a recent multi-tasking training (MTL) approach has similar idea with our proposed DDA approach. In [50], an MTL approach is proposed to simultaneously predict the class label and the clean speech from the noisy speech input.

Please cite this article as: S. Sun et al., An unsupervised deep domain adaptation approach for robust speech recognition, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2016.11.063

We plan to experimentally compare the DDA approach with this MTL approach in our future work.

Acknowledgments

We would like to thank Yaroslav Ganin for the constructive discussions when performing this study. This work was supported by the National Natural Science Foundation of China (Grant No. 61571363) and The National High Technology Research and Development Program of China (Grant No. 2015AA016402).

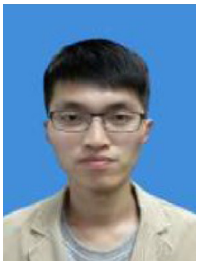
References

[1] D.Y. Li Deng, Deep Learning: Methods and Applications, Technical Report, 2014.
[2] G. Hinton, L. Deng, D. Yu, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.S.G. Dahl, B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition, IEEE Signal Process. Mag. 29 (6) (2012) 82–97.
[3] E. Trentin, M. Gori, A survey of hybrid ANN/HMM models for automatic speech recognition, Neurocomputing 37 (1) (2001) 91–126.
[4] G.E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.
[5] G.E. Dahl, T.N. Sainath, G.E. Hinton, Improving deep neural networks for LVCSR using rectified linear units and dropout, in: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2013, pp. 8609–8613.
[6] M. Xin, H. Zhang, H. Wang, M. Sun, D. Yuan, Arch: adaptive recurrent-convolutional hybrid networks for long-term action recognition, Neurocomputing 178 (2016) 87–102.
[7] P. Miao, Y. Shen, Y. Li, L. Bao, Finite-time recurrent neural networks for solving nonlinear optimization problems and their application, Neurocomputing 177 (2016) 120–129.
[8] M.S. Ali, S. Saravanan, Robust finite-time H^{∞} Control for a class of uncertain switched neural networks of neutral-type with distributed time varying delays, Neurocomputing 177 (2016) 454–468.
[9] N. Nedjah, F.M.G. França, M. De Gregorio, L. de Macedo Mourelle, Weightless neural systems, Neurocomputing 183 (2016) 1–2.
[10] P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, H. Hao, Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification, Neurocomputing 174 (2016) 806–814.
[11] T. Virtanen, R. Singh, B. Raj, Techniques for Noise Robustness in Automatic Speech Recognition, John Wiley & Sons, 2012.
[12] J. Li, L. Deng, Y. Gong, R. Haeb-Umbach, An overview of noise-robust automatic speech recognition, IEEE/ACM Trans. Audio Speech Lang. Process. 22 (4) (2014) 745–777.
[13] V. Peddinti, G. Chen, D. Povey, S. Khudanpur, Reverberation robust acoustic modeling using i-vectors with time delay neural networks, in: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, ISCA, 2015.
[14] K. Kinoshita, M. Delcroix, S. Gannot, E.A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, T. Yoshioka, A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research, EURASIP J. Adv. Signal Process. 2016 (1) (2016) 1–19.
[15] H.-G. Hirsch, D. Pearce, The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions, in: Proceedings of the ASR2000-Automatic Speech Recognition: Challenges for the new Millennium ISCA Tutorial and Research Workshop (ITRW), 2000.
[16] B. Li, Noise-Robust Speech Recognition Using Deep Neural Network, National University of Singapore, 2014 Ph.D. thesis.
[17] M.L. Seltzer, D. Yu, Y. Wang, An investigation of deep neural networks for noise robust speech recognition, in: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2013, pp. 7398–7402.
[18] Y. Qian, T. Tan, D. Yu, An investigation into using parallel data for far–field speech recognition, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5725–5729.
[19] T. Virtanen, R. Singh, B. Raj, Techniques for noise robustness in automatic speech recognition, John Wiley & Sons, 2012.
[20] U. Remes, K.J. Palomaki, M. Kurimo, Robust automatic speech recognition using acoustic model adaptation prior to missing feature reconstruction, in: Proceedings of the 2009 Seventeenth European Signal Processing Conference, IEEE, 2009, pp. 535–539.
[21] V. Gupta, P. Kenny, P. Ouellet, T. Stafylakis, I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription, in: Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014, pp. 6334–6338.
[22] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, J.W. Vaughan, A theory of learning from different domains, Mach. Learn. 79 (1–2) (2010) 151–175.
[23] Y. Ganin, V. Lempitsky, Unsupervised domain adaptation by backpropagation, in: Proceedings of the Thirty Second International Conference on Machine Learning (ICML-15), JMLR, 2015, pp. 1180–1189.
[24] M. Mohri, F. Pereira, M. Riley, Weighted finite-state transducers in speech recognition, Comput. Speech Lang. 16 (1) (2002) 69–88.
[25] M. Westphal, The use of cepstral means in conversational speech recognition., in: Proceedings of the European Conference on Speech Communication and Technology, EUROSPEECH, 1997.
[26] S. Molau, F. Hilger, H. Ney, Feature space normalization in adverse acoustic conditions, in: Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'03), vol. 1, IEEE, 2003, pp. I–656.
[27] F. Hilger, H. Ney, Quantile based histogram equalization for noise robust large vocabulary speech recognition, IEEE Trans. Acoust. Speech Signal Process. 14 (3) (2006) 845–854.
[28] S.F. Boll, Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. Acoust. Speech Signal Process. 27 (2) (1979) 113–120.
[29] J. Koehler, N. Morgan, H. Hermansky, H.G. Hirsch, G. Tong, Integrating rasta-plp into speech recognition, in: Proceedings of the 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-94, vol.1, IEEE, 1994, pp. 1–421.
[30] J.-L. Gauvain, C.-H. Lee, Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains, IEEE Trans. Speech Audio Process. 2 (2) (1994) 291–298.
[31] M.J. Gales, Maximum likelihood linear transformations for HMM-based speech recognition, Comput. Speech Lang. 12 (2) (1998) 75–98.
[32] M. Gales, S. Young, Parallel Model Combination for Speech Recognition in Noise, University of Cambridge, Department of Engineering, 1993.
[33] P.J. Moreno, B. Raj, R.M. Stern, A vector Taylor series approach for environment-independent speech recognition, in: Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-96, vol.2, IEEE, 1996, pp. 733–736.
[34] A.L. Maas, Q.V. Le, T.M. O'Neil, O. Vinyals, P. Nguyen, A.Y. Ng, Recurrent neural networks for noise reduction in robust ASR, in: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2012, pp. 22–25.
[35] J. Du, Q. Wang, T. Gao, Y. Xu, L.-R. Dai, C.-H. Lee, Robust speech recognition with speech enhanced deep neural networks, in: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2014, pp. 616–620.
[36] T. Gao, J. Du, L.-R. Dai, C.-H. Lee, Joint training of front-end and back-end deep neural networks for robust speech recognition, in: Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015, pp. 4375–4379.
[37] K.H. Lee, S.J. Kang, W.H. Kang, N.S. Kim, Two-stage noise aware training using asymmetric deep denoising autoencoder, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5765–5769.
[38] Y. Qian, M. Yin, Y. You, K. Yu, Multi-task joint-learning of deep neural networks for robust speech recognition, in: Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), IEEE, 2015, pp. 310–316.
[39] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, M.S. Lew, Deep learning for visual understanding: a review, Neurocomputing 187 (2016) 27–48. Recent Developments on Deep Big Vision
[40] C. Hong, J. Yu, J. Wan, D. Tao, M. Wang, Multimodal deep autoencoder for human pose recovery, IEEE Trans. Image Process. 24 (2015) 5659–5670.
[41] C. Hong, J. Yu, J. You, X. Chen, D. Tao, Multi-view ensemble manifold regularization for 3d object recognition, Inf. Sci. 320 (2015) 395–405.
[42] L. Bottou, Large-scale machine learning with stochastic gradient descent, in: Proceedings of the International Conference on Computational Statistics, COMPSTAT'2010, Springer, 2010, pp. 177–186.
[43] P.F. Brown, P.V. Desouza, R.L. Mercer, V.J.D. Pietra, J.C. Lai, Class-based n-gram models of natural language, Comput. Linguist. 18 (4) (1992) 467–479.
[44] 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: Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, EPFL-CONF-192584, IEEE Signal Processing Society, 2011.
[45] S.J. Young, N. Russell, J. Thornton, Token Passing: A Simple Conceptual Model for Connected Speech Recognition Systems, Cambridge University, Engineering Department, Cambridge, UK, 1989.
[46] D.B. Paul, J.M. Baker, The design for the wall street journal-based CSR corpus, in: Proceedings of the Workshop on Speech and Natural Language, Association for Computational Linguistics, 1992, pp. 357–362.
[47] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, Librispeech: An ASR corpus based on public domain audio books, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5206–5210.
[48] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, G. Penn, Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition, in: Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2012, pp. 4277–4280.
[49] A. Graves, A.-r. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2013, pp. 6645–6649.
[50] B. Li, T.N. Sainath, R.J. Weiss, K.W. Wilson, M. Bacchiani, Neural network adaptive beamforming for robust multichannel speech recognition, in: Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech, 2016.

Please cite this article as: S. Sun et al., An unsupervised deep domain adaptation approach for robust speech recognition, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2016.11.063

[51] J. Yu, X. Yang, F. Gao, D. Tao, Deep multimodal distance metric learning using click constraints for image ranking, IEEE Trans. Cybern. PP (99) (2016) 1–11.
[52] J. Yu, B. Zhang, Z. Kuang, D. Lin, J. Fan, iPrivacy: image privacy protection by identifying sensitive objects via deep multi-task learning, IEEE Trans. Inf. Forensics Security PP (99) (2016) 1.

Sining Sun received the B.S. degree Computer Science and Technology from the Northwestern Polytechnical University, Xian, China, in 2014. Currently, he is pursuing the Ph.D. degree in the School of Computer Science, Northwestern Polytechnical University, Xian, China. His research interests include speech signal processing, robust speech recognition and machine learning.



Binbin Zhang received the B.S. degree Computer Science and Technology from the Northwestern Polytechnical University, Xian, China, in 2014. Currently, he is pursuing the master degree in the School of Computer Science, Northwestern Polytechnical University, Xian, China. His research interests include deep learning, automatic speech recognition and machine learning.

Lei Xie received the Ph.D. degree in Computer Science from Northwestern Polytechnical University (NPU), Xian, China, in 2004. He is currently a Professor with School of Computer Science, NPU. From 2001 to 2002, he was with the Department of Electronics and Information Processing, Vrije Universiteit Brussel (VUB), Brussels, Belgium, as a Visiting Scientist. From 2004 to 2006, he was a Senior Research Associate in the Center for Media Technology (RCMT), School of Creative Media, City University of Hong Kong, Hong Kong. From 2006 to 2007, he was a Postdoctoral Fellow in the Human Computer Communications Laboratory (HCCL), Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong. He has published more than 120 papers in major journals and proceedings, such as the IEEE Transactions on Audio, Speech, and Language Processing, IEEE Transactions on Multimedia, Information Sciences, Pattern Recognition, ACM Multimedia, ACL, INTERSPEECH, and ICASSP. His current research interests include speech and language processing, multimedia and human–computer interaction.

Yanning Zhang is currently a professor in the School of Computer Science, Northwestern Polytechnical University, China. She received her Ph.D. from the Northwestern Polytechnical University, China in 1996. Her current research interests are in signal processing, multimedia and computer vision. Zhang has been an active member of the technical program committee of several international conferences and a reviewer of several reputed journals and conference, such as reviewer of IEEE Transactions on Systems, Man and Cybernetics (T-SMC), Pattern Recognition Letter. She has also been the organization chair of the Ninth Asian Conference on Computer Vision (ACCV09). She is currently a member of IEEE.

Please cite this article as: S. Sun et al., An unsupervised deep domain adaptation approach for robust speech recognition, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2016.11.063