DOMAIN ADAPTATION OF DNN ACOUSTIC MODELS USING KNOWLEDGE DISTILLATION

Taichi Asami, Ryo Masumura, Yoshikazu Yamaguchi, Hirokazu Masataki, Yushi Aono

NTT Media Intelligence Laboratories, NTT Corporation, Japan
{asami.taichi, masumura.ryo, yamaguchi.yoshikazu, masataki.hirokazu, aono.yushi}@lab.ntt.co.jp

ABSTRACT

Constructing deep neural network (DNN) acoustic models from limited training data is an important issue for the development of automatic speech recognition (ASR) applications that will be used in various application-specific acoustic environments. To this end, domain adaptation techniques that train a domain-matched model without overfitting by leveraging pre-constructed source models are widely used. In this paper, we propose a novel domain adaptation method for DNN acoustic models based on the knowledge distillation framework. Knowledge distillation transfers the knowledge of a teacher model to a student model and offers better generalizability of the student model by controlling the shape of posterior probability distribution of the teacher model, which was originally proposed for model compression. We apply this framework to model adaptation. Our domain adaptation method avoids overfitting of the adapted model trained on limited data by transferring the knowledge of the source model to the adapted model by distillation. Experiments show that the proposed method can effectively avoid the overfitting of convolutional neural network based acoustic models and yield lower error rates than conventional adaptation methods.

Index Terms— Speech recognition, DNN acoustic models, domain adaptation, knowledge distillation

1. INTRODUCTION

Deep neural network (DNN) based acoustic models can significantly improve the performance of automatic speech recognition (ASR) systems, which is enabling various kinds of ASR applications. Each application is designed for an application-specific acoustic environment. For example, voice search is used by the general public in various noisy conditions, applications targeted at children are used by children, and dialectal speech is input to speech assistant applications for particular local areas. This requires application-specific acoustic models to be constructed for each target acoustic environment.

An application-specific DNN acoustic model should match the target acoustic environment and offer good generalizability. However, the amount of training data for each application tends to be small (at most 10 hours in many cases) due to the costs of recording and transcription. For the practical use/development of ASR systems, avoiding the overfitting of acoustic models caused by data scarcity is an important issue. To this end, domain adaptation techniques that create domain-matched acoustic models without overfitting are widely used.

Several approaches have been investigated for the adaptation of DNN acoustic models. Limiting the number of parameters updated by retraining is a widely investigated approach. Most of the methods employing this approach extend neural networks by introducing special adaptation-oriented parameters or structures, such as linear input/hidden layer [1, 2], learning hidden unit contribution [3], and speaker dependent linear transformation network [4]. Expanding the input features is another commonly used approach. Auxiliary features that capture global domain information, such as i-vector [5,6] or speaker code [7], are input with standard acoustic feature vectors as information used in adaptation. Regularization in optimization is a third approach. This approach constrains the gradients during training by the regularization term in the objective function so that information in pre-constructed source models is retained by the trained model. Regularizers include L2 norm [8] and Kullback-Leibler divergence (KLD) [9] between source and trained models.

We focus on the regularization approach since it has a major merit; it makes no assumption as to model structure or input. This means that regularization-based adaptation methods are applicable to any existing or new models. In this paper, we propose a novel adaptation method that uses the knowledge distillation framework for regularization.

Knowledge distillation [10] was originally proposed for model compression. A big (or ensemble) model that achieves high accuracy but requires massive computation time is used as the teacher model, and a small feasible model, called the student model, is trained to imitate the behavior of the teacher model by using output of the teacher model, called the soft target, for computing the cross entropy loss function. The generalizability of the student model can be enhanced by using a temperature parameter to control the smoothness of the soft target. This framework can be viewed as regularization that constrains the gradients so that knowledge in the

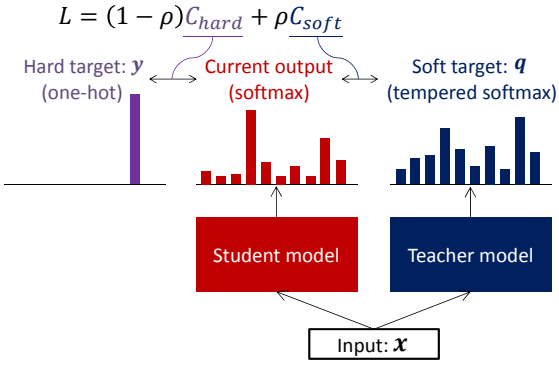

Fig. 1. Knowledge distillation. The loss function is the weighted sum of hard and soft cross entropy.

teacher model is retained by the student model, and when the teacher model is the source model of domain adaptation, this framework straightforwardly becomes regularization-based adaptation. Knowledge distillation has been used in studies of acoustic models for reducing model complexity [11–16] and initializing recurrent neural network based models [17]. To the best of our knowledge, our work is the first to apply knowledge distillation to the acoustic model adaptation task.

The remainder of this paper is organized as follows. In Section 2, we describe the knowledge distillation framework and the proposed adaptation method. Furthermore, we discuss the relationship between the proposed method and the KLD-based regularization method. Section 3 details our experiments on real ASR application data. Finally, Section 4 concludes this paper.

2. KNOWLEDGE DISTILLATION-BASED ADAPTATION

2.1. Knowledge distillation

Knowledge distillation trains the student model to imitate the teacher model. A schematic diagram of knowledge distillation is shown in Fig. 1. The student model is trained so as to minimize following loss function:

$$L = (1-\rho)C_{\text{hard}}(\boldsymbol{x}, \boldsymbol{y}) + \rho C_{\text{soft}}(\boldsymbol{x}, \boldsymbol{q}), \tag{1}$$

$$C_{\text{hard}}(\boldsymbol{x}, \boldsymbol{y}) = -\sum_{i=1}^{K} y_i \log p_i(\boldsymbol{x}), \tag{2}$$

$$C_{\text{soft}}(\boldsymbol{x}, \boldsymbol{q}) = -\sum_{i=1}^{K} q_i \log p_i(\boldsymbol{x}), \tag{3}$$

where ρ is the weight of the hard and soft cross entropy losses, K is the number of output classes, $\boldsymbol{x}$ is an input feature, $p_i(\boldsymbol{x})$ is output (softmax) probability of the i-th class of the student model, $\boldsymbol{y}$, called hard target, is a K dimensional one-hot vector in which y_i is 1 if the i-th class is correct and 0 otherwise. $\boldsymbol{q}$, called soft target, is also a K dimensional vector, and q_i is the *tempered* softmax probability of the i-th class of the teacher model, which is computed as follows:

$$q_i = \frac{\exp(z_i(\boldsymbol{x})/T)}{\sum_{j=1}^{K} \exp(z_j(\boldsymbol{x})/T)}, \tag{4}$$

where $z_i(\boldsymbol{x})$ is pre-softmax output of the i-th class of the teacher model and T is a temperature. As T becomes large, soft target $\boldsymbol{q}$ approaches a uniform distribution. Note that T^2 should be multiplied to the gradient of the second term of Eq. (1) in backpropagation. This keeps the balance between the contribution of C_{hard} and C_{soft} when T is changed since the magnitude of the gradient is scaled by $1/T^2$.

Hard target $\boldsymbol{y}$ is a one-hot vector that indicates the correct class. On the other hand, soft target $\boldsymbol{q}$ is a smooth probability distribution, which contains the knowledge of the teacher model, i.e. not only the correct class but also similarity/correlation between classes. As described in [10], learning the similarity between classes from the teacher model is important to raise the generalizability of the student model. Temperature T allows us to control the importance of the class similarity information in training. When T is set larger than 1, small probabilities of non-target classes are emphasized and the class similarity information is more strongly learned by the student model.

Usually, for model compression, a big or ensemble model that achieves high accuracy but has infeasible computation complexity is used as the teacher model, and a small feasible model is used as the student model. By this setting, the student model achieves both high accuracy and feasible computation complexity.

2.2. Adaptation by distillation

The first term of Eq. (1) is the standard cross entropy loss function. Thus, the second term can be viewed as a regularization term that constrains the student model so as to imitate the teacher model. From this viewpoint, knowledge distillation can be straightforwardly applied to model adaptation.

In the acoustic model adaptation task, we have a source model, θ_{source}, and a small amount (N frames) of target domain data, $\{\boldsymbol{x}_n, \boldsymbol{y}_n\} (1 \leq n \leq N)$. The goal is to obtain an adapted model, θ_{adapted}, that matches the target domain and has better generalizability. Domain adaptation by knowledge distillation is conducted as follows:

1. Initialize θ_{adapted} as a copy of θ_{source}.
2. Set θ_{adapted} as the student model and θ_{source} as the teacher model, then train θ_{adapted} on $\{\boldsymbol{x}_n, \boldsymbol{y}_n\}$ so as to minimize Eq. (1).

Table 1. Size of adaptation and test sets. All data sets contain both male and female utterances, and speakers in the test sets were not included in the adaptation sets for speaker-open testing.

Domain	Adaptation set size	Test set size
Dialect 1 (Osaka)	0.5h/1h/3h/5h	1h
Dialect 2 (Fukuoka)	0.5h/1h/3h/5h	1h
Children	0.2h/0.5h/1h/2h	1h

The first term of Eq. (1) allows the adapted model to be trained so that it can accurately classify in-domain data while the second term, simultaneously, allows the model to acquire generalizability by learning the class similarity information from the source model.

2.3. Relationship to KLD regularization

The KLD regularization proposed in [9] minimizes following loss function:

$$L_{\mathrm{KLD}} = -(1-\rho)\sum_{i=1}^{K} y_i \log p_i(\boldsymbol{x}) - \rho \sum_{i=1}^{K} \tilde{q}_i \log p_i(\boldsymbol{x}), \quad (5)$$

$$\tilde{q}_i = \frac{\exp(z_i(\boldsymbol{x}))}{\sum_{j=1}^{K} \exp(z_j(\boldsymbol{x}))}. \quad (6)$$

Obviously, Eq. (5) is equivalent to Eq. (1) when T in Eq. (4) is set to 1. Therefore, the temperature parameter of our distillation-based adaptation is an extension of KLD regularization, and enables the generalizability of the adapted model to be controlled.

3. EXPERIMENTS

3.1. Data

Training set of the source model consisted of 1200 hours of Japanese utterances recorded in various acoustic environments, and includes real data of voice search application, call center recordings, Corpus of Spontaneous Japanese (academic presentations) [18], and Japanese Newspaper Article Sentences (reading newspapers) [19]. The source model already has robustness to noise/channel variability, but is vulnerable to dialectal and children's speech since the training set does not contain them.

Acoustic model adaptation to three domains, two Japanese dialects (Osaka and Fukuoka) and children's speech, were investigated. 5 hours of dialectal speech and 2 hours of children's speech were recorded for domain adaptation, and speaker-open 1 hour test sets were separately recorded for each domain. Table 1 summarizes the adaptation and test sets of each domain. As shown in Table 1, subsets of each adaptation set were randomly selected and also used for domain adaptation to investigate the relationship between adaptation set size and recognition performance.

All utterances were recorded with 16kHz sampling rate and 16bit resolution.

Table 2. Structure of the NiN-CNN acoustic model in this study. "conv" is a convolutional layer, "pool" is a max pooling layer, and "fc" is a fully-connected layer. Especially "conv{1,2}b" corresponds to the NiN architecture [20].

Layer	Filter size	Input size	#Feature maps
		40x11	3
conv1a	5x11	36x1	180
conv1b	1x1	36x1	180
pool1	2x1	18x1	180
conv2a	5x1	14x1	180
conv2b	1x1	14x1	180
pool2	2x1	7x1	180
fc1		2048	
fc2		2048	
softmax		3072	

3.2. Conditions

The source (baseline) model was a convolutional neural network with network-in-network architecture (NiN-CNN) acoustic model [20]. Its structure is detailed in Table 2. The sigmoid activation function was used for all hidden layers. The acoustic feature consisted of 40 log mel filterbank coefficients appended with delta and acceleration coefficients. Each static and dynamic component was spliced within 11 frames and treated as a feature map, i.e. 3 feature maps of 40x11 size were input to the acoustic model. The source model was trained on the data described in the previous section. The same 3-gram language model was used in all conditions; it has 520K vocabulary size and was trained on various text corpora with a total of 2.3G words. Decoding was performed by the WFST-based decoder VoiceRex [21, 22].

To evaluate our distillation-based adaptation, the following methods were compared:

- **Simple retraining**: All parameters of the source model were simply retrained on the adaptation set.
- **KLD regularization** [9]: The source model was adapted by using Eq. (5). This was implemented by setting the temperature parameter, T in Eq. (4), to 1. Three values (0.1, 0.3 or 0.5) of the weight parameter, ρ, were investigated.
- **Distillation-based adaptation (proposed)**: The source model was adapted by using Eq. (1) with the setting of $T > 1$. Two values (3 or 5) of the temperature were investigated and weight ρ was fixed to 0.1 that yielded the lowest average error rate on KLD regularization.

Table 3. %CER in all conditions. The baseline performance in each domain is shown in the first row. lr is the initial learning rate, and ρ and T are the weight and the temperature parameters of knowledge distillation, respectively. Note that $\rho = 0.0$ means use of simple retraining and $T = 1$ means use of KLD regularization [9] as described in Section 2.3. The best results in each column are highlighted in bold.

				Dialect 1 (baseline: 32.6)				Dialect 2 (baseline: 21.1)				Children (baseline: 13.7)			
Method	ρ	T	lr	0.5h	1h	3h	5h	0.5h	1h	3h	5h	0.2h	0.5h	1h	2h
Retraining	0.0	-	0.08	27.7	26.8	25.8	25.8	18.0	18.2	18.1	18.1	15.7	14.3	13.7	13.0
Retraining	0.0	-	0.008	26.2	25.2	23.6	23.3	16.8	16.4	15.8	15.3	13.8	12.6	12.4	12.3
KLD	0.1	1	0.08	26.8	26.4	25.5	25.4	17.8	18.1	16.6	16.4	15.1	14.0	13.3	12.7
KLD	0.1	1	0.008	26.7	25.6	24.1	23.2	17.1	16.4	15.5	**14.9**	14.3	13.4	12.9	12.8
KLD	0.3	1	0.08	27.3	27.1	26.1	25.6	18.0	17.5	16.6	16.3	14.7	13.8	13.2	12.9
KLD	0.5	1	0.08	28.3	28.4	26.9	25.7	18.5	18.1	17.1	17.0	14.4	13.8	13.7	13.4
Distillation	0.1	3	0.08	**25.6**	25.1	24.2	23.7	16.5	16.2	15.6	15.7	13.2	12.2	**11.8**	**11.6**
Distillation	0.1	3	0.008	25.7	**24.7**	**23.3**	**23.1**	**16.2**	**15.4**	**15.3**	**14.9**	**12.8**	12.0	11.9	11.7
Distillation	0.1	5	0.08	25.7	25.3	24.3	23.9	16.9	16.5	16.3	15.6	13.0	**11.9**	**11.8**	11.7

Stochastic gradient descent with momentum (= 0.9) was used as the optimization algorithm for all methods. Other than the standard value (0.08) of initial learning rate, a small value (0.008) was investigated for the best conditions of each method. 10% of the adaptation set was separated and used as the held-out set for cross-validation (CV). Learning rate was halved when CV frame accuracy decreased, and training was stopped when the learning rate became smaller than 0.0008, and the model achieving the best CV accuracy was used for the test.

The source model was adapted to each adaptation set in Table 1 by the above methods, and the adapted models were evaluated by the character error rate (CER) on the corresponding test set.

3.3. Results

Table 3 shows the CERs for all conditions. Distillation-based adaptation consistently achieved lower error rates than KLD regularization and simple retraining in all conditions. This means that distillation-based adaptation can achieve adapted models with better generalizability when the training data for domain adaptation is limited. Especially in the condition of "Children" domain with 0.2h adaptation set, while simple retraining and KLD regularization could not avoid overfitting and the performance was worse than the baseline, distillation-based adaptation could achieve better results without overfitting. These results indicate that generalizability control by the temperature parameter is effective in avoiding the overfitting caused by data scarcity.

In many conditions, the small initial learning rate yielded lower CERs than the standard value. However, sensitivity to the initial learning rate was reduced when regularization-based methods (both KLD and distillation) were used. This means that local optima of the loss function were moved to appropriate position by the regularization term. This is a good property for practical applications where time and computation resources are limited and dedicated hyperparameter tuning is difficult.

Though further investigation is needed for confirmation, it seems that there is correlation between the baseline performance and the gain by distillation-based adaptation. It can be reasonably argued that the reliability of the soft target impacted the performance of the adapted model, while distillation-based adaptation could achieve useful gain even with the relatively high baseline CER of "Dialect 1" domain.

4. CONCLUSIONS

In this paper we proposed a novel acoustic model adaptation method based on the knowledge distillation framework. The proposed method uses knowledge distillation as regularization that constrains the adapted model into imitating the source model in order to avoid overfitting. We also showed that our distillation-based adaptation is an extension of KLD regularization, and its use of a temperature parameter offers control over the generalizability of the adapted model.

Experiments on three real acoustic domains (dialectal and children's speech) showed that distillation-based adaptation could effectively avoid overfitting when training data was scarce and achieve lower error rates than conventional KLD regularization and simple retraining.

Though we examined adaptation of the NiN-CNN acoustic model in this work, other kinds of neural networks, such as a long short-term memory or a highway network, can be effectively enhanced by adoption of distillation-based adaptation since it makes no assumption as to the model structure used.

5. REFERENCES

[1] "Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system," in *Proc. of EUROSPEECH*, 1995, pp. 2171–2174.

[2] R. Gemello, F. Mana, S. Scanzio, P. Laface, and R.D. Mori, "Linear hidden transformations for adaptation of hybrid ANN/HMM models," *Speech Communication*, vol. 49(10-11), pp. 827–835, 2007.

[3] P. Swietojanski and S. Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *Proc. of SLT*, 2014, pp. 171–176.

[4] T. Ochiai, S. Matsuda, H. Watanabe, X. Lu, C. Hori, and S. Katagiri, "Speaker adaptive training for deep neural networks embedding linear transformation networks," in *Proc. of ICASSP*, 2015, pp. 4605–4609.

[5] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proc. of ASRU*, 2013, pp. 55–59.

[6] M. Delcroix, K. Kinoshita, A. Ogawa, T. Yoshioka, D. Tran, and T. Nakatani, "Context adaptive neural network for rapid adaptation of deep CNN based acoustic models," in *Proc. of INTERSPEECH*, 2016, pp. 1573–1577.

[7] O. Abdel-Hamid and H. Jiang, "Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code," in *Proc. of ICASSP*, 2013, pp. 7942–7946.

[8] H. Liao, "Speaker adaptation of context dependent deep neural networks," in *Proc. of ICASSP*, 2013, pp. 7947–7951.

[9] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Proc. of ICASSP*, 2013, pp. 7893–7897.

[10] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. of NIPS Deep Learning and Representation Learning Workshop*, 2014.

[11] J. Li, R. Zhao, J.T. Huang, and Y. Gong, "Learning small-size DNN with output-distribution-based criteria," in *Proc. of INTERSPEECH*, 2014, pp. 1910–1914.

[12] R. Price, K. Iso, and K. Shinoda, "Wise teachers train better DNN acoustic models," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 10, pp. 1–19, 2016.

[13] W. Chan, N.R. Ke, and I. Lane, "Transferring knowledge from a RNN to a DNN," in *Proc. of INTERSPEECH*, 2015, pp. 3264–3268.

[14] Y. Chebotar and A. Waters, "Distilling knowledge from ensembles of neural networks for speech recognition," in *Proc. of INTERSPEECH*, 2016, pp. 3439–3443.

[15] G. Tucker, M. Wu, M. Sun, S. Panchapagesan, G. Fu, and S. Vitaladevuni, "Model compression applied to small-footprint keyword spotting," in *Proc. of INTERSPEECH*, 2016, pp. 1878–1882.

[16] K. Markov and T. Matsui, "Robust speech recognition using generalized distillation framework," in *Proc. of INTERSPEECH*, 2016, pp. 2364–2368.

[17] Z. Tang, D. Wang, and Z. Zhang, "Recurrent neural network training with dark knowledge transfer," in *Proc. of ICASSP*, 2016, pp. 5900–5904.

[18] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *Proc. of LREC*, 2000, pp. 947–952.

[19] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, and T. Kobayashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of the Acoustical Society of Japan (E)*, vol. 20(3), pp. 199–206, 1999.

[20] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W.J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. of ASRU*, 2015, pp. 436–443.

[21] H. Masataki, D. Shibata, Y. Nakazawa, S. Kobashikawa, A. Ogawa, and K. Ohtsuki, "VoiceRex - Spontaneous speech recognition technology for contact center conversations," *NTT Technical Review*, vol. 5(1), pp. 22–27, 2007.

[22] T. Hori, C. Hori, Y. Minami, and A. Nakamura, "Efficient WFST based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition," *IEEE Transaction on Audio, Speech and Language Processing*, vol. 15(4), pp. 1352–1365, 2007.