BLHUC: BAYESIAN LEARNING OF HIDDEN UNIT CONTRIBUTIONS FOR DEEP NEURAL NETWORK SPEAKER ADAPTATION

Xurong Xie[1,2], *Xunying Liu*[1,2], *Tan Lee*[1], *Shoukang Hu*[1], *Lan Wang*[2]

[1]Chinese University of Hong Kong, Hong Kong, China
[2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
{xrxie, tanlee}@ee.cuhk.edu.hk, {xyliu, skhu}@se.cuhk.edu.hk, lan.wang@siat.ac.cn

ABSTRACT

Speaker adaptation techniques play a key role in reducing the mismatch between speech recognition systems and target users. In order to robustly learn speaker-dependent adaptation parameters, model based DNN adaptation techniques often require a significant amount of data. For example, in the commonly used learning hidden unit contributions (LHUC) based DNN adaptation, speaker-dependent high-dimensional hidden layer output scaling vectors are used. When limited adaptation data are available, the standard LHUC is prone to over-fitting and poor generalization. To address the issue, Bayesian learning of hidden unit contributions (BLHUC) is proposed in this paper. A posterior distribution over the LHUC scaling vectors is used to explicitly model the uncertainty associated with the adaptation parameters. An efficient variational inference based approach is adopted to estimate the LHUC parameter posterior distribution. Experiments conducted on a 300-hour Switchboard setup showed that the proposed BLHUC method outperformed the baseline speaker-independent DNN systems and LHUC adapted DNN systems by up to 1.4% and 1.1% absolute reductions of word error rate respectively, when only using 1 utterance of adaptation data from each speaker. Consistent performance improvements were also obtained over the baseline, LHUC adapted and LHUC SAT systems when increasing the amount of adaptation data.

Index Terms— Bayesian learning, LHUC, speaker adaptation

1. INTRODUCTION

Speaker adaptation techniques play a vital role in speech recognition systems to reduce the mismatch against target users. For current deep neural network (DNN) based speech recognition systems,
three categories of speaker adaptation methods can be used. In auxiliary input feature based DNN adaptation techniques, such as i-vectors [1, 2, 3], speaker codes [4, 5, 6], and bottleneck features [7], speaker-dependent (SD) characteristics are encoded in a compact vector used to facilitate model adaptation. Inspired by the speaker cluster based adaptation techniques originally proposed for GMM-HMM systems [8], interpolation based DNN adaptation methods with multiple basis of sub-network hidden outputs have been proposed [9, 10]. Model based DNN speaker adaptation techniques directly estimate SD parameters represented by, for example, speaker-dependent hidden layers, or input layer linear transforms applied by the input features of each speaker that are either separately learned from GMM-HMM systems [11, 12, 13] or jointly estimated with the remaining DNN parameters [14, 15]. It is also possible to use a set of scaling vectors to learn the varying hidden layer units contributions (LHUC) [16, 17, 18] among diverse speakers.

In common with other model based adaptation techniques that require a significant number of SD parameters to be robustly estimated, the standard LHUC based DNN adaptation method can lead to over-fitting and poor generalization when using limited speaker specific data. One solution to address this issue with parameter uncertainty is to use a Bayesian learning approach. In machine learning community, a series of previous research were conducted in this direction. A practical Bayesian framework for back-propagation networks was introduced in [19]. Efficient variational learning based inference was later proposed for Bayesian neural networks [20]. However, limited research has been conducted to apply Bayesian learning for speech recognition. In [21], a Bayesian recurrent neural network (RNN) using variational inference based training was evaluated on the TIMIT speech corpus. A Bayesian learning approach for RNN language models was proposed in [22].

In this paper, a novel Bayesian learning of hidden unit contributions (BLHUC) is proposed. A posterior distribution over the LHUC scaling vectors is used to explicitly model the uncertainty associated with the adaptation parameters. An efficient variational inference based approach is adopted to estimate the LHUC parameter posterior distribution. The proposed BLHUC method is investigated for both unsupervised test time DNN model speaker adaptation, and DNN model speaker adaptive training (SAT) [23]. To the best of our knowledge, this is the first work about using Bayesian learning for DNN speaker adaptation.

In the experiments, speaker-independent (SI) DNN-HMM systems, LHUC adapted DNN systems, and the proposed BLHUC adapted DNN systems were built on a 300-hour Switchboard setup, and evaluated on speech recognition task of Hub5' 00 data set. By using only the beginning 1 utterance from each speaker as adaptation data, the BLHUC adapted systems outperformed the baseline SI systems and standard LHUC adapted systems by up to 1.4% and 1.1% absolute in word error rate (WER) respectively. Consistent performance improvements were also obtained by BLHUC adaptation over the baseline SI, LHUC adapted, and LHUC SAT systems, when the adaptation data amount from each speaker was increased.

The rest of this paper is organized as follows. The standard LHUC method will be reviewed in section 2. Then, section 3 introduces the proposed BLHUC technique. The variational inference based estimation of the BLHUC speaker-dependent parameters will be described in section 4. In section 5, various DNN-HMM systems with or without LHUC/BLHUC adaptation are evaluated on the

This research was partially supported by a direct grant from Research Committee of the Chinese University of Hong Kong (CUHK), Hong Kong Research Grants Council General Research grant Fund No.14227216 and No.14200218, the CUHK grant No.4055065, Natural Science Foundation of China U1736202, and ShenZhen Fundamental Research Program JCYJ20160429184226930 and KQJSCX20170731163308665.

speech recognition tasks of the Switchboard databases. Section 6 draws the conclusion and the future works.

2. LEARNING HIDDEN UNIT CONTRIBUTIONS

The key idea of using learning hidden unit contributions (LHUC) [16] for speaker adaptation is to modify the amplitudes of DNN hidden unit activations for each speaker. This speaker-dependent (SD) modification can be explicitly parameterized by using scaling vectors. Then, for speaker s and the lth hidden layer, letting $\boldsymbol{r}^{l,s} \in \mathbb{R}^D$ denotes the parameter set of scaling vector, the hidden layer output can be computed by

$$\boldsymbol{h}^{l,s} = \xi(\boldsymbol{r}^{l,s}) \otimes \psi(\boldsymbol{W}^{l\top}\boldsymbol{h}^{l-1,s} + \boldsymbol{b}^l) \quad (1)$$

where $\boldsymbol{W}^l$ and $\boldsymbol{b}^l$ denote the DNN weight matrix and bias vector, ψ is the hidden unit activation function (which is sigmoid in this work), and $\otimes$ denotes the Hadamard product. The scaling vectors are modeled by function $\xi : \mathbb{R}^D \rightarrow \{\mathbb{R}^+\}^D$ on parameters $\boldsymbol{r}^{l,s}$. Typically, ξ can be sigmoid[16, 17], linearity or RELU [18] functions. In this work the element-wise function $\xi(\cdot) = 2\text{sigmoid}(\cdot)$ is utilized, such that the hidden unit amplitude scaling are constrained in $(0, 2)$. Figure 1 is an example of using LHUC adaptation in a DNN acoustic model.

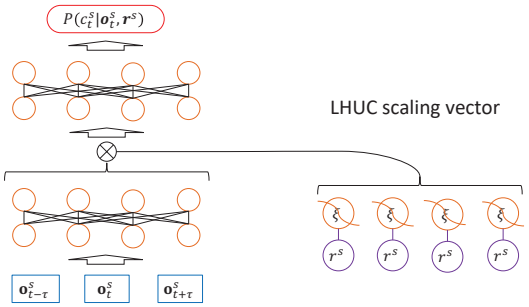

Fig. 1. An example of using LHUC adaptation in a DNN acoustic model.

When using LHUC adaptation, the hidden layer output becomes speaker-dependent by using D^l dimensional of SD parameters, which can be significantly less than the transformation based adaptation techniques. However, in practice D^l can still be large and the available adaptation data may be limited, thus it may be insufficient to give a robust estimation for the D^l dimensional parameters.

3. BAYESIAN LEARNING OF HIDDEN UNIT CONTRIBUTIONS

In standard LHUC adaptation, given adaptation data $\boldsymbol{o}^s$ and the corresponding alignment c^s for speaker s, the inference for data $\boldsymbol{o}_t^s$ is

$$P(c_t^s|\boldsymbol{o}_t^s, \boldsymbol{o}^s, c^s) = \int P(c_t^s|\boldsymbol{o}_t^s, \boldsymbol{r}^s)p(\boldsymbol{r}^s|\boldsymbol{o}^s, c^s)d\boldsymbol{r}^s \quad (2)$$

$$\approx P(c_t^s|\boldsymbol{o}_t^s, \hat{\boldsymbol{r}}^s) \quad (3)$$

where $\hat{\boldsymbol{r}}^s = \arg\max_{\boldsymbol{r}^s} P(c^s|\boldsymbol{o}^s, \boldsymbol{r}^s)$ is the maximum likelihood (ML) estimate of SD parameters $\boldsymbol{r}^s$. The ML approximation by equation (3) makes sense only when the strong assumption $p(\hat{\boldsymbol{r}}^s|\boldsymbol{o}^s, c^s) \approx 1$ on the posterior distribution makes sense. However, the limited amount of available adaptation data may lead to uncertainty on the SD parameters, thus this approximation may not be trustworthy.

Therefore, Bayesian learning of hidden unit contributions (B-LHUC) is proposed to change the deterministic SD parameters $\boldsymbol{r}^s$ in standard LHUC to probabilistic for modeling their uncertainty, by given $\boldsymbol{r}^s \sim p(\boldsymbol{r}^s)$ where $p(\boldsymbol{r}^s)$ is a speaker-independent (SI) prior distribution. However, computing the integral in equation (2) for inference on BLHUC adapted DNN model may be slower than using standard LHUC adaptation. An alternative more efficient approximation is to exploit the expectation of $\boldsymbol{r}^s$ following the posterior distribution $p(\boldsymbol{r}^s|\boldsymbol{o}^s, c^s)$ for the inference in equation (2) by

$$\int P(c_t^s|\boldsymbol{o}_t^s, \boldsymbol{r}^s)p(\boldsymbol{r}^s|\boldsymbol{o}^s, c^s)d\boldsymbol{r}^s \approx P(c_t^s|\boldsymbol{o}_t^s, \mathbb{E}[\boldsymbol{r}^s|\boldsymbol{o}^s, c^s]) \quad (4)$$

where $\mathbb{E}[\cdot]$ denotes the expectation. The approximation makes the inference on BLHUC adapted DNN model has similar form and computational complexity to using the standard LHUC adaptation. Figure 2 shows an example of inference on BLHUC adapted DNN model. The essential factor for using BLHUC adaptation is to compute the posterior distribution $p(\boldsymbol{r}^s|\boldsymbol{o}^s, c^s)$ of the scaling vectors $\boldsymbol{r}^s$.

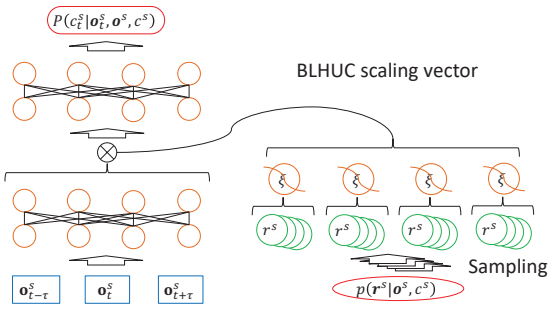

Fig. 2. An example of inference on BLHUC adapted DNN model.

4. VARIATIONAL ESTIMATION FOR BLHUC PARAMETERS

In order to estimate the posterior distribution of the BLHUC scaling vector $\boldsymbol{r}^s$ by variational approximation, the cross entropy (CE) of the adaptation data can be written as

$$\begin{aligned} \text{Loss} &= -\log P(c^s|\boldsymbol{o}^s) \\ &= -\log \int P(c^s|\boldsymbol{o}^s, \boldsymbol{r}^s)p(\boldsymbol{r}^s)d\boldsymbol{r}^s \\ &\leq -\int q_s(\boldsymbol{r}^s)\log P(c^s|\boldsymbol{o}^s, \boldsymbol{r}^s)d\boldsymbol{r}^s + KL(q_s||p) \end{aligned} \quad (5)$$

where $q_s(\boldsymbol{r}^s)$ is the variational approximation of the posterior distribution $p(\boldsymbol{r}^s|\boldsymbol{o}^s, c^s)$, and $KL(q_s||p) = \int q_s(\boldsymbol{r}^s)\log\frac{q_s(\boldsymbol{r}^s)}{p(\boldsymbol{r}^s)}d\boldsymbol{r}^s$ denotes the Kullback-Leibler (KL) divergence between distributions q_s and p. For simplification, both q_s and p are assumed to be normal distributions, namely $q_s(r_d^s) = \mathcal{N}(r_d^s; \mu_{s,d}, \sigma_{s,d}^2)$ and $p(r_d^s) = \mathcal{N}(r_d^s; \mu_{0,d}, \sigma_{0,d}^2)$ on the dth dimension. Then, the expectation in equation (4) can be simply computed as

$$\mathbb{E}[\boldsymbol{r}^s|\boldsymbol{o}^s, c^s] = \boldsymbol{\mu}_s. \quad (6)$$

Moreover, the KL divergence can be explicitly calculated by

$$KL(q_s||p) = \frac{1}{2}\sum_{d=1}^{D}\left\{\frac{(\mu_{s,d}-\mu_{0,d})^2 + \sigma_{s,d}^2}{\sigma_{0,d}^2} - \log\frac{\sigma_{s,d}^2}{\sigma_{0,d}^2} - 1\right\} \quad (7)$$

where D denotes the number of hidden units for adaptation.

BLHUC scaling vector posterior can then be parameterized by

$$\theta_s^{\mathrm{B}} = \{\boldsymbol{\mu}_s, \boldsymbol{\gamma}_s\} \tag{8}$$

where $\boldsymbol{\sigma}_s = \exp(\boldsymbol{\gamma}_s)$. In order to make the θ_s^{B} updatable in the integral in equation (5), the integral can be rewritten as

$$\begin{aligned}
&\int q_s(\boldsymbol{r}^s) \log P(c^s|\boldsymbol{o}^s, \boldsymbol{r}^s) d\boldsymbol{r}^s \\
&= \int \mathcal{N}(\boldsymbol{\epsilon}; 0, I) \log P(c^s|\boldsymbol{o}^s, \boldsymbol{\mu}_s + \exp(\boldsymbol{\gamma}_s) \otimes \boldsymbol{\epsilon}) d\boldsymbol{\epsilon} \\
&\approx \frac{1}{J} \sum_{j=1}^{J} \log P(c^s|\boldsymbol{o}^s, \theta_s^{\mathrm{B}}, \boldsymbol{\epsilon}_j)
\end{aligned} \tag{9}$$

where $\boldsymbol{\epsilon}_j$ is the jth Monte Carlo sample drawed from the standard normal distribution. Then, the gradient of θ_s^{B} in one data batch m can be computed by

$$\begin{aligned}
\frac{\partial \mathrm{Loss}_m}{\partial \theta_s^{\mathrm{B}}} \approx & -\frac{N_s}{J N_{m,s}} \sum_{j=1}^{J} \frac{\partial \log P(c_m^s|\boldsymbol{o}_m^s, \theta_s^{\mathrm{B}}, \boldsymbol{\epsilon}_j)}{\partial \theta_s^{\mathrm{B}}} + \frac{\partial KL(q_s||p)}{\partial \theta_s^{\mathrm{B}}} \\
= & \alpha \left\{ -\frac{1}{J} \sum_{j=1}^{J} \frac{\partial \log P(c_m^s|\boldsymbol{o}_m^s, \theta_s^{\mathrm{B}}, \boldsymbol{\epsilon}_j)}{\partial \theta_s^{\mathrm{B}}} + \frac{N_{m,s}}{N_s} \frac{\partial KL(q_s||p)}{\partial \theta_s^{\mathrm{B}}} \right\}
\end{aligned} \tag{10}$$

where $N_{\{\}}$ denotes the number of frames, and $\alpha = \frac{N_s}{N_{m,s}}$ which can be absorbed by the learning rate. To be specific, the gradients of loss function over $\boldsymbol{\mu}_s$ and $\boldsymbol{\gamma}_s$ on the dth dimension can be computed as

$$\begin{aligned}
\frac{\partial \mathrm{Loss}_m}{\partial \gamma_{s,d}} &= \alpha \left\{ \frac{1}{J} \sum\nolimits_{j=1}^{J} G_{s,j,d}^{\mathrm{LHUC}} \sigma_{s,d} \epsilon_{j,d} + \frac{N_{m,s}}{N_s} \left(\frac{\sigma_{s,d}^2}{\sigma_{0,d}^2} - 1 \right) \right\} \\
\frac{\partial \mathrm{Loss}_m}{\partial \mu_{s,d}} &= \alpha \left\{ \frac{1}{J} \sum\nolimits_{j=1}^{J} G_{s,j,d}^{\mathrm{LHUC}} + \frac{N_{m,s}}{N_s} \frac{\mu_{s,d} - \mu_{0,d}}{\sigma_{0,d}^2} \right\}
\end{aligned} \tag{11}$$

where $G_{s,j,d}^{\mathrm{LHUC}} = -\frac{\partial \log P(c_m^s|\boldsymbol{o}_m^s, \boldsymbol{r}_j^s)}{\partial h_d^s} \frac{\partial \xi}{\partial r_{j,d}^s} \psi_d$ is the gradient in standard LHUC. Therefore, the update of θ_s^{B} over one batch is a trade off between the ML learning of DNN and the penalty for distance to the prior. If the data amount N_s is small, q_s is more likely to approach the SI prior. Thus the variational Bayesian estimation in BLHUC should be more robust than the ML estimation in standard LHUC. Furthermore, an efficient update by setting $J = 1$ can be used instead of computing the summation [24]. Then, the gradient equation (10) has a similar form to that for conventional DNN adaptation using KL-divergence regularization [25]. However, the main difference is, for DNN adaptation using BLHUC, the variable $\boldsymbol{\epsilon}_j$ is randomly selected by normal distribution and thus various for each time of update. The variances on difference entries of BLHUC scaling vector can be tied together such that it has similar parameter number to the standard LHUC.

An important issue of BLHUC adaptation is selection of the prior distribution. As the speaker-independent (SI) DNN can be viewed as using BLHUC with $\boldsymbol{r}^s = 0$, zero mean and unit variance can be used by the prior distribution in BLHUC test time SI DNN model adaptation. When applying BLHUC test time adaptation on BLHUC speaker adaptive training (SAT) [23] DNN model, the BLHUC prior can be directly estimated by computing gradients over $\mu_{0,d}$ and $\sigma_{0,d}$ derived from equations (5) and (7). Alternatively, BLHUC test time adaptation can be applied to the LHUC SAT model. Then, the BLHUC prior will be estimated by replacing the SD LHUC parameters $\boldsymbol{r}^s$ with SI LHUC parameters $\boldsymbol{r}^0$ and globally learned with all training data, similar to the prior estimation in [26].

5. EXPERIMENTS

5.1. Experimental setup

The proposed BLHUC method is investigated for both unsupervised test time speaker-independent (SI) DNN model adaptation, and DNN model speaker adaptive training (SAT). DNN-HMM systems with 8929 tied tri-phone states were built on a 300-hour Switchboard setup. A four-gram language model with 30-thousand words was employed for evaluation on the Hub5' 00 data set with **SWBD** and **CallHome** test sets. DNNs with 6 hidden layers were trained under the minimum cross entropy (CE) criterion and minimum phone error (MPE) criterion. Each of the hidden layer contained 2000 hidden nodes. 9 successive frames of 80 dimensional filter-bank features with the first order difference were concatenated and used as the DNN input features. Back propagation with stochastic gradient descent was employed to update each mini-batches with 800 frames. In the previous research [27], under a similar configuration the performance of speaker-independent DNN system trained with CE on the **SWBD** test set was 15.5% in word error rate (WER). The baseline SI DNN system here got a similar 15.3% of WER. All systems were trained and evaluated with a modification of Kaldi [28] and HTK [29].

When using LHUC or BLHUC adaptation, the SD parameters were employed on the first hidden layer only. Alignments decoded by the baseline SI DNN systems were used for estimating the SD parameters.

5.2. Performance of BLHUC adaptation

Performance of different DNN-HMM systems using or without using test time speaker adaptation were evaluated on the **SWBD** and **CallHome** test sets and shown in table 1. Systems (1)-(3) were trained by minimizing the cross entropy, and systems (4)-(6) were trained based on MPE. Systems (1.1) and (2.1) were baselines in the previous research [16] with similar setup to systems (1) and (2). The standard LHUC adapted DNN systems (Sys (2) and (5)) significantly outperformed the corresponding baseline SI DNN systems (Sys (1) and (4)) by about 4.5% relative WER reductions on **SWBD** and 6%-10% relative WER reductions on **CallHome**, respectively. Moreover, consistent improvements were obtained by using BLHUC adaptation (Sys (3) and (6)) against the standard LHUC adaptation (Sys (2) and (5)) on both CE and MPE trained DNN systems, while similar numbers of SD parameters were exploited.

Sys	DNN criterion	Test adapt	#SD params	WER (%) SWBD	WER (%) CallHome
(1)	CE	-	-	15.3	27.6
(1.1)		-	-	15.2 [16]	28.2 [16]
(2)		LHUC	2k	14.6	25.8
(2.1)		LHUC	-	14.7 [16]	26.6 [16]
(3)		BLHUC	2k	**14.2**	**25.3**
(4)	MPE	-	-	13.4	26.8
(5)		LHUC	2k	12.8	24.0
(6)		BLHUC	2k	**12.4**	**23.1**

Table 1. Performance of baseline SI, LHUC adapted, and BLHUC adapted DNN systems evaluated on **SWBD** and **CallHome** test sets.

Table 2 shows the performance of LHUC adapted and BLHUC adapted DNN systems when the beginning 100%, 50%, 25%, 10% and 1 utterances from the **SWBD** and **CallHome** test sets were utilized as adaptation data for each speaker. Comparing system (3) to system (2), and system (6) to system (5), BLHUC adaptation con-

sistently outperformed the standard LHUC adaptation, when various amounts of adaptation data as low as only 1 utterance were used. On the more difficult **CallHome** test set, significant improvements could be obtained by the BLHUC adapted DNN system (Sys (6)) over the standard LHUC adapted DNN system (Sys (5)) based on MPE training. When only 1 utterance (2 seconds on average) was utilized as adaptation data from each speaker, BLHUC adaptation significantly outperformed the standard LHUC adaptation by absolute WER reduction of 1.1%. Figure 3 and 4 show the performance contrast of LHUC adaptation and BLHUC adaptation on MPE trained DNN systems using various amount of adaptation data.

Test set	Sys	DNN criterion	Test adapt	WER (%) w.r.t. adapt data amount				
				1 utt† 4s§	10%‡ 22s	25% 51s	50% 98s	100% 185s
SWBD	(1)	CE	-	15.3	15.3	15.3	15.3	15.3
	(2)		LHUC	15.2	15.0	14.8	14.8	14.6
	(3)		BLHUC	**14.9**	**14.6**	**14.4**	**14.3**	**14.2**
	(4)	MPE	-	13.4	13.4	13.4	13.4	13.4
	(5)		LHUC	13.3	13.1	13.2	13.0	12.8
	(6)		BLHUC	13.3	13.1	**12.9**	**12.7**	**12.4**
Test set	Sys	DNN criterion	Test adapt	WER (%) w.r.t. adapt data amount				
				1 utt 2s	10% 15s	25% 33s	50% 70s	100% 140s
Call-Home	(1)	CE	-	27.6	27.6	27.6	27.6	27.6
	(2)		LHUC	27.4	27.1	26.6	26.0	25.8
	(3)		BLHUC	**27.2**	**26.5**	**26.0**	**25.7**	**25.3**
	(4)	MPE	-	26.8	26.8	26.8	26.8	26.8
	(5)		LHUC	26.5	25.2	24.6	24.5	24.0
	(6)		BLHUC	**25.4**	**24.1**	**23.6**	**23.4**	**23.1**

Table 2. Performance of LHUC and BLHUC test time adaptation using various amounts of adaptation data.
†: "1 utt" means that 1 utterance from each speaker in the test set was used as adaptation data; ‡: "10%" means that 10% utterances from each speaker in the test set was used as adaptation data; §: "4s" means the average time of adaptation data used for each speaker is 4 seconds.

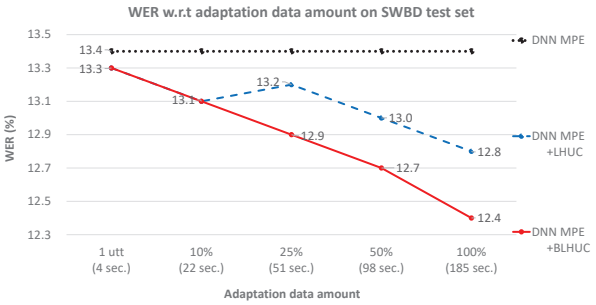

Fig. 3. Performance contrast of LHUC adapted and BLHUC adapted DNN MPE systems using various amounts of adaptation data on **SWBD** test set.

5.3. Performance of BLHUC SAT systems

Performance of using LHUC and BLHUC test time speaker adaptation on different SAT DNN-HMM systems were evaluated on the **SWBD** and **CallHome** test sets and shown in table 3. Here alignments for test time adaptations were decoded by the baseline SI DNN system (Sys (1)). The standard LHUC SAT DNN system (Sys (2)) significantly outperformed the baseline SI DNN system (Sys (1)) by 13.7% and 14.9% relative WER reductions on the **SWBD** and **CallHome** test sets respectively. Consistent improvements could be

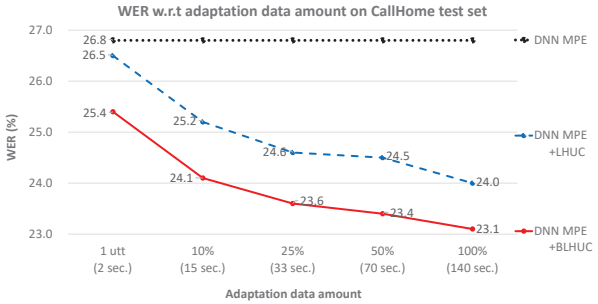

Fig. 4. Performance contrast of LHUC adapted and BLHUC adapted DNN MPE systems using various amounts of adaptation data on **CallHome** test set.

obtained by using BLHUC test time adaptation on both the LHUC SAT (Sys (3)) and BLHUC SAT (Sys (4)) DNN systems, compared to the standard LHUC adaptation. Finally, the BLHUC SAT DNN system (Sys (4)) achieved the best performance among these SAT DNN systems by using the whole test set as adaptation data. However, reducing the amount of adaptation data in BLHUC SAT DNN system led to no improvement over the LHUC SAT DNN system. This might be caused by the inappropriate settings for the BLHUC SAT DNN system. Further investigations will focus on improving the BLHUC SAT system in the future.

Sys	DNN criterion	SAT	Test adapt	WER (%)	
				SWBD	CallHome
(1)	CE	-	-	15.3	27.6
(2)		LHUC	LHUC	13.2	23.5
(3)		LHUC	BLHUC	13.0	23.4
(4)		BLHUC	BLHUC	**12.8**	**22.9**

Table 3. Performances of SAT DNN systems using LHUC and B-LHUC adaptation.

6. CONCLUSION

This paper proposed the Bayesian learning of hidden unit contributions (BLHUC) technique for DNN speaker adaptation. A posterior distribution over the LHUC scaling vectors is used to explicitly model the uncertainty associated with the adaptation parameters. An efficient variational inference based approach is adopted to estimate the LHUC parameter posterior distribution. Experiments conducted on a 300-hour Switchboard setup showed that, when only using 1 utterance from each speaker as adaptation data, the proposed BLHUC adaptation outperformed the baseline speaker-independent DNN systems and LHUC adapted DNN systems by up to 1.4% and 1.1% in absolute WER reduction respectively. When increasing the adaptation data amount, consistent performance improvements could be obtained by BLHUC adaptation systems over the baseline SI, LHUC adapted and LHUC SAT systems. To the best of our knowledge, this is the first work about using Bayesian learning for DNN speaker adaptation. Future works will focus on improving the BLHUC SAT system and other forms of adaptation technique using Baysian learning.

7. REFERENCES

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

[2] A. Senior and I. Lopez-Moreno, "Improving dnn speaker independence with i-vector inputs," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 225–229.

[3] V. Gupta, P. Kenny, P. Ouellet, and T. Stafylakis, "I-vector-based speaker adaptation of deep neural networks for french broadcast audio transcription," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6334–6338.

[4] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7942–7946.

[5] S. Xue, O. Abdel-Hamid, H. Jiang, and L. Dai, "Direct adaptation of hybrid dnn/hmm model for fast speaker adaptation in lvcsr based on speaker code," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6339–6343.

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

[7] H. Huang and K. C. Sim, "An investigation of augmenting speaker representations to improve speaker normalisation for dnn-based speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4610–4613.

[8] M. J. Gales, "Cluster adaptive training of hidden markov models," *IEEE transactions on speech and audio processing*, vol. 8, no. 4, pp. 417–428, 2000.

[9] C. Wu and M. J. Gales, "Multi-basis adaptive neural network for rapid adaptation in speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4315–4319.

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

[11] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[12] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, "Speaker adaptation using constrained estimation of gaussian mixtures," *IEEE Transactions on speech and Audio Processing*, vol. 3, no. 5, pp. 357–366, 1995.

[13] M. J. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[14] B. Li and K. C. Sim, "Comparison of discriminative input and output transformations for speaker adaptation in the hybrid nn/hmm systems," 2010.

[15] J. Trmal, J. Zelinka, and L. Müller, "Adaptation of a feedforward artificial neural network using a linear transform," in *International Conference on Text, Speech and Dialogue*. Springer, 2010, pp. 423–430.

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

[17] P. Swietojanski and S. Renals, "Differentiable pooling for unsupervised speaker adaptation," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4305–4309.

[18] C. Zhang and P. C. Woodland, "Dnn speaker adaptation using parameterised sigmoid and relu hidden activation functions," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5300–5304.

[19] D. J. C. Mackay, "A practical bayesian framework for backpropagation networks," *Neural Computation*, vol. 4, no. 3, pp. 448–472, 1992.

[20] D. Barber and C. M. Bishop, "Ensemble learning in bayesian neural networks," *NATO ASI SERIES F COMPUTER AND SYSTEMS SCIENCES*, vol. 168, pp. 215–238, 1998.

[21] A. Graves, "Practical variational inference for neural networks," in *Advances in neural information processing systems*, 2011, pp. 2348–2356.

[22] J. T. Chien and Y. C. Ku, "Bayesian recurrent neural network for language modeling," *IEEE Transactions on Neural Networks & Learning Systems*, vol. 27, no. 2, pp. 361–374, 2016.

[23] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 2. IEEE, 1996, pp. 1137–1140.

[24] D. P. Kingma and M. Welling, "Stochastic gradient vb and the variational auto-encoder," *Journal of Beijing Administrative College*, 2013.

[25] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7893–7897.

[26] C. Chesta, O. Siohan, and C.-H. Lee, "Maximum a posteriori linear regression for hidden markov model adaptation," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[27] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[28] 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 *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[29] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The htk book," *Cambridge university engineering department*, vol. 3, p. 175, 2002.