Bayesian Learning for Deep Neural Network Adaptation

Xurong Xie, Xunying Liu, Member, IEEE, Tan Lee, Member, IEEE, Lan Wang, Member, IEEE

Abstract—(*This manuscript is an updated version of the article published in TASLP, with appendix of released codes (in https://github.com/XIEXurong/kaldi.git) based on Kaldi toolkit, and appendix of updated experiment results on the Switchboard, HKUST, and DementiaBank Pitt corpuses. Please still cite the published article in the TASLP, DOI: 10.1109/TASLP.2021.3084072.)

A key task for speech recognition systems is to reduce the mismatch between training and evaluation data that is often attributable to speaker differences. Speaker adaptation techniques play a vital role to reduce the mismatch. Model-based speaker adaptation approaches often require sufficient amounts of target speaker data to ensure robustness. When the amount of speaker level data is limited, speaker adaptation is prone to overfitting and poor generalization. To address the issue, this paper proposes a full Bayesian learning based DNN adaptation framework to model speaker-dependent (SD) parameter uncertainty given limited speaker specific adaptation data. This framework is investigated in three forms of model based DNN adaptation techniques: Bayesian learning of hidden unit contributions (BLHUC), Bayesian parameterized activation functions (BPAct), and Bayesian hidden unit bias vectors (BHUB). In the three methods, deterministic SD parameters are replaced by latent variable posterior distributions for each speaker, whose parameters are efficiently estimated using a variational inference based approach. Experiments conducted on 300-hour speed perturbed Switchboard corpus trained LF-MMI TDNN/CNN-TDNN systems suggest the proposed Bayesian adaptation approaches consistently outperform the deterministic adaptation on the NIST Hub5’00 and RT03 evaluation sets. When using only the first five utterances from each speaker as adaptation data, significant word error rate reductions up to 1.4% absolute (7.2% relative) were obtained on the CallHome subset. The efficacy of the proposed Bayesian adaptation techniques is further demonstrated in a comparison against the state-of-the-art performance obtained on the same task using the most recent systems reported in the literature.

Index Terms—Bayesian learning, adaptation, LHUC, TDNN, Switchboard

I. INTRODUCTION

In recent decades the performance of automatic speech recognition (ASR) systems have been significantly improved with the successful application of deep learning techniques. In the conventional deep neural network-hidden Markov models (DNN-HMMs) based hybrid ASR systems [1], [2], artificial neural networks are used to estimate the conditional probabilities of HMM states given acoustic features. In these systems, word sequences are transformed into phoneme sequences using a pronunciation lexicon. The resulting phoneme sequences are modeled by HMMs representing phonetic units. In order to sufficiently model long range temporal contexts and spectral characteristics of continuous speech, advanced forms of DNNs including convolutional neural networks (CNNs) [3], [4], time-delay neural networks (TDNNs) [5], recurrent neural networks (RNNs) [6] and their long short-term memory variants [7] as well as self-attention [8] are widely used in current hybrid DNN-HMM systems. In order to reduce the mismatch between the conventional frame level cross-entropy (CE) based loss function and recognition error rate normally measured at word level over a sentence, sequence level training criteria based on maximum mutual information (MMI) [9], minimum phone error (MPE) [10], segmental minimum Bayes-risk (sMBR) [11], or lattice-free MMI (LF-MMI) [12] can be used to further improve the performance of hybrid ASR systems.

More recently there has also been a significant trend in the speech technology field moving from hybrid HMM-DNN system architecture to an all neural end-to-end (E2E) modeling paradigm. E2E systems represented by connectionist temporal classification (CTC) [13], RNN transducers (RNN-T) [14], listen, attend and spell (LAS) [15] or encoder-decoder with attention (AED) [16], LF-MMI [17] and transformers [18], utilize a single DNN model transforming an input acoustic feature sequence directly to a target stream of output tokens based on characters, word pieces or words.

A key task facing all ASR systems is to learn the systematic and latent variation among diverse speech data. This often creates a large mismatch between the training and evaluation data leading to recognition performance degradation. A major source of such variability is attributable to speaker level characteristics representing factors such as accent and idiosyncrasy, or physiological differences manifested in, for example, age or gender. To this end, speaker adaptation techniques play a vital role in current speech recognition systems. Speaker adaptation techniques used in current neural network based ASR systems can be characterized into several broad categories: auxiliary speaker embedding based approaches, feature transformation based methods, and model-based adaptation techniques. The model-based adaptation approaches exploit specially designed speaker-dependent (SD) DNN model parameters to compensate speaker level variability, for examples, the learning hidden unit contributions (LHUC) [19], [20], [21] and parameterized activation functions (PAct) [22], [23], [24] techniques. More speaker adaptation approaches are introduced in Section I-A.

Corresponding authors: Xunying Liu and Lan Wang.

Xurong Xie is with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, and he was with Chinese University of Hong Kong, Hong Kong, China when he worked on this work, e-mail: xrxie@siat.ac.cn

Xunying Liu, and Tan Lee are with Chinese University of Hong Kong, Hong Kong, China, e-mail: xyliu@se.cuhk.edu.hk, tanlee@ee.cuhk.edu.hk.

Lan Wang is with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, e-mail: lan.wang@siat.ac.cn.
Model-based speaker adaptation approaches often require sufficient amounts of target speaker data for robust SD parameter estimation. However, the available amount of speaker-specific data is often limited in practical applications. Even when using compact SD parameter representation based on the aforementioned structured transforms, for example, an LHUC scaling vector of several thousand dimensions, speaker adaptation can still be prone to overfitting and poor generalization performance. This issue is particularly prominent when rapid adaptation to individual speakers’ voice characteristics using very little data, for example, a few seconds of speech, is required at the onset of user enrollment to ASR systems.

A solution proposed in this paper to address the inherent SD parameter uncertainty resulted from limited adaptation data is using Bayesian learning approaches. In the machine learning community, Bayesian learning has been established as a mathematically well formulated framework to account for modelling uncertainty in neural network systems including current deep learning models. A practical Bayesian framework to account for parameter uncertainty using Laplace approximation in back-propagation neural networks was first introduced in [24]. More efficient inference methods based on variational learning and Monte Carlo parameter sampling were proposed for Bayesian neural networks in [25], before being further extended to auto-encoder neural networks in [26]. In contrast, very limited previous research on Bayesian deep learning approaches has been conducted in the area of speech recognition. In [27], variational inference was adopted in the training of a Bayesian recurrent neural network (RNN) based phone recognition system using the TIMIT speech corpus. Bayesian learning for RNN language modeling was investigated in [28]. More recently, a Bayesian learning framework was used to account for model uncertainty in sequence discriminative training of factored TDNN acoustic models [29], [30]. Stochastic noise injection to model parameters [31] was also exploited to improve the generalization performance of E2E ASR systems [32], [33]. The above application of Bayesian learning to current ASR systems focused on modelling parameter uncertainty in speaker independent systems, while the often more prominent data sparsity problem encountered during speaker adaptation on very limited data remains unvisited.

In order to address this issue, a full Bayesian learning based DNN speaker adaptation framework is proposed in this paper to account for SD parameter uncertainty given limited speaker specific adaptation data. The generic nature and advantages of this full Bayesian DNN adaptation framework is demonstrated via its application to three exemplar types of structured DNN speaker adaptation transforms: Bayesian learning of hidden unit contributions (BLHUC) [34], Bayesian parameterized activation functions (BPAct) and Bayesian hidden unit bias vectors (BHUB). In these Bayesian adaptation methods, deterministic SD parameters are replaced by latent variable posterior distributions to be learned for each speaker, thus allowing the parameter uncertainty to be explicitly modeled to reduce the risk of overfitting. The parameters of these posterior distributions are efficiently estimated using variational inference based approach further benefited from the use of Monte Carlo parameter sampling and re-parameterization. The proposed Bayesian DNN adaptation methods are investigated for both unsupervised test time speaker adaptation and speaker adaptive training.

Experiments conducted on state-of-the-art 300-hour speed perturbed Switchboard corpus trained LF-MMI factored TDNN and CNN-TDNN systems featuring i-vector speaker adaptation suggest the proposed Bayesian adaptation approaches consistently outperform the adapted system constructed using deterministic parameter based LHUC, parameterized activation functions (PAct) or hidden unit bias vectors (HUB) on the NIST Hub5’00 and RT03 evaluation sets in unsupervised test time speaker adaptation and speaker adaptive training. In particular, when using only the first five utterance from each speaker as the adaptation data, statistically significant absolute (relative) reduction up to 1.4% (7.2%) in word error rate (WER) were obtained on the CallHome subset. Similar consistent performance improvements were also retained after LSTM recurrent neural network language model rescoring. The efficacy of the proposed Bayesian adaptation techniques is further demonstrated in a comparison against the state-of-the-art performance obtained on the same task using most recent hybrid and E2E systems reported in the literature.

The main contributions of this paper are summarized below:

1) To the best of our knowledge, this paper presents the first use of a full Bayesian framework to account for model uncertainty in speaker adaptation of DNN acoustic model for speech recognition. In contrast, previous research on Bayesian neural network techniques were restricted to the estimation of un-adapted acoustic models or language models in hybrid systems [27], [28], [29], [30] or E2E systems [32], [33].

2) The efficient variational inference approach developed for the proposed Bayesian DNN adaptation uses a very few number of parameter samples (as low as one) to ensure their computational cost comparable to that of the baseline adaptation techniques. Their generic nature allows them to be applied to E2E approaches to address similar model uncertainty issues during adaptation to new speakers using limited data.

3) Statistically significant performance improvements over multiple model-based DNN adaptation methods given varying amounts of adaptation data were obtained on state-of-the-art baseline LF-MMI factored TDNN systems using i-vector feature based speaker adaptation [35], [36]. These demonstrate the power of the proposed Bayesian adaptation approaches to facilitate both rapid, instantaneous adaptation to individual speakers’ voice, and batch mode adaptation performed using more data. Such flexibility is crucial for speech based conversational agents in commercial applications to ensure user satisfaction from their enrollment to the system very early on.

The rest of this paper is organized as follows. Section I-A introduces the related works of speaker adaptation. Section II reviews three types of structured DNN adaptation based on LHUC, hidden unit bias vectors and parameterized activation functions. A full Bayesian learning framework and its application to these three types of adaptation methods is proposed in Section III. Section IV presents the experimental results and analysis of the Bayesian adaptation approaches conducted on state-of-the-art sequence discriminatively trained TDNN and CNN-TDNN acoustic models on a 300-hour Switchboard.
setup. The last section draws the conclusion and discusses possible future works.

A. Related Works of Speaker Adaptation

In the auxiliary speaker embedding based approaches, auxiliary speaker embedding based DNN adaptation techniques encode speaker-dependent (SD) characteristics in a compact vector representation. The resulting SD embedding vectors are used as auxiliary input features to be fed into DNNs and facilitate speaker adaptation during both system training and evaluation. The SD auxiliary feature estimation is performed independently of the remaining recognition system components. For example, i-vectors [35, 36] are learned from Gaussian mixture model (GMM) based universal background models (UBMs).

In the feature transformation based methods, feature transforms can be applied to acoustic front-ends and aim to produce canonical, speaker invariant input features. These are then fed into the back-end DNN based recognition systems to model other remaining sources of variability, for example, longer span phonetic and phonological context dependency in speech. Feature-space maximum likelihood linear regression (f-MLLR) transforms [37] are estimated at speaker level from GMM-HMM based ASR systems first [38, 39] before being applied to acoustic front-ends to produce canonical, speaker invariant like input features that are then fed into DNN based recognition systems. In order to compensate for vocal tract length differences between speakers, physiologically motivated feature space adaptive transformations based on vocal tract length normalization (VTLN) can also be used [40, 41]. Speaker level VTLN normalized features can be obtained using either piecewise linear frequency warping factors directly applied to the spectrum, or affine linear transformations similar to f-MLLR.

In the model-based adaptation techniques, the amount of speaker specific adaptation data practically determines the SD parameters’ modelling granularity. In order to ensure good generalization performance and mitigate the risk of overfitting, a trade-off between moving from adapting only parts of the whole acoustic model towards complete speaker dependent systems [42, 43] needs to be adjusted. To this end, existing researches focused on deriving compact forms of SD parameter representations, predominantly based on structured transforms that are positioned in various parts of the DNN acoustic model. These include the use of SD linear input networks (LIN) [44, 45], linear output networks (LON) [45], linear hidden networks (LHN) [46], learning hidden unit contributions (LHUC) [19, 20, 21], parameterized activation functions (PAct) [23, 21], subspace factorized speaker-independent (SI) and SD affine transformations [47], speaker code based bias vectors [48], and adaptive linear interolation of basis sub-networks’ hidden outputs [49, 50], akin to cluster based speaker adaptation for GMM-HMM systems [51]. In addition to only modelling the variability among speakers found in the test data during recognition, the estimation of SD parameters in both the system training and evaluation stages leads to speaker adaptive training (SAT) [52], thus allowing a joint optimization of both the SD and SI parameters shared among all speakers in the training data.

Other forms of model-based approaches learn to compensate speaker level variability without explicitly using SD parameters. For example, adversarial learning [53, 54] of DNN acoustic models can be used to generate speaker-invariant hidden layer outputs. Although most of the above adaptation approaches were first developed for hybrid ASR systems, recent researches suggest these can be further applied to E2E based ASR systems [55, 56, 57].

A series of speaker adaptation techniques have been developed in the past to address the data sparsity issue. Among these, data augmentation methods were used to expand the target speaker or domain data using, for example, perturbation of speaking rates [58, 59] or vocal tract length [60], speaker mappings [61] and synthesis techniques [62]. Another category of solutions to this problem is based on regularized adaptation methods. For example, L2 regularization was used to prevent the SD parameters from diverging too far from those of the SI models [22]. This is similar to maximum a posteriori (MAP) adaptation [63, 64, 65, 66] which restricts the estimation of SD parameters using a prior distribution empirically estimated, for example, over a set of SD adapted parameters obtained from the training data. Furthermore, Kullback-Leibler (KL) divergence between the output distributions of SI model and adapted model was also exploited for regularization of DNN adaptation [43]. Hyper-parameter grid-search and meta-learning based adaptation methods regularized by early stopping and appropriate learning rate settings have also been investigated [67].

II. MODEL-BASED SPEAKER ADAPTATION

In this section, we introduce several forms of widely used DNN model-based speaker adaptation approaches. A common idea of applying model-based speaker adaptation to a DNN model is to modify a set of SD parameters for each speaker. A widely used form of SD parameters is an affine transformation deployed over hidden vectors in a DNN layer. However, the use of full matrix affine linear transformation with thousands of parameters makes the adaptation prone to overfitting. An alternative idea is to use less SD parameters for adaptation, for example, in form of vectors with significantly lower dimension than affine transformations.

A. Learning Hidden Unit Contributions

An intuitive solution is to reduce the number of SD parameters by restricting the affine transformation as a diagonal matrix. The SD parameters can then be realized as an SD scaling vector to the DNN hidden layer vector. This approach is known as the learning hidden unit contributions (LHUC) adaptation [20]. In a certain DNN layer and for a certain speaker s, let \( h_t \in \mathbb{R}^D \) be a hidden layer vector at the tth instant, and \( r^s \) be the SD parameters for speaker s, the adapted hidden vector is computed as

\[
h_t^s = \xi(r^s) \odot h_t
\]
where $\xi(r^s)$ is the scaling vector parameterized by $r^s \in \mathbb{R}^{D_r}$, and $\otimes$ denotes the Hadamard product operation. In practice, $h_t$ can be the hidden vector computed after or prior to a nonlinear activation function. In this study, $h_t$ is a ReLU activated hidden vector. The LHUC adaptation is shown as the left part in figure 1.

The dimension $D_r$ is commonly not larger than the layer dimension $D$. The LHUC activation function $\xi: \mathbb{R}^{D_r} \rightarrow \mathbb{R}^D$ constrains the ranges of the LHUC scaling. If $D_r = D$, a widely used form of LHUC activation is 2Sigmoid such that the scaling range is $(0, 2)$ [19]. Other functions can be ReLU [21], exponential [20], or identity functions. If $D_r$ is less than $D$, $\xi$ can be a linear projection matrix mapping lower-dimensional SD parameters to high-dimensional scaling vectors [68].

### B. Hidden Unit Bias Vector

Another way to reduce the number of SD parameters in affine transformation is to use the bias vector only. The SD parameters serve as a bias or offset to the hidden vector in a certain DNN layer. This is referred to hidden unit bias vectors (HUB) adaptation in this study. The adapted hidden vector is then computed as

$$h^*_t = h_t + \xi(r^s)$$

(2)

where $\xi(r^s)$ is the bias vector parameterized by $r^s \in \mathbb{R}^{D_r}$. The speaker code based adaptation [48] is one of the approaches using only the bias vector as SD parameters. The HUB adaptation is shown as the middle part in figure 1.

Similar to LHUC adaptation, the bias activation function $\xi$ can be function of Tanh, identity, or a linear projection matrix mapping lower-dimensional SD parameters to high-dimensional bias vectors [48].

### C. Parameterized Activation Functions

The third type of structured DNN transforms based adaptation considered in this paper uses parameterised activation functions (PAct) [23]. When using parameterised ReLU (PReLU) as activation function, and defining the SD piecewise linear slope parameters as $r^s = [\alpha^s, \beta^s]^T \in \mathbb{R}^{2D}$, the $d$th element of speaker-specific activation vector after adaptation is computed as

$$h^*_{t,d} = \begin{cases} 
\alpha^s z_{t,d}, & z_{t,d} > 0 \\
\beta^s z_{t,d}, & z_{t,d} \leq 0 
\end{cases}$$

(3)

where $\alpha^s_{d}$ and $\beta^s_{d}$ are the $d$th elements of $\alpha^s$ and $\beta^s$ respectively, and $z_{t,d}$ is the hidden node value prior to the PReLU activation. The PAct adaptation is shown as the right part in figure 1. The standard ReLU activation can be viewed as a PReLU activation with $\alpha^s = 1$ and $\beta^s = 0$.

### D. Estimation with Minimum Cross Entropy

The LHUC, HUB, and PAct adaptation approaches share the same framework for SD parameter estimation. The SD parameters can be estimated by minimizing cross entropy (CE). The minimum CE estimator can be given by

$$\hat{r}^s_{CE} = \arg\min_r \{-\log P(C^s|O^s, r^s)\}$$

(4)

where $O^s$ denotes the adaptation data set for speaker $s$, and $C^s$ is the corresponding supervision labels, e.g., HMM state alignments. Without loss of generality, the SI parameters in the model is omitted.

In the context of unsupervised test-time adaptation, the adaptation data $O^s$ is commonly a part of test data, hence the supervision $C^s$ is not available during adaptation. In order to handle the problem, $C^s$ can be derived by applying decoding process to the adaptation data $O^s$. After SD parameters estimated, if necessary, the adaptation data can be decoded again to re-generate the supervision $C^s$ for re-estimation of the SD parameters. Finally, the estimated SD parameters are employed by decoding process.

In the speaker adaptive training (SAT) framework [52], the SI parameters associated with speakers in training data is estimated in the training stage. This enables the SI parameters to be learned more effectively with less distortion by speaker...
variability. In this stage, the SD parameters are updated with the same loss function as that for DNN model training.

In the decoding process, given test data $\tilde{O}^s$, the corresponding states $\tilde{C}^s$ is inferred by the adapted model with probability
\[
P(\tilde{C}^s|\tilde{O}^s, O^s, C^s) \approx P(\tilde{C}^s|\tilde{O}^s, \hat{r}_{CE}^s) \quad (5)
\]
where $\hat{r}_{CE}^s$ is the minimum CE estimate of $r^s$ in equation (4).

III. BAYESIAN LEARNING OF ADAPTATION APPROACH

This section presents a framework for applying full Bayesian learning to DNN model-based speaker adaptation approaches. This framework is demonstrated via applications to the aforementioned three DNN adaptation approaches: Bayesian learning of hidden unit contributions (BLHUC), Bayesian hidden unit bias vectors (BHUB), and Bayesian parameterized activation functions (BPAct).

A. Inference of Bayesian Learning

In contrast to the conventional fixed parameter based decoding process determining the most likely word or HMM state sequence, the comparable Bayesian inference incorporating the adaptation parameter uncertainty is given as the following
\[
P(\tilde{C}^s|\tilde{O}^s, O^s, C^s) = \int P(\tilde{C}^s|\tilde{O}^s, r^s)p(r^s|O^s, C^s)dr^s. \quad (6)
\]
The inference processes of minimum CE and MAP are special cases of the Bayesian inference. When using the minimum CE estimate for inference as equation (5), an approximation is employed by assuming $P(\hat{r}_{CE}^s|O^s, C^s) \approx 1$. Similarly, the MAP adaptation makes use of $\hat{r}_{MAP} = \arg \max_r \log P(r^s|O^s, C^s)$ by assuming $P(\hat{r}_{MAP}^s|O^s, C^s) \approx 1$. Under these assumptions, the point estimates may introduce estimation bias to the inference process.

In Bayesian learning, the deterministic SD parameters are replaced by probabilistic latent variables. The same notation $r^s$ is used for the latent variable without loss of generality. Assuming that the posterior distribution $p(r^s|O^s, C^s)$ is given, the equation (6) can be directly used for decoding. The Monte Carlo approximation of the integral in equation (6) with $J_{inf}$ samples is computed as
\[
P(\tilde{C}^s|\tilde{O}^s, O^s, C^s) \approx \frac{1}{J_{inf}} \sum_{j=1}^{J_{inf}} P(\tilde{C}^s|\tilde{O}^s, r_j^s) \quad (7)
\]
where $r_j^s$ denotes the $j$th random sampling of $r^s$ drawn from the posterior distribution $p(r^s|O^s, C^s)$.

However, computing the equation (7) over $J_{inf}$ random samples could be computationally more expensive than using minimum CE point estimate. The expectation of $r^s$ is an alternative point estimate to approximate the integral for inference. This is given by
\[
P(\tilde{C}^s|\tilde{O}^s, O^s, C^s) \approx P(\tilde{C}^s|\tilde{O}^s, E[r^s|O^s, C^s]) \quad (8)
\]
where $E[\cdot]$ denotes the expectation. Using the expectation takes a similar complexity and procedure to using the minimum CE estimate by assuming that $P(E[r^s|O^s, C^s]|O^s, C^s) \approx 1$. Moreover, it normally represents an unbiased average over the possible latent variable values. In order to use Bayesian learning for adaptation, the key issue is to compute the posterior distribution $p(r^s|O^s, C^s)$ over the SD latent variable $r^s$.

B. Framework for Variational Approximation

Variational distribution $q_s(r^s)$ can be used to approximate the true posterior distribution $p(r^s|O^s, C^s)$. For the speaker $s$, the marginal cross entropy over the adaptation data set $\{C^s, O^s\}$ can be computed as
\[
-\log P(C^s|O^s) = -\log \int P(C^s, r^s|O^s)dr^s \leq -\int q_s(r^s) \log P(C^s|O^s, r^s)dr^s + KL(q_s||p_0)
\]
\[
\mathcal{L}(q_s) \triangleq KL(q_s||p_0) \quad (9)
\]
where $KL(q_s||p_0) = \int q_s(r^s) \log \frac{q_s(r^s)}{p_0(r^s)} dr^s$ denotes the KL divergence between the variational distribution $q_s(r^s)$ and prior distribution $p_0(r^s)$ of $r^s$. The gap between the cross entropy $-\log P(C^s|O^s)$ and its upper bound $\mathcal{L}(q_s)$ is actually the KL divergence between the true posterior $p(r^s|O^s, C^s)$ and the variational distribution $q_s(r^s)$, which is expressed as
\[
KL(q_s(r^s)||p(r^s|O^s, C^s)) = \int q_s(r^s) \log \frac{q_s(r^s)}{p(r^s|O^s, C^s)} dr^s
\]
\[
\mathcal{L}(q_s) = \{-\log P(C^s|O^s)\}. \quad (10)
\]
Obviously, minimizing the upper bound $\mathcal{L}(q_s)$ is equivalent to minimizing the distance between $p(r^s|O^s, C^s)$ and $q_s(r^s)$. Thereby, the goal is to estimate $q_s$ for minimizing $\mathcal{L}(q_s)$.

For simplicity, both $q_s$ and $p_0$ are assumed to be normal distributions parameterized by $\theta_s^B = \{\mu_s, \sigma_s\}$ and $\theta_s^U = \{\mu_0, \sigma_0\}$ respectively. These distributions can be written as
\[
q_s(r_0^s) = N(r_0^s; \mu_s, \sigma_s^2)
\]
\[
p_0(r_0^s) = N(r_0^s; \mu_0, \sigma_0^2) \quad (11)
\]
where $\mu_s$ and $\mu_0$ denote the $d$th elements of the mean vectors $\mu_s$ and $\mu_0$, and $\sigma_s$ and $\sigma_0$ denote the $d$th elements of the standard deviation vectors $\sigma_s$ and $\sigma_0$, respectively.

By this assumption, the random sample $r_0^s$ for inference in equation (11) is drawn from the distribution $N(\mu_s, \sigma_s^2)$, and the inference in equation (8) is written as
\[
P(\tilde{C}^s|\tilde{O}^s, O^s, C^s) \approx P(\tilde{C}^s|\tilde{O}^s, \mu_s) \quad (12)
\]
where the expectation of posterior distribution is simply approximated by $\mu_s$, which is presented as
\[
E[r^s|O^s, C^s] \approx \mu_s. \quad (13)
\]
Meanwhile, the KL divergence $KL(q_s||p_0)$ has a closed form integral that is explicitly calculated as
\[
KL(q_s||p_0) = \frac{1}{2} \sum_{d=1}^{D} \left( \frac{(\mu_s - \mu_0)^2}{\sigma_s^2} + \frac{(\mu_s - \mu_0)^2}{\sigma_0^2} - \log \frac{\sigma_s^2}{\sigma_0^2} - 1 \right) \quad (14)
\]
However, directly updating $\theta_s^B$ in the integral of upper bound $\mathcal{L}(q_s)$ is nontrivial. The re-parameterization introduced in 26
is an efficient technique to obtain a differentiable estimator of the upper bound. The parameters $\theta^b_s$ can be re-parameterized as
\[ r^s = \mu_s + \sigma_s \otimes \epsilon \] (15)
where $\epsilon \sim \mathcal{N}(0, I)$. Applying Monte Carlo approximation, the integral in equation (9) is rewritten as
\[ \mathcal{L}_1(q_s) = \int q_s(r^s) \log P(C^s|O^s, r^s) dr^s \]
\[ = \int \mathcal{N}(\epsilon; 0, I) \log P(C^s|O^s, \mu_s + \sigma_s \otimes \epsilon) d\epsilon \approx \frac{1}{J_{\text{est}}} \sum_{j=1}^{J_{\text{est}}} \log P(C^s|O^s, \theta^b_s, \epsilon_j) \] (16)
where $\epsilon_j$ is the $j$th random sampling of $\epsilon$ drawn from the standard normal distribution $\mathcal{N}(0, I)$.

To minimize the upper bound $\mathcal{L}(q_s)$ in equations (13) computed by equations (14) and (16), the gradient of $\theta^b_s$ at the $m$th update for speaker $s$ is computed as
\[ \frac{\partial \mathcal{L}^{(m)}(q_s)}{\partial \theta^b_s} \approx -\frac{N_s}{N_{m,s, \text{est}}} \frac{1}{J_{\text{est}}} \sum_{j=1}^{J_{\text{est}}} G_{m,s,j,d} \epsilon_{j,d} + \lambda \frac{\partial K L(q_s || p_0)}{\partial \theta^b_s} \] (17)
where $N_s$ denotes the number of frames for speaker $s$ in all adaptation data, $N_{m,s}$ denotes the number of frames for speaker $s$ used by the $m$th update, $\{C^s_m, O^s_m\} \subseteq \{C^s, O^s\}$ is the adaptation data used in the $m$th update for speaker $s$, and $\lambda \in (0, 1]$ makes the gradient parts computed from the DNN and KL divergence have similar scales. Specifically, the gradients of upper bound $\mathcal{L}^{(m)}(q_s)$ over parameters $\mu_s$ and $\sigma_s$ on the $d$th dimension can be calculated as
\[ \frac{\partial \mathcal{L}^{(m)}(q_s)}{\partial \sigma_{s,d}} = \frac{N_s}{N_{m,s, \text{est}}} \frac{1}{J_{\text{est}}} \sum_{j=1}^{J_{\text{est}}} G_{m,s,j,d} \epsilon_{j,d} + \lambda \frac{1}{\sigma_{0,d}} \frac{\partial \sigma_{0,d}}{\partial \sigma_{s,d}} \] (18)
where $G_{m,s,j,d} = -\frac{\partial \log P(C^s_m|O^s_m, \epsilon_j)}{\partial \epsilon_j}$ is the gradient computed from DNN model.

The gradient $G_{m,s,j,d}$ in equation (18) has the same form as that derived by minimum CE estimation using parameters $r^s_j = \mu_s + \sigma_s \otimes \epsilon_j$, thus has different forms for different adaptation approaches:

1) **Bayesian LHUC adaptation**: For using LHUC adaptation approaches as introduced in Section II-A the Bayesian learning based adaptation approach is referred to Bayesian LHUC (BLHUC) adaptation. The corresponding gradient $G_{m,s,j,d}$ is given by
\[ G_{\text{LHUC}} = -\sum_{t_m}^{D} \frac{\partial \log P(C^s_m|O^s_m, r^s_j)}{\partial r^j} \frac{\partial \xi_d(r^s_j)}{\partial r^j} \sigma_{m,d} (t_m) \] (19)

2) **Bayesian HUB adaptation**: For HUB adaptation as introduced in Section II-B the Bayesian learning approach is referred to Bayesian HUB (BHUB) adaptation. The gradient $G_{m,s,j,d}$ is given by
\[ G_{\text{HUB}} = -\sum_{t_m}^{D} \frac{\partial \log P(C^s_m|O^s_m, r^s)}{\partial r^j} \frac{\partial \xi_d(r^s)}{\partial r^j} \] (20)

The BHUB adaptation is shown as the middle part in figure 2.

3) **Bayesian PAct adaptation**: For PAct adaptation as introduced in Section II-C the gradient $G_{m,s,j,d}$ becomes
\[ G_{\text{PAct}} = -\sum_{t_m}^{D} \frac{\partial \log P(C^s_m|O^s_m, r^s)}{\partial r^j} \frac{\partial \xi_d(r^s)}{\partial r^j} \] (21)

with $r^s = [\alpha^s, \beta^s]^T$. This Bayesian learning approaches is referred to Bayesian PAct (BPAct) adaptation, and is shown as the right part in figure 2.

**C. Implementation Details**

1) **Choices of Activation Functions**: The activation function $\xi$ in Bayesian adaptation affects its performance. Different forms of $\xi$ can be used in BLHUC and BHUB adaptation as mentioned in Sections II-A and II-B. In this paper we assume $D_r = D$ (e.g., the same dimensionality for both the hidden vector and LHUC SD parameters in one layer), and the forms of $\xi$ for BLHUC and BHUB adaptation are given in Table I

| Bayesian adaptation | Latent variable | Activation $\xi$ | Prior $p_0$ |
|---------------------|----------------|-----------------|-------------|
| BLHUC               | $r^s$          | Identity        | $\mathcal{N}(0, 1)$ |
| BHUB                | $r^s$          | Sigmoid         | $\mathcal{N}(0, 0.01)$ |
| BPAct               | $\alpha^s$     | Tanh            | $\mathcal{N}(0, 1)$ |

2) **Choices of Prior Distribution**: One important issue of applying Bayesian adaptation is the choice of prior distribution, which is dependent on the form of activation function mentioned in Section II-C. The prior distribution $p_0$ can be directly given as a certain simple distribution, which is shown as in Table I. Alternatively, an empirical prior distribution can be used similarly to that in MAP adaptation. The empirical prior is estimated using training data by applying minimum CE estimation to the SD parameters. Letting $r^s$ be the SD parameter estimates to training data, the $d$th element of the mean and variance vectors of the prior distribution are computed over all training speakers as
\[ \mu_{0,d} = \frac{1}{S} \sum_{s=1}^{S} r^s_d \]
\[ \sigma_{0,d}^2 = \frac{1}{S} \sum_{s=1}^{S} (r^s_d - \mu_{0,d})^2 \] (23)
where $S$ is the number of speakers in the training data.

3) Test-Time Adaptation and SAT: During test-time adaptation using Bayesian learning, SD latent variable distributions are learned using variational inference and applied to the ReLU activated hidden vectors for different layers. All the other parameters in the baseline SI system shared among all speakers are fixed. Cross entropy is minimized by back-propagation based on canonical gradients in 7 epochs. The HMM state alignments used as supervision labels for adaptation are obtained by decoding using the baseline SI system. In the first $n$ layer applying Bayesian adaptation, the batch level gradient scaling $\lambda$ in equation (18) is set empirically to $\min\{10^{n-5}, 1\}$. In the SAT framework, the adaptation approaches using deterministic SD parameters are applied in the training stage, because data for training speakers should be sufficient such that Bayesian learning is unnecessary.

4) Computational Efficiency in Inference: The variational inference based Bayesian adaptation approaches presented in Section III-A requires Monte Carlo sampling in order to approximate integral in equations (6) and (7) for inference, and to approximate the first term of the marginal log-likelihood shown in equations (9) and (16). The number of samples drawn has a direct impact on the computational efficiency during inference. In order to maintain the overall computational cost of Bayesian adaptation to that of the conventional deterministic parameter adaptation methods, the following approaches can be used. First, as mentioned in Section III-A, taking the parameter mean or expectation during inference as equation (8) is more efficient than drawn samples using equation (9). Second, it is feasible to use only one sample in each update (26) of Bayesian adaptation in equation (18), i.e., $J_{out} = 1$. Finally, in each layer the Gaussian variances $\sigma^2_{s,d}$ of SD latent variable in a posterior distribution can be tied together over each of the dimension $d$. This allows for each speaker, when using any of three adaptation methods (LHUC, HUB or PAct), the effective number of SD parameters of their latent posterior distributions are comparable to that of using conventional deterministic parameter adaptation. All these techniques allow Bayesian adaptation to have comparable computational efficiency to the deterministic adaptation.

IV. EXPERIMENTS

In this experiment section, the performance of the Bayesian learning based speaker adaptation approaches including BLHUC, BHUB, and BPAct proposed in Section III are investigated on the commonly used 300-hour Switchboard conversational telephone speech dataset and its augmented 900-hour version derived using speed perturbation [58]. This section is organized as follows. Section IV-A introduces the Switchboard speech data used in the experiments. A series of experiments conducted on the 300-hour setup are then presented in Section IV-B. The baseline TDNN system trained on the 300-hour data is described in Section IV-B1. An initial set of experiments in Section IV-B2 serve to confirm the implementation details and settings previously given in Section III-C, which are sensible to use for the subsequent experiments in the rest of this paper. The robustness of our Bayesian adaptation methods is then demonstrated in Section IV-B3 by performing adaptation using varying amounts of speaker level data, as little as five utterances of speech. The advantages of the Bayesian adaptation techniques of this paper are further shown in Section IV-B4 through an extensive comparison drawn against existing regularised adaptation methods. In Section IV-C1 the performance of Bayesian adaptation methods is further evaluated on speed perturbed 900-hour setup featuring more advanced baseline system configurations using i-vector adapted CNN-TDNNS and LSTM-RNN language models. A final performance contrast against a set of latest hybrid and end-to-end systems reported in the literature is presented in Section IV-C2 to further demonstrate the superiority of Bayesian adaptation approaches.

A. Data Description

Conversational telephone speech is a widely investigated ASR task in which speaker variability is considered a major challenge. The Switchboard-1 Release 2 (LDC97S62) corpus [70] was used in the training stage. Near 286 hours of speech collected from 4804 speakers were utilized as training
set. Approximate 3.6 minutes of speech data was recorded from each speaker on average. This training set was divided into 191,602 utterances of which each lasted for about 5.4 seconds on average.

For evaluation, the results were reported on the NIST Hub5’00 (LDC2002S09) and RT03 (LDC2007S10) data sets. The Hub5’00 set consisted of two subsets with different domain properties: Switchboard (SWBD) and CallHome (CHE). The SWBD subset with 2 hours of speech was better matched to the training data. The CHE subset with 1.5 hours of speech had larger mismatch thus was a more challenging task. Both of these two subsets were collected from 40 speakers. Approximate 3.1 and 2.5 minutes of speech data were recorded from each speaker in the SWBD and CHE subsets respectively on average. Each utterance lasted for about 4.1 and 2.3 seconds on average. The RT03 data set consisted of two subsets: Fisher (FSH) and SWBD. Both of them contained 72 speakers. Approximate 2.5 and 2.7 minutes of speech data were recorded from each speaker respectively on average. The data amount for each speaker in the evaluation sets was thus significantly less than that in the training set.

B. Experiments on 300-Hour Switchboard Setup

1) TDNN Baseline System: The baseline TDNN system was built mainly following the recipe of the Kaldi toolkit [71]. The network consisted of 15 context-splicing layers with 1536 nodes per layer. A 160-dimensional factored linear projection was employed prior to affine transformation in each context-splicing layer other than the first one. The ReLU function was used in the context-splicing layer, followed by batch normalization and dropout operations. The hidden layer input prior to context splicing was scaled and added to the hidden layer output by a skip connection. The splicing indices \{-1, 0, 1\} and \{-3, 0, 3\} were employed in the first 4 and the last 10 context-splicing layers respectively. After two 256 dimensional linear projections and batch normalization operations, the output layer generated the probabilities of 4456 HMM states.

For TDNN parameter optimization, the LF-MMI criterion [12] was applied together with the minimum CE regularization. The back-propagation update was based on natural gradient [72]. Different from the recipe, 40-dimensional static filter-bank features were transformed by linear discriminative analysis to produce the input features. Speed perturbation was not used in this section. Due to the limited computing resource, the baseline TDNN was trained with 4 epochs of iterations on single job with one TITAN X (Pascal) GPU. No parallel models combination was used in each iteration.

In the decoding process, each utterance was first decoded by a 3-gram language model (LM) with a 30K word vocabulary which was built with the transcripts of training data. Lattice-based rescoring was then applied using a 4-gram LM built with the transcripts of both the Switchboard training data and the Fisher English corpora (LDC2004T19 & LDC2005T19) to obtain the best recognition result.

It has been shown that i-vector and LHUC adaptation could be jointly used to achieve complementary performance improvement in state-of-the-art ASR systems [73]. Hence i-vector based adaptive training was used by the baseline TDNN system to investigate the complementary effect between i-vector and Bayesian adaptation. The i-vector extractor was built by exploiting the first 100,000 utterances of training data.

Significance test based on the matched pairs sentence-segment word error was applied with the NIST SCTK [74]. Two different models given the same evaluation set were utilized for the test, by setting the null hypothesis as no difference of average segmental errors was yielded between the two models [75]. The significance level is defined as $P \leq 0.05$.

2) Analysis of Bayesian Adaptation Configurations: This section investigates the appropriate configurations of Bayesian adaptation approaches including the BLHUC, BHUB, and BPAct adaptation. Table II shows the performance of applying Bayesian adaptation with different configurations to TDNN systems without using i-vector on the Hub5’00 data set. All speaker level data was used as adaptation data in this investigation. For each update, the mini-batch size was 64. The learning rates for LHUC and PAct were set to 0.01, and that for HUB was 0.000001, which were not normalized by the mini-batch size. All Bayesian adapted systems used the same random seed for fair comparison. The systems applying Bayesian adaptation consistently outperformed the corresponding systems applying deterministic adaptation.

| Adapt method | Adapt #layers | Activation (\(\epsilon\)) | Tied var | \(\lambda_{\text{inf}}\) | Method of inference | WER (%) of Hub5’00 |
|--------------|---------------|--------------------------|----------|----------------|-------------------|-------------------|
| (1)          | -             | -                        | -        | -              | -                 | 10.1              |
| (2)          | -             | -                        | -        | -              | -                 | 20.6              |
| (3)          | -             | -                        | -        | -              | -                 | 15.4              |
| (4)          | LHUC          | 1                        | Identity | -              | -                 | 9.7               |
| (5)          | LHUC          | 1                        | 2Sigmoid | Exponential   | 1                 | 19.3              |
| (6)          | LHUC          | 1                        | 3        | 2Sigmoid Exponential | 1 | 19.4              |
| (7)          | LHUC          | 1                        | Identity | -              | -                 | 9.8               |
| (8)          | LHUC          | 1                        | 2Sigmoid | Exponential   | 1                 | 19.2              |
| (9)          | LHUC          | 1                        | Identity | -              | -                 | 9.7               |
| (10)         | LHUC          | 1                        | 2Sigmoid | Exponential   | 1                 | 19.1              |
| (11)         | LHUC          | 14                       | Identity | -              | -                 | 9.9               |
| (12)         | LHUC          | 14                       | 2Sigmoid | Exponential   | 1                 | 19.3              |
| (13)         | LHUC          | 14                       | 3        | 2Sigmoid Exponential | 1 | 19.4              |
| (14)         | LHUC          | 14                       | Identity | -              | -                 | 9.7               |
| (15)         | LHUC          | 14                       | 2Sigmoid | Exponential   | 1                 | 19.1              |
| (16)         | LHUC          | 14                       | Identity | -              | -                 | 9.8               |
| (17)         | LHUC          | 14                       | 2Sigmoid | Exponential   | 1                 | 19.1              |
| (18)         | LHUC          | 14                       | Identity | -              | -                 | 9.8               |
| (19)         | LHUC          | 14                       | 2Sigmoid | Exponential   | 1                 | 19.4              |
| (20)         | LHUC          | 14                       | Identity | -              | -                 | 9.7               |
| (21)         | LHUC          | 14                       | 2Sigmoid | Exponential   | 1                 | 19.1              |
| (22)         | LHUC          | 14                       | Identity | -              | -                 | 9.8               |
| (23)         | PAct          | 1                        | -        | -              | -                 | 9.8               |
| (24)         | BPAct         | 1                        | -        | -              | -                 | 19.6              |
| (25)         | BPAct         | 1                        | -        | -              | -                 | 19.4              |
| (26)         | BPAct         | 1                        | Identity | -              | -                 | 9.5               |

The results shown by Systems (5) to (8.6) in Table II confirmed the combined use of tied variances, single sample in
variational update ($J_{\text{est}} = 1$ in equation (18)) and taking the latent variable distribution expectation (equation (8)) during inference, as suggested previously in Section III-C4 produced the best trade-off between computational efficiency and performance during Bayesian adaptation. When the expectation in equation (8) was not used, a small number of samples in equation (7) was used such that the computational cost would not increase significantly. Comparison between System (7) and Systems (8.1) to (8.6) show that, during inference, using the expectation is a good approximation and no statistically significant performance difference was observed when compared with drawing varying numbers of samples from 3 up to 128. System (5) using $J_{\text{est}} = 1$ (equation (18)), expectation for inference (equation (8)), and tied variances made the adaptation simpler and more efficient while producing performance comparable to Systems (6) to (8.6). Hence this configuration would be used by other Bayesian adapted systems in the rest of this paper.

Two trends could be observed from comparison between LHUC and BLHUC adapted systems (Systems (2) to (16) in Table III). The first trend suggested that applying BLHUC adaptation to the first 14 layers (Systems (14) to (16)) achieved better performance than to the first layer only (Systems (5) to (10)), while no significant difference was obtained by applying LHUC adaptation to varying numbers of layers. Second, using the identity activation as $\xi$ (in the left part of figure 2) achieved the best performance of BLHUC adapted to different numbers of layers (Systems (5) and (14)), while 2Sigmoid was the best one for standard LHUC. Figure 3 shows more details of the trends on the CHE data of $\text{Hub5'00}$ with systems applying LHUC or BLHUC adaptation to varying number of layers and using different activation functions for LHUC scaling vectors. The adapted layers were ranging from the first 1 to the first 14 layers. In this figure the BLHUC adaptation was repeated 10 times with different random seeds (for sampling in equation (19)) to show the average performance in WER and standard deviation. The performance of BLHUC adaptation went better when more layers were adapted, while more adapted layers showed overfitting in LHUC adaptation. This demonstrated the robustness of BLHUC adaptation against parameter uncertainty that linearly increased with respect to the number of deterministic SD parameters in the LHUC adapted systems.

Similarly, applying BHUB adaptation to 14 layers (Systems (22) in Table III) achieved better performance than to the first layer only (Systems (19) and (20)), while no significant difference was obtained by applying HUB adaptation to varying number of layers. Moreover, using identity or Tanh activation functions as $\xi$ (in the middle part of figure 2) obtained similar performance, hence Tanh activations would be used for HUB/BHUB adaptation. In contrast, no significant improvement was obtained by applying PAct adaptation to more layer than only the first layer (Systems (24) v.s. (26)), while significant performance degradation resulted from applying PAct adaptation to 14 layers. This also demonstrated the robustness of PAct adaptation against parameter uncertainty that linearly increased with respect to the number of deterministic SD parameters in the PAct adapted systems.

Fig. 3. Performance contrast of applying LHUC and BLHUC adaptation to varying number of hidden layers and using different activation functions for LHUC scaling vectors, including 2Sigmoid, identity, and exponential. The evaluation was conducted on the CHE subset of $\text{Hub5'00}$. The BLHUC adaptation was repeated 10 times with different random seeds (for sampling in equation (16)) to show the average performance in word error rate (WER) and standard deviation.

3) Robustness of Adaptation Using Varying Amounts of Speaker Data: The robustness of Bayesian adaptation against varying amounts of adaptation data is investigated in this section. Table III shows the TDNN systems adapted with various amounts of adaptation data ranging from the first 5, 10, 20, 40, to all utterances of the CHE subset of $\text{Hub5'00}$. The aforementioned best configurations in Table III were used for the adaptation approaches (i.e., tied variances, $J_{\text{est}} = 1$ (equation (18)), and expectation (equation (8)) for inference were used for all Bayesian adapted systems; LHUC/BLHUC adaptation with identity activations were applied to the first 14 layers; HUB/BHUB adaptation with Tanh activations and were applied to the first 14 layers; PAct/BPAct adaptation were applied to the first layer). It could be seen that the Bayesian adapted systems consistently outperformed the systems adapted with deterministic SD parameters. Moreover, using only 5 utterances (11 seconds on average) as adaptation data for Bayesian test-time adaptation led to significant absolute WER reduction of 0.7% against the corresponding deterministic parameter adapted systems (Systems (3) v.s. (2), Systems (5) v.s. (4), and Systems (7) v.s. (6) in Table III respectively).

When incorporating i-vectors and SAT, similar trends could be observed by using the Bayesian adaptation approaches. When using only the first 5 utterances as adaptation data, the improvements achieved by Bayesian adaptation over deterministic adaptation were increased up to 1.4% absolute (7.2% relative) WER reduction (Systems (14) v.s. (13) in Table III). When using all speaker level data as adaptation data, the BLHUC adaptation (System (10)) significantly outperformed the LHUC adaptation (System (9)) by 2.1% absolute (11.2% relative) WER reduction. However, when incorporating i-vectors, the performance of applying Bayesian adaptation to SAT systems (Systems (16), (18), and (20)) produced no significant improvement over the non-SAT systems (Systems (10), (12), and (14)). This indicated the similar effects of i-vector based adaptive training and applying model-based SAT.

4) Advantages over Other Regularized Adaptation Methods: This section compares the BLHUC adaptation to
applying other regularizations to LHUC adaptation. Table IV shows the performance contrasts of these adapted TDNN systems without using i-vector. The regularization approaches included f-MLLR (Systems (2) and (3)) in Table IV which generated speaker-invariant input features and was shown to have good complementary effect to LHUC adaptation in [20]. KL divergence regularization [43] (System (4)), lattice based supervision [76] (System (5)) which alleviated the sensitivity to errors in the supervision, MAP adaptation [65] (Systems (6) and (7)), and noisy LHUC adaptation (Systems (8) and (9)) which added noise sampled from the fixed distribution prior distribution to the LHUC SD parameters. The noisy LHUC could be viewed as a special case of BLHUC with fixed variances and non-regularized update of SD distribution means. Among these regularization approaches, all configurations were tuned to produce the best performance. When using \( \mathcal{N}(1, 1) \) as the prior distribution, the MAP adaptation was simplified to L2 regularization [22]. The empirical prior distribution was computed by equation [23] in Section III-C2. The Table IV shows that the BLHUC adapted systems (Systems (10) and (11)) consistently outperformed all the other regularization approaches. The noisy LHUC (Systems (8) and (9)) yielded comparable performance to BLHUC (Systems (10) and (11)) when the amount of adaptation data is small (≤ 10 utterances). This is expected as limited updates were performed in this setting and the estimated posterior in BLHUC would not diverge too far from the prior. Moreover, no significant improvement was obtained by using an empirical prior (Systems (11)) against using \( p_0 = \mathcal{N}(1, 1) \) (Systems (10)) for BLHUC. Joint use of f-MLLR and BLHUC (System (12)) produced no improvement over using BLHUC alone.

### Table III

| Adapt method | i-vec | WER (%) w.r.t. #Adapt utterances |
|--------------|-------|----------------------------------|
| (1) -        | ×     | 20.6, 20.6, 20.6, 20.6, 20.6    |
| (2) LHUC     |       | 19.7, 19.2, 19.4, 19.3, 19.8    |
| (3) BLINEC   |       | 19.05, 18.9, 18.75, 18.7, 18.67 |
| (4) HUB      | ×     | 19.9, 19.5, 19.1, 18.7, 18.6    |
| (5) BHUB     |       | 19.2, 18.97, 18.8, 18.07, 17.97 |
| (6) PAct     |       | 19.9, 19.8, 19.6, 19.4, 19.6   |
| (7) BPAAct   |       | 19.2, 19.4, 19.2, 18.8, 18.5    |
| (8) -        | √     | 19.2, 19.2, 19.2, 19.2, 19.2    |
| (9) LHUC     |       | 19.1, 18.7, 18.5, 18.3, 18.8    |
| (10) BLHUC   |       | 18.1, 17.7, 17.3, 16.6, 16.7    |
| (11) HUB     | √     | 19.1, 18.6, 18.5, 18.0, 17.8    |
| (12) BHUB    |       | 17.8, 17.5, 17.2, 17.1, 16.9    |
| (13) PAct    |       | 19.5, 19.3, 18.6, 18.3, 18.4    |
| (14) BPAAct  |       | 18.15, 18.45, 18.3, 17.6, 17.51 |
| (15) SAT LHUC|       | 18.8, 18.7, 18.6, 18.3, 18.6    |
| (16) SAT BLHUC|      | 17.97, 17.7, 17.3, 16.9, 16.6   |
| (17) SAT HUB |       | 20.8, 19.7, 19.4, 18.0, 17.8    |
| (18) SAT BHUB|       | 19.07, 19.2, 18.6, 17.5, 17.1   |
| (19) SAT PAct|       | 18.9, 19.2, 18.9, 18.5, 18.3    |
| (20) SAT BPAAct|     | 17.97, 18.27, 17.6, 17.7, 17.5   |

### Table IV

| Adapt method | Prior \( p_0 \) | WER (%) w.r.t. #Adapt utterances |
|--------------|------------------|----------------------------------|
| (1) LHUC     | -                | 19.7, 19.2, 19.4, 19.3, 19.8    |
| (2) f-MLLR   | -                | 27.4, 25.2, 23.1, 19.8, 19.5    |
| (3) f-MLLR + LHUC | -             | 27.1, 23.8, 21.0, 19.6, 19.0    |
| (4) KL-regularized LHUC | -        | 19.6, 19.3, 19.2, 19.2, 19.8    |
| (5) Lattice based LHUC | -        | 19.9, 19.9, 19.7, 19.3, 19.2    |
| (6) MAP LHUC | \( \mathcal{N}(1, 1) \) | 19.6, 19.3, 19.2, 19.1, 19.4 |
| (7) MAP LHUC | Empirical | 19.7, 19.1, 19.3, 19.2, 19.8 |
| (8) Noisy LHUC (\( \sim p_0 \)) | \( \mathcal{N}(0, 0.5) \) | 19.2, 18.9, 19.0, 18.8, 19.6 |
| (9) Noisy LHUC (\( \sim p_0 \)) | \( \mathcal{N}(0, 1) \) | 19.2, 19.2, 19.0, 18.0, 18.0 |
| (10) BLHUC   | \( \mathcal{N}(1, 1) \) | 19.0, 18.9, 18.7, 18.5, 17.6 |
| (11) BLHUC   | Empirical | 19.4, 18.8, 18.4, 17.7, 17.5 |
| (12) f-MLLR + BLHUC | \( \mathcal{N}(1, 1) \) | 25.3, 22.4, 19.9, 18.3, 17.7 |

5) Sensitivity of Adaptation to Different Learning Rates: In the previous sections, the learning rates of adaptation were empirically adjusted and fixed (i.e., set to 0.01 for LHUC/BLHUC and PAct/BPAAct, and 0.00001 for HUB/BHUB) to prevent overfitting. Figure 2 shows the analysis about the effect of learning rate tuning on different adaptation techniques based on the systems using i-vector adaptation in Table III (Systems (9) to (20)). It shows that the Bayesian adaptation consistently outperforms the comparable deterministic adaptation using different learning rates (by comparing curves of the same colour with “×” and “○” legends in the sub-figure, representing each baseline deterministic adaptation method and its comparable Bayesian counterpart).

### C. Experiments on Data Augmented 900-Hour Switchboard Setup

In this section, speed perturbation on the rates of 0.9 and 1.1 was used to augment the 300-hour Switchboard data such that three times of training data totalling 864 hours could be used. In addition to TDNN systems, a baseline CNN-TDNN system was built following the standard Kaldi recipe. The CNN-TDNN employed 2-dimensional convolutional layers as the first 6 layers, where the two dimensions of the feature maps were the “height” and “time” axes. The input–output sizes of these convolutional layers in form of “filter×height” were \( 6×40-64×40-64×40-128×20-128×20-256×10-256×10 \) respectively. Intuitively at the first convolutional layer when using identity filter weights, the symbol “filter” represented the number of filters being used, while “height” denoted the spectral dimensional context. Each filter had the size \( 3×3 \) in form of “height×time”. ReLU activations and batch normalizations were utilized following the 2560-dimensional convolution outputs. Other configurations of the CNN-TDNN system were the same as the TDNN systems. The

<https://egs/swbd/s5c/local/chain/run_cnn_tddd_4a.sh>
same configurations of the Bayesian and deterministic adaptation approaches as in Table III were used in this section, except that, for the adapted CNN-TDNN systems, both identity and 2Sigmoid were used as activation functions in LHUC/BLHUC adaptation, and the second to sixth convolutional layers would not be adapted.

Moreover, two unidirectional (forward/backward) context based long short-term memory recurrent neural network language models (LSTM LMs) were built following the standard Kaldi recipe. The transcripts of both the Switchboard training data and the Fisher English corpora were encoded as 1024 dimensional embedding vectors to train the LSTM LMs. Each LSTM LM consisted of 2 unidirectional LSTM layers with 1024 cells. Projections were used inside the LSTMs to reduce the output dimensions to 512. A context splicing layer with ReLU activation was exploited before and after each LSTM layer. The splicing indices of the three context splicing layers were \{-1, 0\}, \{-3, 0\}, and \{-3, 0\}. Finally, an affine transformation was used to generate the output embedding vectors. For decoding, the two LSTM LMs were employed successively to rescore the lattices generated by the 4-gram LM. The supervision labels used in both deterministic and Bayesian adaptation were generated by the 1-best paths of LSTM LMs.

1) Performance of Bayesian Adaptation on 900-Hour Perturbed Setup: Table VI presents the results of adapted TDNN systems using varying amounts of adaptation data. Consistent improvements were achieved by the Bayesian adaptation approaches over the deterministic adaptation approaches. Among these results, the BLHUC adaptation (Systems (3) and (10) in Table VI) significantly outperformed the comparable LHUC adapted baseline system (System (9)) when using 10 utterances (22 seconds on average) of speech or more as the adaptation data.

Table V Performance of Applying Bayesian Adaptation to the Data Augmented 900-Hour TDNN System Using Varying Amounts of Adaptation Data Evaluated on the CHE Subset of Hub5’00. The 1-Vectors Were Used in These Systems. “†” Means That the Bayesian Adaptation Significantly Outperformed (P ≤ 0.05) the Corresponding Deterministic Adaptation. The Results in Bold Fonts Are the Lowest WERs in the Corresponding Columns.

| Adapt method | LSTM LM | WER (%) w.r.t. #Adapt utterances |
|--------------|---------|----------------------------------|
| (1) -        |         | 17.7                             |
| (2) LHUC     | ×       | 17.5                             |
| (3) BLHUC    | ×       | 16.9†                            |
| (4) HUB      | ×       | 17.8                             |
| (5) BHUB     | ×       | 17.6                             |
| (6) PAct     | ×       | 17.7                             |
| (7) BPAct    | ×       | 17.7                             |
| (8) -        |         | 15.4                             |
| (9) LHUC     | √       | 15.0                             |
| (10) BLHUC   | √       | 14.8†                            |
| (11) HUB     | √       | 15.2                            |
| (12) BHUB    | √       | 15.2                            |
| (13) PAct    | √       | 15.3                            |
| (14) BPAct   | √       | 15.4                            |

Table VI shows the results of applying Bayesian adaptation approaches to both TDNN and CNN-TDNN systems evaluated on the Hub5’00 and RT03 data. All speaker level data was used as adaptation data. It can be found that the Bayesian learning approaches, i.e., BLHUC, BHUB, and BPAct adaptation consistently outperformed both the corresponding deterministic approaches, i.e., LHUC, HUB, and PAct adaptation, as well
as the un-adapted baseline system. Especially on the RT03 set, all the improvements produced by the Bayesian adaptation over deterministic adaptation are significant. In the CNN-TDNN systems, there was no significant difference between using 2Sigmoid activation and identity activation for BLHUC adaptation (Systems (17.1) v.s. (17.2)). Therefore, only the 2Sigmoid activation was used for LHUC/BLHUC adaptation in the CNN-TDNN systems in the rest of this paper. When using LSTM LM rescoring, the BLHUC adapted systems significantly outperformed the comparable LHUC adapted systems by relative WER reduction up to 10% (The Avg. WERs on RT03 in Systems (10) v.s. (9) in Table VI).

### TABLE VI

OVERALL PERFORMANCE OF APPLYING BAYESIAN ADAPTATION TO THE DATA AUGMENTED 900-HOUR TDNN AND CNN-TDNN SYSTEMS EVALUATED ON THE Hub5’00 and RT03 DATA SETS. THE I-VECTORS WERE USED IN THESE SYSTEMS. THE ADAPTED SYSTEMS DENOTED BY “LHUC*”/“BLHUC*” USED IDENTITY ACTIVATION FUNCTIONS, AND THOSE DENOTED BY “LHUC**”/“BLHUC**” USED 2SIGMOID ACTIVATION FUNCTIONS. “+” MEANS THAT THE BAYESIAN ADAPTATION SIGNIFICANTLY OUTPERFORMED (P < 0.05) THE CORRESPONDING DETERMINISTIC ADAPTATION. THE RESULTS IN BOLT FONTS ARE THE LOWEST WERs IN THE CORRESPONDING COLUMNS.

| Model | Adapt method | LSTM LM | WER (%) of Hub5’00 | WER (%) of RT03 |
|-------|--------------|---------|---------------------|-----------------|
| (1)   |              |         |                     |                 |
| (2)   |              | ×       |                     |                 |
| (3)   |              | ×       |                     |                 |
| (4)   |              | ×       |                     |                 |
| (5)   |              | ×       |                     |                 |
| (6)   |              | ×       |                     |                 |
| (7)   |              | ×       |                     |                 |
| (8)   |              | ×       |                     |                 |
| (9)   |              | ×       |                     |                 |
| (10)  |              | ×       |                     |                 |
| (11)  |              | ×       |                     |                 |
| (12)  |              | ×       |                     |                 |
| (13)  |              | ×       |                     |                 |
| (14)  |              | ×       |                     |                 |

### Table VII

### V. DISCUSSION AND CONCLUSIONS

This paper proposes a DNN speaker adaptation framework using full Bayesian learning to account for SD parameter uncertainty given limited speaker specific adaptation data. Three exemplar approaches of DNN based speaker adaptation including Bayesian learning of hidden unit contributions (BLHUC), Bayesian parameterized activation functions (BPAct), and Bayesian hidden unit bias vectors (BHUB) are obtained by applying the Bayesian learning framework to the approaches using deterministic parameters including LHUC, parameterized activation functions (PAct), and hidden unit bias vectors (HUB). Latent variable posterior distributions are learned for each speaker in all of these Bayesian adaptation methods. An efficient variational inference based approach benefited from the Monte Carlo parameter sampling and re-parameterization is used to estimate the posterior parameters. Experiments conducted on the 300-hour speed perturbed Switchboard setup of LF-MMI factored TDNN/CNN-TDNN systems using i-vector speaker adaptive training suggest the proposed Bayesian adaptation consistently outperforms the deterministic adaptation on the NIST Hub5’00 and RT03 evaluation sets. The efficacy of the proposed Bayesian adaptation techniques is demonstrated in a comparison against the state-of-the-art performance obtained on the most recent hybrid and end-to-end systems reported in the literature. The best proposed BLHUC adapted CNN-TDNN system achieves the state-of-the-art performance with 10.7% WER on the RT03 evaluation set conducted on this task. Future works may focus on the non-variational inference of the Bayesian adaptation.

### APPENDIX A

**CODES OF THE BAYESIAN ADAPTATION BASED ON KALDI TOOLKIT**

The codes of the Bayesian adaptation is available in https://github.com/XIEXueling/kaldi.git. They are written based on the Kaldi toolkit [71]. Please cite the published version of this article in the TASLP (DOI: 10.1109/TASLP.2021.3084072) if you use these codes.
TABLE VII

Performance contrasts of our BLHUC adapted CNN-TDNN systems rescored by larger LSTM LMs against other state-of-the-art systems conducted on the 300-hour Switchboard setup. The overall WERs in "(c)" are not reported in the literature and are estimated using the subset WERs. The results in bold fonts are the lowest WERs in the corresponding columns.

| System | #Param | WER (%) of Hub5'00 | WER (%) of RT03 |
|--------|--------|---------------------|---------------------|
|        |        | SWRD | CHE | Avg. | SWRD | CHE | Avg. |
| (1) SMBR BLSTM hybrid + i-vectors [72] | - | 6.7 | 14.7 | 10.7 | - | - | - |
| (2) + Affine transform based environment adaptation | - | 6.7 | 13.5 | 10.2 | - | - | - |
| (3) Listen, attend and spell network + SpecAugment [72] | - | 6.8 | 14.1 | (10.5) | - | - | - |
| (4) LSTM based attention encoder-decoder + SpecAugment [73] | 28.5M | 7.4 | 14.6 | (11.0) | - | - | - |
| (5) + LHUC based speaker adaptation | 75.1M | 6.8 | 13.4 | (10.1) | - | - | - |
| (6) 2801M | 6.4 | 12.5 | (9.5) | 14.8 | 8.4 | (11.7) |
| (7) Hybrid encoder-decoder + CTC + SpecAugment [79] | 6.3 | 13.3 | (9.8) | - | - | - |
| (8) LF-MMI CNN-TDNN + i-vectors (ours) | 15.1M | 7.1 | 14.1 | 10.6 | 15.9 | 9.4 | 12.8 |
| (9) + LHUC based speaker adaptation | + 11.8k | 6.8 | 13.5 | 10.2 | 14.5 | 8.6 | 11.7 |
| (10) + BLHUC based speaker adaptation | + 11.8k | 6.8 | 12.8 | 9.7 | 13.8 | 8.2 | 11.1 |
| (11) LF-MMMI CNN-TDNN + i-vectors + SpecAugment (ours) | 15.1M | 7.2 | 13.9 | 10.6 | 15.1 | 9.1 | 12.2 |
| (12) + LHUC based speaker adaptation | + 11.8k | 6.8 | 13.0 | 9.9 | 14.0 | 8.3 | 11.3 |
| (13) + BLHUC based speaker adaptation | + 11.8k | 6.7 | 12.7 | 9.7 | 13.4 | 7.9 | 10.7 |

APPENDIX B

Updated Results Using The Bayesian Adaptation

This part shows some new results obtained after the article submitted.

A. Bayesian Learning of Linear Hidden Network

As mentioned in Section I-A linear hidden networks (LHN) [40] is a widely used transform based adaptation approach. This part shows the performance of applying Bayesian learning to the LHN, which is called Bayesian LHN (BLHN). This experiment was conducted on the 300-hour Switchboard setup without using the i-vector based adaptive training, as shown in Table II. The LHN adaptation was applied following the 160-dimensional linear projection of the second layer of the TDNN system, and estimated a 160 × 160 dimensional transformation for each speaker. The prior distribution for both MAP LHN and BLHUN was empirically estimated as in Section IV-B. On the CHE subset, applying Bayesian learning to LHN adaptation achieved significant WER reduction compared to the standard LHN and MAP LHN. This shows the effectiveness of the proposed Bayesian learning framework on transform based adaptation approach.

TABLE VIII

Performance of linear hidden network (LHN) adaptation and its improvements using maximum a posteriori (MAP) adaptation [63] and Bayesian learning based adaptation.

| Adapt method | Initialization | WER (%) of Hub5'00 |
|--------------|---------------|---------------------|
|              |               | SWRD | CHE | Avg. |
| (1) LHN      | Random        | 10.1 | 20.6 | 15.4 |
| (2) MAP LHN  | Random        | 10.0 | 19.8 | 15.0 |
| (3) BLHUN     | Random        | 9.8  | 19.3 | 14.6 |
| (4) LHN      | Identity      | 9.9  | 19.9 | 14.9 |
| (5) MAP LHN  | Identity      | 9.9  | 19.7 | 14.8 |
| (6) BLHUN     | Identity      | 9.8  | 19.0 | 14.4 |

B. Results on Baseline Systems Using Parallel Models Combination

In the published article, the baseline systems were trained on single job without parallel models combination in each iteration due to the limited computing resource. As a complement, this part shows the performance of the BLHUC adaptation applied to the TDNN and CNN-TDNN systems trained with parallel models combination in each iteration. The Table IX is corresponding to the BLHUC results in Table VII. The default parallel setting in the Kaldi toolkit was used, where the baseline systems were trained with 6 epochs beginning with 3 jobs and ending with 16 jobs. It shows that the Bayesian adaptation can be used together with the parallel models combination to obtain further improvement. For this purpose, Table IX shows the new results corresponding to the Table VII with parallel models combination used in systems (8)–(13). It shows that the BLHUC adapted systems with models combination (systems (10) and (13)) achieve the best overall performance on both Hub5'00 and RT03 data sets.

TABLE IX

Overall performance of applying BLHUC to the data augmented 900-hour TDNN and CNN-TDNN systems evaluated on the Hub5'00 and RT03 data sets. The i-vectors were used in these systems.

| System | Adapt method | LSTM LM | WER (%) of Hub5'00 | WER (%) of RT03 |
|--------|--------------|---------|---------------------|---------------------|
|        |              | SWRD | CHE | Avg. | SWRD | CHE | Avg. |
|        |              |       |       |       |       |       |       |
| TDNN   | -             | 9.4  | 17.7 | 13.6  | 19.8 | 12.3 | 16.2  |
|        | BLHUC        | 9.5  | 17.3 | 13.4  | 19.4 | 12.0 | 15.7  |
|        | LHUC         | 8.9  | 16.0 | 12.5  | 17.2 | 10.8 | 14.2  |
|        | BLHUC + Model Combine | 7.9 | 15.9 | 11.7 | 17.4 | 10.4 | 14.0  |
|        | LHUC + Model Combine | 7.5 | 15.3 | 10.9 | 17.0 | 10.1 | 13.5  |
|        | LHUC         | 8.6  | 16.8 | 12.8  | 18.1 | 11.2 | 14.8  |
|        | BLHUC        | 8.2  | 15.5 | 11.9  | 15.6 | 10.2 | 13.0  |
|        | LHUC + Model Combine | 7.1 | 14.4 | 10.8 | 15.8 | 9.5  | 12.8  |
|        | LHUC         | 7.0  | 13.1 | 10.1 | 13.6 | 8.5  | 11.1  |
|        | BLHUC        | 8.3  | 16.7 | 12.9  | 19.1 | 11.7 | 15.5  |
|        | LHUC         | 8.5  | 16.2 | 12.4  | 17.9 | 10.8 | 14.5  |
|        | BLHUC        | 8.3  | 15.7 | 12.0  | 17.2 | 10.3 | 13.9  |
|        | LHUC         | 7.1  | 13.9 | 10.5 | 15.3 | 8.9  | 12.2  |
|        | BLHUC        | 6.9  | 13.2 | 10.1 | 14.5 | 8.5  | 11.6  |
|        | LHUC         | 8.4  | 16.0 | 12.3  | 17.4 | 10.6 | 14.2  |
|        | BLHUC        | 8.4  | 15.3 | 11.9 | 16.5 | 10.1 | 13.5  |
|        | LHUC         | 7.3  | 13.7 | 10.3 | 14.8 | 8.8  | 11.9  |
|        | BLHUC        | 7.2  | 13.3 | 10.3 | 14.0 | 8.4  | 11.3  |
TABLE X

Performance contrasts of out BLHUC adapted CNN-TDNN systems rescored by larger LSTM LMs against other state-of-the-art systems conducted on the 300-hour Switchboard setup. The overall WERs in “()” are not reported in the literature and are estimated using the subset WERs.

| System | #Param | WER (%) of Hub5’00 | WER (%) of RT03 |
|--------|--------|---------------------|----------------|
|        | SWRD   | CMC                | SWRD           |
|        |        |        | Avg. |        |        |        | Avg. |
| (1) SMBR BLSTM hybrid + i-vectors [79] | - | 6.7 | 14.7 | 10.7 | - | - | - |
| (2) + Affline transform based environment adaptation | - | 6.7 | 13.5 | 10.2 | - | - | - |
| (3) Listen, attend and spell network + SpecAugment [82] | - | 6.8 | 14.1 | (10.3) | - | - | - |
| (4) LSTM based attention encoder-decoder + SpecAugment [79] | 28.5M | 7.4 | 14.6 | (11.0) | - | - | - |
| (5) 75.1M | 6.8 | 13.4 | (10.1) | - | - | - |
| (6) 280.1M | 6.4 | 2.5 | (9.9) | 14.8 | 8.4 | (11.7) |
| (7) Hybrid encoder-decoder + CTC + SpecAugment [80] | 6.3 | 13.3 | (9.8) | - | - | 11.4 |
| (8) LF-MMI CNN-TDNN + i-vectors + SpecAugment + Model Combine (ours) | 15.1M | 6.9 | 13.2 | 10.1 | 14.6 | 8.4 | 11.7 |
| (9) + LHUC based speaker adaptation | + 11.8k | 6.6 | 12.5 | 9.6 | 13.6 | 8.1 | 11.0 |
| (10) + BLHUC based speaker adaptation | + 11.8k | 6.5 | 12.2 | 9.4 | 12.8 | 7.6 | 10.3 |

C. Experiment on The 150-hour HKUST Task

The published article conducted experiments only on the Switchboard data. As a complement, this part shows the performance of LHUC/BLHUC adaptation on the 150-hour HKUST Mandarin conversational telephone speech database (LDC2005S15, LDC2005T32). The TDNN and CNN-TDNN system were built with the same configurations as the Switchboard setup. Parallel models combination, speed perturbation (to 450 hours), and i-vector adaptation were used for training. The systems were evaluated on data sets of development set (Dev) released in 2014 [81], NIST RT03S (LDC2007S10), and 1997 NIST Hub5 Mandarin evaluation set [23] (EVAL97). The LSTM LM was trained with the transcription of training data. Table [XII] shows the performance of applying LHUC/BLHUC adaptation to the TDNN and CNN-TDNN systems.

| Model | Adap method | LSTM LM | Speed perturbation | Adapt method | LHUC SAT | WER (%) | Participant | Investigator | WER (%) | Average |
|-------|-------------|---------|--------------------|--------------|----------|---------|-------------|-------------|---------|---------|
| (1)  | -           |         | (1)               |               |          |        |             |             |         |         |
| (2)  | LHUC        | ×       | (2)               | LHUC         | ×        | 38.4   | 17.6       | 29.3        |         |         |
| (3)  | BLHUC       |         | (3)               | BLHUC        | ×        | 37.5   | 17.2       | 28.6        |         |         |
| (4)  | -           |         | (4)               | LHUC         | ×        | 37.3   | 17.1       | 28.4        |         |         |
| (5)  | LHUC        | √       | (5)               | LHUC         | √        | 36.4   | 16.6       | 27.8        |         |         |
| (6)  | -           |         | (6)               | ×            |          |        |             |             |         |         |
| (7)  | LHUC        | ×       | (7)               | LHUC         | ×        | 36.9   | 17.1       | 28.3        |         |         |
| (8)  | BLHUC       |         | (8)               | BLHUC        | ×        | 36.0   | 16.6       | 27.5        |         |         |
| (9)  | LHUC        | √       | (9)               | LHUC         | √        | 35.4   | 16.4       | 27.1        |         |         |
| (10) | BLHUC       |         | (10)              | BLHUC        | √        | 34.9   | 16.3       | 26.8        |         |         |

D. Experiment on The DementiaBank Pitt Elderly Speech

This part shows the performance of LHUC/BLHUC adaptation on the DementiaBank Pitt corpus [82]. In the experiment, 15.75 hours of training data with 9.72 hours elderly participant data and 6.03 hours investigator data was used after silence stripping, and 3.14 hours of test data with 1.93 hours elderly participant data and 1.21 hours investigator data was used. CNN-TDNN was trained with the same configuration of Appendix B-C. Table [XII] shows the performance of applying LHUC/BLHUC adaptation as well as SAT to the CNN-TDNN system.

ACKNOWLEDGMENT

This research is supported by Hong Kong Research Grants Council GRF grant No. 14200218, 14200220, Theme-based Research Scheme T45-407/19N, Innovation & Technology Fund grant No. ITS/254/19, PhH/350/20, Shun Hing Institute of Advanced Engineering grant No. MMT-p1-19, National Natural Science Foundation of China (U1736202), and National Key R&D Program of China (2020YFC2004100)

REFERENCES

[1] G. E. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,” IEEE Transactions on audio, speech, and language processing, vol. 20, no. 1, pp. 30–42, 2011.
[2] H. A. Bourlard and N. Morgan, Connectionist speech recognition: a hybrid approach. Springer Science & Business Media, 2012, vol. 247.
[3] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, and G. Penn, “Applying convolutional neural networks concepts to hybrid nn-hmms model for speech recognition,” in 2012 IEEE international conference on Acoustics, speech and signal processing (ICASSP), 2012, pp. 4277–4280.
[4] O. Abdel-Hamid, L. Deng, and D. Yu, “Exploring convolutional neural network structures and optimization techniques for speech recognition,” in Interspeech, vol. 11, 2013, pp. 73–5.
Xunying Liu (Member, IEEE) received the undergraduate degree from Shanghai Jiao Tong University, Shanghai, China, and the M.Phil. degree in computer speech and language processing and the Ph.D. degree in speech recognition from the University of Cambridge, Cambridge, U.K. He was a Senior Research Associate with the Machine Intelligence Laboratory, Cambridge University Engineering Department, and since 2016, has been an Associate Professor with the Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong. His current research interests include machine learning, large vocabulary continuous speech recognition, statistical language modeling, noise robust speech recognition, audio-visual speech recognition, computer aided language learning, speech synthesis, and assistive technology. He is a Member of the ISCA. He and his students were the recipient of the number of best paper awards and nominations, including the Best Paper Award at ISCA Interspeech 2010 for the paper titled “Language Model Cross Adaptation for LVCSR System Combination”, and the Best Paper Award at IEEE ICASSP2019 for their paper titled “BLHUC: Bayesian Learning of Hidden Unit Contributions for Deep Neural Network Speaker Adaptation”.

Tan Lee (Member, IEEE) is an Associate Professor at the Department of Electronic Engineering and concurrently the Assistant Dean for Education at the Faculty of Engineering, the Chinese University of Hong Kong. His research covers speech and audio signal processing, automatic speech recognition, text-to-speech, paralinguistics in speech, and neurological basis of speech and language. Tan Lee led the effort on developing Cantonese-focused spoken language technologies that have been widely licensed for industrial applications. His recent work is focused on applying signal processing techniques and deep learning models to atypical speech in the context of low-resource languages and human communication disorders across different age groups. Tan Lee is an Associate Editor for the IEEE/ACM Transactions on Audio, Speech and Language Processing. He was an Area Chair of the Technical Programme Committees of INTERSPEECH 2014, 2016, and 2018, the General Co-Chair of ISCSLP 2018 and 2021. He is the Vice Chair of ISCA Special Interest Group on Chinese Spoken Language Processing.

Lan Wang (Member, IEEE) received the M.S. degree from the Center of Information Science, Peking University, Beijing, China, and the Ph.D. degree from the Machine Intelligence Laboratory, Department of Engineering, University of Cambridge, Cambridge, U.K. She worked on the autonomous Global Integrated Language Exploitation project funded under DARPA’s Global Autonomous Language Exploitation program. She is currently a Research Professor with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. Her research interests include large vocabulary continuous speech recognition, speech visualization, and audio-visual speech processing.