Bayesian Unsupervised Batch and Online Speaker Adaptation of Activation Function Parameters in Deep Models for Automatic Speech Recognition

Zhen Huang, *Student Member, IEEE*, Sabato Marco Siniscalchi, *Member, IEEE*, and Chin-Hui Lee, *Fellow, IEEE*

***Abstract*—We present a Bayesian framework to obtain maximum *a posteriori* (MAP) estimation of a small set of hidden activation function parameters in context-dependent-deep neural network-hidden markov model (CD-DNN-HMM)-based automatic speech recognition (ASR) systems. When applied to speaker adaptation, we aim at transfer learning from a well-trained deep model for a "general" usage to a "personalized" model geared toward a particular talker by using a collection of speaker-specific data. To make the framework applicable to practical situations, we perform adaptation in an unsupervised manner assuming that the transcriptions of the adaptation utterances are not readily available to the ASR system. We conduct a series of comprehensive batch adaptation experiments on the Switchboard ASR task and show that the proposed approach is effective even with CD-DNN-HMM built with discriminative sequential training. Indeed, MAP speaker adaptation reduces the word error rate (WER) to 20.1% from an initial 21.9% on the full NIST 2000 Hub5 benchmark test set. Moreover, MAP speaker adaptation compares favorably with other techniques evaluated on the same speech tasks. We also demonstrate its complementarity to other approaches by applying MAP adaptation to CD-DNN-HMM trained with speaker adaptive features generated through constrained maximum likelihood linear regression and further reduces the WER to 18.6%. Leveraging upon the intrinsic recursive nature in Bayesian adaptation and mitigating possible system constraints on batch learning, we also proposed an incremental approach to unsupervised online speaker adaptation by simultaneously updating the hyperparameters of the approximate posterior densities and the DNN parameters *sequentially*. The advantage of such a sequential learning algorithm over a batch method is not necessarily in the final performance, but in computational efficiency and reduced storage needs, without having to wait for all the data to be processed. So far, the experimental results are promising.**

***Index Terms*—Automatic speech recognition, Bayesian learning, deep neural networks, online adaptation, prior evolution, transfer learning, unsupervised speaker adaptation.**

I. Introduction

In RECENT years, context-dependent, deep neural network, hidden Markov models (CD-DNN-HMMs) have become the state-of-the-art acoustic modeling technique to build automatic speech recognition (ASR) engines outperforming conventional discriminatively trained Gaussian mixture model (GMM) based systems in different tasks and datasets [1]. DNNs provide a scaled state likelihood estimate by replacing the GMM in the acoustic modeling component [2]. Unfortunately, CD-DNN-HMMs, similarly to conventional CD-GMM-HMMs [3], also suffer from a performance drop under potential mismatched conditions between training and testing, and a degradation of the recognition accuracy is typically observed when channel conditions change, or when moving from a speaker-dependent (SD) to a speaker-independent (SI) environment due to inter-speaker variability [4]. A possible, and simple approach to enhance ASR robustness is to collect a huge amount of training material in an attempt to create acoustic models to cover all possible acoustic variabilities in speech. However, robust speech recognition cannot be solved simply by collecting more data, since the data-labeling process can be very expensive to carry out. Moreover, it can also be quite complicated to collect the training material in many real-world applications, e.g., under-resourced languages, e.g., [5]. Since the amount of data available for the specific working condition is usually quite limited, re-building a new ASR from scratch would definitely lead to over-fitting on these data.

Acoustic model adaptation has demonstrated to be a valid and effective approach to modify the acoustic model parameters to better resemble the evaluation data, as testified by the great deal of DNN adaptation techniques available in the recent literature, e.g., [6]–[18]. The key idea of any adaptation algorithm is like this: starting from a pre-trained (e.g., speaker and/or task independent) speech recognition system, for a new user (or a group of users) to use the system for a specific task, a small amount of *adaptation data* is collected from the user. These data are employed to construct a speaker adaptive system for the speaker in the particular environment for that specific application. By doing so, the mismatch between training and testing can be generally reduced. However, adapting parameters in a CD-DNN-HMM is much more challenging than in earlier connectionist ASR systems due to the large number of network nodes, and branches along with the huge number of tied HMM states, often referred to as senones [19]. Furthermore, the posterior probabilities for the unobserved and scarcely seen senones are often pushed towards zero during adaptation, because DNN parameters are adapted by every sample frame regardless of its senone class. The latter phenomenon is commonly referred to as catastrophic forgetting [20].

Manuscript received August 4, 2016; revised October 18, 2016 and October 21, 2016; accepted October 21, 2016. Date of publication October 26, 2016; date of current version November 28, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiaodong Cui.

Z. Huang is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: huangzhenee@gmail.com).

S. M. Siniscalchi is with the Faculty of Architecture and Engineering, University of Enna "Kore," Enna 94100, Italy, and also with the Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: marco.siniscalchi@unikore.it).

C.-H. Lee is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: chl@ece.gatech.edu).

Digital Object Identifier 10.1109/TASLP.2016.2621669

Conservative ad-hoc solutions for CD-DNN-HMMs have been proposed to address the aforementioned catastrophic forgetting phenomenon. For example, a Kullback-Leibler divergence (KLD) based objective criterion to be used during adaptation was devised in [6]. A variation to the standard method of assigning the target values, referred to as conservative training, was instead discussed in [21]. Bayesian adaptation techniques developed within the CD-GMM-HMM framework, such as maximum *a posteriori* (MAP) [22], and structured MAP (SMAP) [23], have also shown to be the optimal candidates to address data scarcity issues while providing a highly desirable asymptotic behavior as the number of adaptation sentences increases toward infinity; for example, MAP adapted acoustic models approach speaker-dependent models trained with maximum likelihood (ML) estimation [22]. The key idea of MAP adaptation scheme for deep models, as discussed in [24], is to assume that same of the DNN parameters of the ASR system are distributed according to a prior distribution that summarises all knowledge learned by the DNN to address the source task. This prior allows us to find the most probable model with respect to the target data under MAP.

Besides the over-fitting problem, environment personalisation with a neural model creates a storage problem as well, since a different set of neural parameters has to be stored for each specific target speaker. Indeed, a good adaptation technique should (i) minimize the storage requirements, since a different set of adaptation parameters needs to be stored for each different condition, (ii) combat catastrophic forgetting issues, and (iii) allow a meaningful system performance improvement even when a small amount of adaptation data is available. To meet these challenging requirements, we present a Bayesian adaptation framework to obtain maximum *a posteriori* (MAP) estimation of a small set of hidden activation function parameters in CD-DNN-HMM based ASR systems. When applied to speaker adaptation, we aim at transfer learning from a well-trained deep model for a "general" usage to a "personalized" model geared towards a particular talker using a collection of speech data provided by that speaker. To make the framework applicable to practical situations, we perform adaptation in an unsupervised manner assuming the transcription information of the adaptation utterances are not readily available to the ASR system.

Unsupervised learning is usually more realistic, and desirable, we therefore consider unsupervised adaptation in this paper, as aforementioned. Nonetheless, a sequential algorithm is even more attractive in real production, since it allows to adaptively track the varying parameters [25]. This learning scheme is often referred to as the *online adaptation* and makes the ASR system capable of continuously adjusting itself to a new operational environment without the requirement of storing a large set of previously used training data. The Bayesian inference theory again provides a good vehicle to formulate and solve this problem. Therefore, we also present a sequential version of the proposed MAP adaptation technique, which allows us to perform unsupervised, online speaker adaptation. The implied algorithm is adaptive in nature, and it can be used to perform a full-scale online adaptive learning of the CD-DNN-HMM parameters only using the current available data to continuously track the varying acoustic conditions through the prior evolution mechanism.

In the speech recognition community, the term *unsupervised* adaptation has often been used loosely, and it actually refers to *semi-supervised* learning in machine learning, as clearly pointed out in [26]. To avoid possible confusion while assessing our experimental result, we briefly discuss here the terminology used in this work. Unsupervised adaptation accounts for using a seed ASR engine to decode un-transcribed data for a specific test condition, e.g. target speaker. New acoustic models, which should better resemble the test condition, are then built using these automatic transcripts as the label during acoustic model adaptation. Furthermore, the adaptation algorithm can be carried out in a *batch*, or *online incremental* fashion. In the first case, all adaptation data are available at the same time; whereas, spoken utterances are processed one-by-one (or in mini-batches) in the latter case. It is also a common practice in the speech community to adapt and test on the same data, e.g., [10]. In this work, we therefore use the same NIST 2000 Hub5 evaluation material for adaptation and evaluation in order to make our experimental environment comparable to other investigations. Moreover, we analyze the *self-adaptation* performance of the proposed online MAP adaptation approach. Self-adaptation is similar to the *self-training* concept in machine learning [27], namely: the adaptation algorithm iteratively adapts a seed classifier by making predictions on the unlabeled test data, which are processed one-by-one, to expand the adaptation set [26].

We will assess the feasibility of the proposed solution on a *speaker adaptation* task and demonstrate consistent recognition error reductions on the Switchboard spontaneous speech recognition benchmark [28]. We will show through a series of comparative experiments that (i) the proposed solution is effective even when applied to already high-accuracy CD-DNN-HMMs trained in a sequence discriminative manner [29], [30], (ii) the proposed solution compares favorably against conventional linear network based adaptation schemes, and with other techniques evaluated on the same speech tasks. We also demonstrate its complementarity to other approaches by applying MAP adaptation to CD-DNN-HMM trained with speaker adaptive features, which is generated through constrained maximum likelihood linear regression (fMLLR) and obtain a WER of 18.6%.

It is important to point out that the speaker adaptation technique presented in this work is not limited to a mere extension of the MAP-based adaptation approach proposed in [24], and the two approaches differ in many respects. We highlight here some of the chief differences. First, speaker adaptation is confined to a special linear hidden layer injected right before the softmax layer in [24]; whereas, we propose to parametrize the activation function in this paper. These learnable activation functions are adjusted during the adaptation phase, and this solution offers two advantages: (i) The number of learnable parameters is limited to only twice the number of hidden nodes, and that minimize the storage requirements without introducing any special constraint that keeps the amount of parameters within reasonable limits. In fact, a bottleneck non-linear top layer had to be employed to constraint the amount of adaptation parameters in [24], and (ii) Adaptation equally affects all layers in the deep acoustic model

rather than some specific, heuristically chosen layers, as in [24]. Second, the hierarchical adaptation scheme presented in [24] is limited to the spatial dimension, and no temporal evolution of the prior distribution is investigated. In this paper, we fully leverage the Bayesian framework and exploit the hierarchical relationships among prior parameters in time. That involves the time dimension through the evolution of the prior information. Third, a spontaneous speech recognition task and unsupervised adaptation is addressed in this work; whereas, a much simpler read speech task and supervised adaptation were studied in [24]. Finally, self-learning is investigated in this paper.

We would like to finally remark that that feed-forward deep neural networks equipped with memory blocks [31] have been proven competitive to long short-term memory (LSTM) networks [32], [33], which represent the state-of-the-art acoustic modeling technique on the Switchboard task [34], [35]. Hence the use of the feedforward deep models is not an oversimplification with respect to the final goal of this paper. Nevertheless, to further emphasize the effectiveness of the proposed technique and prove that an improvement can be demonstrated even using ASR engines attaining state-of-the-art word accuracies, we evaluate our approach against a very challenging experimental scenario by employing CD-DNN-HMMs trained on speaker compensated features, which have been obtained by applying feature space transformations - referred to as fMLLR [36], [37], to the input features.

The rest of the paper is organized as follows. In Section II, we describe related work in adaptive learning for ASR. In Section III, a brief overview on a connectionist approach to ASR based on CD-DNN-HMM is presented. In Section IV, we present our proposed Bayesian learning approach to adapting a small set of activation function parameters in deep models. A series of comprehensive experiments is then conducted using the Switchboard ASR corpus on unsupervised, batch and online, speaker adaptation. The results are presented in Section V. Finally we summarize our findings in Section VI.

II. Related Work

Broadly speaking, we can categorize speaker adaptation techniques for connectionist ASRs into two main groups, namely: feature and model spaces. In recent years, there has also been the tendency to augment the conventional speech vector with speaker-specific features that hopefully confer robustness against speaker variability. We refer to these techniques simply as *other approaches*, e.g., i-vectors [38] are employed in [39], and speaker discriminative vectors are implemented in [40]. A brief overview of the most representative adaptation techniques in each group is given in the following.

A. Feature Space Adaptation

The most representative example of feature space adaptation of deep models is the constrained maximum likelihood linear regression (MLLR) technique, referred to as CMLLR or fMLLR [36], [37]. fMLLR estimates a set of affine transforms to be applied to input acoustic features and generate speech vectors more robust to training/testing mismatches. The affine transform is found maximizing the log-likelihood that the model generates the adaptation data based on first pass alignments. fMLLR has proven to be effective for adaptation of hybrid ASR engines in different tasks, e.g., [35], [41]. A major limit of fMLLR is that a CD-GMM-HMM system has to be built to generate a single input transform per each speaker. The transformed feature vectors can then be used to train the CD-DNN-HMM in a speaker adaptive manner, and another set of transforms is estimated (using the available GMMs) during evaluation for unseen speakers. The latter technique is commonly referred to as speaker adaptive training (SAT).

To simulate fMLLR without the burden of building CD-GMM-HMM systems, a linear transformation network can be added to the input of the DNN, which can be directly estimated by minimising the error at the output of the neural architecture while keeping all other DNN parameters frozen. Such a transformation rotates the input space to reduce the discrepancy between testing and training conditions. This approach is commonly referred to as linear input network (LIN), and its goal is to map the speaker dependent (SD) input vectors to the speaker independent (SI) ASR system [42]. LIN, and its feature-space discriminative linear regression (fDLR) variant have been tested with success, e.g., [11], [43], [44].

B. Model Space Adaptation

The simplest approach to adapt a hybrid ASR system is to modify all parameters of the connectionist block using some adaptation data. Unfortunately, this solution easily leads to overfitting on the adaptation material when the amount of data is limited [42]. A successful regularization based method to address over-fitting issues has been proposed in [6], where the Kullback-Liebler divergence (KLD) between the speaker-independent output distribution and the speaker-adapted output distributions was used during adaptation. In [45]–[47], a factorization technique based on singular value decomposition was instead devised to reduce the number of parameters to be adapted. In [21], the authors proposed to add a linear transformation network before the output layer, referred to as linear hidden network (LHN). The rationale behind LHN is that the added linear layer generates discriminative features of the input pattern suitable for the classification performed at the output of the DNN. Over-fitting issues, can be further reduced by adapting the DNN top layer in a maximum a posteriori (MAP) fashion [24]. Another way to address the the data sparsity problem in CD-DNN-HMM adaptation was proposed in [9], where one or more small auxiliary output layers modeling broad acoustic units, such as mono-phones or senone-clusters, were added to the original DNN structure. DNN parameters were then adapted through multi-task learning.

In [46], it was indeed argued that the large number of DNN parameters for ASR makes adaptation very challenging, and it also limits the use of environmental personalization due to the huge storage cost in large-scale deployments, and it may require a computationally intensive adaptation process. In [13], the authors proposed an ingenious solution to perform speaker adaptation for hybrid ASR systems that can simultaneously al-

low us to (i) reduce the computational requirements, (ii) address overfitting issues, and (iii) store the adapted parameters in a small-sized storage space. The key idea was to adapt the shape of the hidden activation functions rather than some network parameters. To this end, Hermite polynomial activation functions were used in the hidden neurons. Later, it was demonstrated that slope and bias parameters introduced in the sigmoid activation function can be also learned in a speaker adaptive fashion [48]. Following the same line of research, an adaptive linear factor associated with each hidden unit is used to scale the unit output value and create a speaker dependent model was proposed in [49]. In practice, DNN adaptation becomes a re-weighting of the importance of different hidden units for every speaker. In the learning hidden unit contributions (LHUC) technique [10], an additional amplitude parameter is added for each hidden unit. These amplitude parameters are tied for each speaker, and are learned using unsupervised adaptation. In [43], only output layer biases have been adapted.

In the proposed approach, we assume that the activation function parameters are distributed according to a prior distribution that summarise all knowledge learned to address the source task. This prior allows us to find the most probable model with respect to the target data under MAP, which strengths robustness to data scarcity, as demonstrated in [25].

C. Other Approaches

It has been shown that robustness to speaker variability can be gained by appending speaker-specific features, computed for each speaker at both training and test stages, to the conventional speech vectors. In particular, i-vectors [38], which can be regarded as basis vectors of a speaker variability subspace, have been tested, e.g., [12], [50], [51]. Miao *et al.* [39] use an auxiliary DNN to build speaker-specific transforms of the original feature vectors.

Speaker discriminative codes, that capture speaker variabilities in trainable vectors to be used in addition to the conventional feature vectors for DNN, have been proposed in [40]. In practice, the speaker code vector is connected to a large speaker-independent neural network through a separate set of connection weights. These new weights and codes for all speakers in the training set can be jointly learned based on the available training data. Speaker codes often require speaker adaptive (re-)training, owing to the additional connection weights between codes and the hidden units.

III. Connectionist ASR With Deep Models

A. CD-DNN-HMM: Topology

In CD-DNN-HMM systems, the deep model estimates the *a posteriori probability*, $P(q_t|\mathbf{o_t})$, of the q_t state given the speech observation $\mathbf{o_t}$. Next, the Bayes' rule is used to obtain the observation probability, $p(\mathbf{o}_t|q_t) = P(q_t|\mathbf{o}_t)p(\mathbf{o}_t)/P(q_t)$, where $P(q_t)$ is the prior probability of each state estimated from the training set, and $p(\mathbf{o}_t)$ is independent of the word sequence and thus can be ignored.

The input to the deep architecture is typically a splice of a central frame (whose label is that for the splice) and its n context frames on both left and right sides. The hidden non-linear layers are constructed by sigmoid units, and the output layer is a softmax layer. The softmax output predicts the posterior probabilities of thousands of senones. In this work, an $(L+1)$-layer DNN, consisting of L hidden nonlinear layers ($l = 1, ..., L$) and one output layer ($l = L+1$), is used to model the posterior probability of an HMM state given an observation vector. Thus the output at the l-th hidden layer, $\mathbf{h}^l$, can be recursively defined as the nonlinear transformation of the $(l-1)$-th layer, namely:

$$\mathbf{h}^l = \sigma(\mathbf{x}^l) = \sigma(\mathbf{W}^l\mathbf{h}^{(l-1)} + \mathbf{b}^l) \tag{1}$$

where $\mathbf{W}^l$ and $\mathbf{b}^l$ are the weight matrix, and the bias vector for layer l, respectively. $\sigma(\mathbf{x}^l) = 1/(1+e^{-\mathbf{x}^l})$ is an element-wise operation. Moreover, $\mathbf{h}^l$, and $\mathbf{x}^l$ correspond to the activation and excitation of the l-th layer, respectively. Finally, $\mathbf{h}^0$ corresponds to the input observation vector.

B. CD-DNN-HMM: Training

DNN can be trained by maximizing the log posterior probability over the training frames. This is equivalent to minimizing the cross-entropy (CE) objective function. Let $\mathcal{X}$ be the whole training set, which contains T frames, i.e., $\{\mathbf{o}_1, \ldots, \mathbf{o}_T\} \in \mathcal{X}$, then the loss with respect to $\mathcal{X}$ is given by

$$\mathcal{L}^{\mathrm{CE}} = -\sum_{t=1}^{T}\sum_{j=1}^{J} \tilde{P}_t(j)\log P(q_t^j|\mathbf{o}_t), \tag{2}$$

where $P(q_t^j|\mathbf{o}_t)$ is an estimate of the $j\mathrm{th}$ HMM state (i.e., senone state) at time t, q_t^j, and $\tilde{P}_t$ is the target probability of frame t. In practice, the target probability $\tilde{P}_t$ is often obtained by a forced alignment with an existing system resulting in only the target entry that is equal to 1. The objective function is minimized by using error backpropagation, which is a gradient-descent based optimization method developed for artificial neural networks (see [52] for detail). In this work, we adopt mini-batch stochastic gradient descent with a chosen batch size of 256.

In [29], [30], it has been shown that sequence training can significantly boost ASR performance by incorporating acoustic models, lexicon and language models constraints in the objective function. In this work, we therefore adopt sequence training to estimate the DNN parameters of the speaker independent CD-DNN-HMM systems. Specifically, we follow [30], and the minimum Bayes risk, e.g., [53], [54], objective function at a state-level (sMBR) is used:

$$\mathcal{L}^{\mathrm{sMBR}} = -\sum_{u}\frac{\sum_{W} p(\mathbf{O}_u|S_u)^k P(W) A(W, W_u)}{\sum_{W'} p(\mathbf{O}_u|S_u)^k P(W')}, \tag{3}$$

where $O_u = \{o_{u1}, \ldots, o_{uT_u}\}$ is the sequence of observation for the u-th utterance, W_u is the word-sequence in the reference for utterance u, $S_u = \{s_{u1}, \ldots, s_{uT_u}\}$ is the sequence of state corresponding to W_u, and k is the acoustic scaling factor. Finally, $A(W, W_u)$ is the raw number of correct state labels corresponding to the word sequence W with respect to the reference word sequence W_u.

Training deep neural networks from a set of randomly initialised parameters may result in a poor local optimum when performing error backpropagation. To cope with this, pre-training methods have been proposed for a better initialization of the parameters, e.g., [55]. In restricted Boltzmann machine (RBM) based pre-training, the chief idea is to grow the DNN layer by layer without using the label information [55]. Each pair of layers in the DNN is treated as an RBM and is trained using an objective criterion called contrastive divergence [55]. In this work, RBM pre-training is always performed.

IV. Bayesian Learning of Hidden Activation Functions

The underpinning of the proposed MAP adaptation of trainable hidden activation functions is now presented. The first step to accomplish CD-DNN-HMM adaptation through the activation function is to parametrize the sigmoid with a slope, d, and a bias, c, that can be simultaneously learned using the adaptation data. The slope and bias terms are initialized to 1, and 0, respectively, and trained in a speaker adaptive fashion. It has been shown that adding a slope and bias term to each sigmoid accounts for pre-appending an affine linear layer, $\mathbf{\Phi} = \{\mathbf{D}, \mathbf{c}\}$, to each hidden nonlinear layer [48], where $\mathbf{D}$ is a diagonal matrix with the activation slopes, d, on the diagonal, and $\mathbf{c}$ is the vector of bias terms, c. Eq. (1), which represent the output of the l-th hidden layer, would therefore become:

$$\mathbf{h}^l = \sigma(\mathbf{x}^l, \mathbf{\Phi}^l) = \sigma(\mathbf{D}^l \mathbf{x}^{(l)} + \mathbf{c}^l) \tag{4}$$

A. Activation Function Learning

The error backpropagation algorithm can be employed to learn the aforementioned $\mathbf{\Phi} = \{\mathbf{D}, \mathbf{c}\}$ matrix while keeping all of the other CD-DNN-HMM parameters unchanged. The activation function parameters could be estimated by minimizing Eq. (2): For notational simplicity, we could expand the output vector, $\mathbf{h}^l$, in each layer by adding an additional dimension of constant 1 to incorporate the bias vector, $\mathbf{c}$, into the diagonal matrix, $\mathbf{D}$. To avoid introducing new symbols, we refer to this new weight matrix as $\mathbf{\Phi}^l$. Thus backpropagated error, $\boldsymbol{\epsilon}^l$, vector for the generating slope-bias layer is:

$$\boldsymbol{\epsilon}^l = (\mathbf{\Phi}^{l+1})^T \boldsymbol{\epsilon}^{l+1} \circ (\mathbf{h}^l)' \, (l = L, ..., 1), \tag{5}$$

where $\circ$ denotes element-wise multiplication, $(\mathbf{\Phi}^{l+1})^T$ is the transpose of $\mathbf{\Phi}^{l+1}$, and $(\mathbf{h}^l)' = \mathbf{h}^l \circ (1 - \mathbf{h}^l)$ for sigmoid. With the error vectors at certain hidden layers, the gradient over the whole training set with respect to the weight matrix $\mathbf{\Phi}^l$ is given by

$$\frac{\partial \mathcal{L}^{\mathrm{CE}}}{\partial \mathbf{\Phi}^l} = \boldsymbol{E}^l (\mathbf{H}^{l-1})^T. \tag{6}$$

Note that in the above equation, both $\mathbf{H}^{l-1}$ and $\boldsymbol{E}^i$ are matrices, which are formed by concatenating vectors corresponding to all the training frames from frame 1 to T, *e.g.,* $\boldsymbol{\epsilon}^l = [\boldsymbol{\epsilon}_1^l \ldots, \boldsymbol{\epsilon}_t^l, \ldots, \boldsymbol{\epsilon}_T^l]$.

B. MAP Adaptation of Activation Functions

The maximum a posteriori (MAP) adaptation framework can be established by defining a prior distribution over the set of affine transformations, $\mathbf{\Phi}^l$s, holding the slope and bias terms for each hidden neuron.

1) Empirical Bayes Estimation: To analyze and estimate the prior density, we have used the training data. Moreover, we have adopted an *Empirical Bayes Approach*, e.g., [56] and treated each speaker in the training set as a sample speaker. Supervised speaker adaptive training has been performed using the data for each training speaker, separately. After that, we have obtained a set of L matrices, $\mathbf{\Phi}^l$, for each speaker. We have assumed the distribution of the $\mathbf{\Phi}^l$ matrix to be joint Gaussian.

By expressing the generic $\mathbf{\Phi}^l$ as a a vector $\mathbf{w}^l$ with each entry representing a particular slope/bias parameter, we have the prior density in a *multivariate Gaussian distribution* with the following form:

$$p(\mathbf{\Phi}^l) = \frac{1}{(2\pi)^{M/2} |\mathbf{\Sigma}^l|^{1/2}} \exp(-\frac{1}{2} (\mathbf{w}^l - \boldsymbol{\mu}^l)^T \mathbf{\Sigma}^{l^{-1}} (\mathbf{w}^l - \boldsymbol{\mu}^l)) \tag{7}$$

where only the diagonal entries of the covariance matrix $\mathbf{\Sigma}$ are non-zero (from the independence assumption). With S adapted speaker vectors, the maximum likelihood estimation of the mean $\boldsymbol{\mu}^l$ and variance $\mathbf{\Sigma}^l$ can be expressed as:

$$\boldsymbol{\mu}_{ML}^l = \frac{1}{S} \sum_{i=1}^{S} \mathbf{w}_i^l \tag{8}$$

$$\mathbf{\Sigma}_{ML}^l = \frac{1}{S} \sum_{i=1}^{S} (\mathbf{w}_i^l - \boldsymbol{\mu}_{ML}^l)(\mathbf{w}_i^l - \boldsymbol{\mu}_{ML}^l)^T \tag{9}$$

where $\mathbf{w}_i$ is the vector consisting of the adapted transformation weights of speaker i.

2) MAP Formulation: Eq. (10) formulates the MAP learning idea by adding the terms concerning prior densities, $p(\mathbf{\Phi}^l)$, to the plain cross entropy objective function:

$$\mathcal{L}^{\mathrm{MAP}} = -\sum_{l=1}^{L} \lambda^l \log p(\mathbf{\Phi}^l) + \mathcal{L}^{\mathrm{CE}} \tag{10}$$

where λ^l controls the importance of the prior term.

Applying the prior form of Eq. (7), the objective function for MAP adaptation is in the form of Eq. (11).

$$\mathcal{L}^{\mathrm{MAP}} = \sum_{l=1}^{L} \frac{\lambda^l}{2} (\mathbf{w}^l - \boldsymbol{\mu}^l)^T \mathbf{\Sigma}^{l^{-1}} (\mathbf{w}^l - \boldsymbol{\mu}^l) + \mathcal{L}^{\mathrm{CE}} \tag{11}$$

where only the diagonal entries of the covariance matrix $\mathbf{\Sigma}$ are non-zero (from the independence assumption of the weights).

A close look at Eq. (11), when the prior density is a standard Gaussian $N(0|I)$, MAP learning will degenerate to conventional L2-regularized training. The gradient of $\mathcal{L}^{\mathrm{MAP}}$ with respect to $\mathbf{w}^l$ can now be expressed as:

$$\frac{\partial \mathcal{L}^{\mathrm{MAP}}}{\partial \mathbf{w}^l} = \lambda^l (\mathbf{w}^l - \boldsymbol{\mu}^l)^T \, diag(\mathbf{\Sigma}^{l^{-1}}) + \frac{\partial \mathcal{L}^{\mathrm{CE}}}{\partial \mathbf{w}^l}, \tag{12}$$

where $diag(\mathbf{\Sigma}^{-1})$ consists of the diagonal entries of $\mathbf{\Sigma}^{-1}$.

C. Prior Evolution & Online Adaptation

The MAP adaptation method previously discussed implies batch algorithms that requires processing the available adaptation data as a whole. It is often more desirable and more realistic to process the data sequentially. As discussed in [25], the advantage of a sequential algorithm over a batch method is not necessarily in the final performance, but in computational efficiency, reduced storage requirements, and the fact that an outcome may be provided without having to wait for all the data to be processed. In addition, different data segments often correspond to different parameter values, so it is no longer desirable to process all the available adaptation samples, even if we can afford the computational load of the batch algorithm. A sequential algorithm can instead be designed to adaptively track the varying parameters, and that leads to an attractive adaptation scenario, which is known as the online (or incremental, sequential) adaptation.

Sequential methods make use of observations one at a time, or in small batches, and then discard them before the next observations are used. These methods can be used, for example, in real-time learning scenarios, where a steady stream of data is arriving, and predictions must be made before all of the data is seen. Sequential approach to learning arises naturally within the MAP adaptation framework proposed in this paper, and we here present an online adaptation version based on a key concept called *prior evolution* [25]. In addition to evolution in time, priors can also be evolved in space as done in tree-based structural MAP (SMAP) [23].

In a Bayesian framework, the uncertainty of the DNN parameters is taken into account by treating these parameters, namely $\mathbf{w}^l$, as random variables. Thus our prior knowledge about $\mathbf{w}^l$, is assumed to be summarized in a known joint *a priori* pdf $p(\mathbf{w}^l|\phi^{(0)})$ with *hyper-parameters* $\phi^{(0)}$, where $\mathbf{w}^l \in \Omega$, and Ω denotes an admissible region of the hidden activation function parameters. The prior information can be derived from previous experience, e.g., the training data, as discussed in the previous sections. It can also be derived from previous experiences, e.g., training data, $\mathcal{X}$, as we discussed in Section IV-B. Now, let $\mathcal{X}^n = \{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$, be n independent sets of observation samples, which are incrementally obtained and used to update our knowledge about $p(\mathbf{w}^l)$. There exist many ways to *evolve* the prior. The central idea is that the intended evolving prior pdf $p_{\text{intended}}(\mathbf{w}^l)$ summarizes the information inherited from the prior knowledge and learned from the observation data.

Online MAP adaptation can be accomplished as follows: Given a new block of feature vector sequences, the current set of CD-DNN-HMMs is used to recognize this feature vector sequence. After the recognition of the current block of utterances, the prior pdfs for the DNN parameters, which are the results of the previous prior evolution step, are evolved to derive a set of intended posterior distributions, which will be served as the prior for the next round of prior evolution. By taking a MAP estimate from the evolved prior distributions, the hidden activation function parameters are adapted, and the updated models are used to recognize the future input utterance(s). The prior evolution algorithm requires the senone-level transcription of the speech utterances. In this work, such a transcription is derived directly from the recognition results, i.e., unsupervised adaptation.

1) Prior Evolution: The implementation of this learning procedure for incremental CD-GMM-HMM training raises some serious computational difficulties because of the nature of the missing data problem, and a quasi-Bayes learning formulation was proposed in [25]. In this work, we focus on prior evolution of the hidden activation parameters only, which have a multivariate Gaussian distribution, $\mathcal{N}(\cdot)$. Assuming that mean and the precision in the l-th hidden non-linear layer are unknown, then the conjugate prior for the n-th independent set is given by the normal-Wishard distribution, $\mathcal{W}(\cdot)$, [57]:

$$p(\boldsymbol{\mu}_n^l, \boldsymbol{\Lambda}_n^l) = \mathcal{N}(\boldsymbol{\mu}_n^l|\boldsymbol{\mu}_{n-1}^l, (\tau^l W_{n-1}^l)^{-1})\mathcal{W}(\boldsymbol{\Lambda}_n^l|W_{n-1}^l, \nu) \tag{13}$$

where $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ is the precision matrix, $\nu^l < k-1$, and W^l are called the number of degrees of freedom, and the scale matrix of the Wishart distribution (k is the dimension of the precision matrix).

The prior evolution scheme for mean and precision can now be established with ease leveraging the conjugacy properties. The mean prior evolution can be accomplished as follows (see [57] for details):

$$\boldsymbol{\mu}_n^l = \frac{\tau\boldsymbol{\mu}_{n-1}^l + |\mathcal{X}_n|\mathbf{w}^{\text{CE}}}{\tau + |\mathcal{X}_n|} \tag{14}$$

$$\tau = \tau + |\mathcal{X}_n| \tag{15}$$

where $|\mathcal{X}_n|$ is the number of observations used to perform cross-entropy based adaptation of the activation function parameters, and τ indicates the pseudo observations used to estimate $\boldsymbol{\mu}_{n-1}^l$. Eq. (8) can be used to computed $\boldsymbol{\mu}_0^l$.

The precision matrix evolution can be obtained as follows (see [57] for details):

$$W_n^l = \Big((W_{n-1}^l)^{-1} + \boldsymbol{C}_{n-1}^l + \frac{\tau|\mathcal{X}_n|}{\tau + |\mathcal{X}_n|} \times (\mathbf{w}^{\text{CE}} - \boldsymbol{\mu}_n^l)(\mathbf{w}^{\text{CE}} - \boldsymbol{\mu}_n^l)'\Big)^{-1} \tag{16}$$

$$\nu = \nu + |\mathcal{X}_n| \tag{17}$$

where $\mathbf{C}_0^l$ is initialized by multiplying Eq. (9) by S.

2) Online MAP Adaptation: The online MAP (OMAP) formula is similar to the batch one discussed in Section IV-B2 except that the prior evolution effect has to be taken into account. The corresponding objective function is defined as:

$$\mathcal{L}_n^{OMAP} = \sum_{l=1}^{L} \frac{\lambda^l}{2}(\mathbf{w}^l - \boldsymbol{\mu}_{n-1}^l)^T(\boldsymbol{\Sigma}_{n-1}^l)^{-1}(\mathbf{w}^l - \boldsymbol{\mu}_{n-1}^l) + \mathcal{L}^{\text{CE}} \tag{18}$$

The gradient of $\mathcal{L}_n^{OMAP}$ with respect to $\mathbf{w}^l$ can be easily computed.

V. Experiments

A. Experimental Setup and SI Benchmarks

Two CD-DNN-HMM baseline systems were built using the 309-hour Switchboard corpus [28] - a conversational telephone

speech corpus, and the Kaldi toolkit [58]. The key difference between the two baseline ASR systems relied on input speech feature vectors used to train the connectionist component. In the first system, non-adaptive features, namely filter banks, were employed. In particular, the feature vector is a 23-dimension mean-normalized log-filter bank feature with up to second-order derivatives and a context window of 11 frames, forming a vector of 759-dimension (69×11) input. In the second system, the DNN was trained over 40-dimension adaptive feature vector, namely fMLLR, plus a context window of 11 frames, forming a vector of 440-dimension (40×11) input. As previously mentioned, fMLLR features can be generated by training a complete GMM-based system, which is then used to estimate a single input transform per speaker. The transformed feature vectors are then used to train a DNN in a speaker adaptive manner and another set of transforms is estimated (using GMM) during the testing stage for unseen speakers.

Aside from feature parametrization, the connectionist component was initialized with stacked RBMs by using layer-by-layer generative pre-training. An initial learning rate of 0.01 was used to train the Gaussian-Bernoulli RBM, and a learning rate of 0.4 was applied to the Bernoulli-Bernoulli RBMs. The DNN has six hidden layers, each having 2048 sigmoid units, with bias and slope set to zero and one, respectively, during training. The output layer has 8806 softmax units corresponding to tied-parameter context-dependent acoustic states, known as senones. Following common practices, these senone-based target units were inherited from an already-trained CD-GMM-HMM system, which had to be built to address the same task. Senone-state classes were generated using decision tree based state tying and a clustering algorithm based on the maximum likelihood criterion using the statistics collected with Gaussian models. The transition probabilities, a_{ij}, were also borrowed from the CD-GMM-HMM system [1]. The cross-entropy (CE) objective function in Eq. (2) with an initial learning rate of 0.008 was employed to start fine-tuning of the DNN parameters, which is then finalized by three iterations of sMBR sequence training. A 3-gram language model estimated from the Switchboard corpus, was used in decoding.

Experimental results are reported on the NIST 2000 Hub5 evaluation set [59], which covers 80 different speakers. The Hub5 set contains two types of data, Switchboard (SWBD) and CallHome English (CHE). SWBD data match better with the training data, since the SWBD set covers 40 speakers and 1831 utterances. Therefore, the speaker independent models generalize sufficiently well on the SWBD dataset, and the system performance improvement due to speaker adaptation is expected to be lower than that attainable in mismatched conditions. The number of spoken segments differs among test speakers, and varies between a minimum of 25 utterances to a maximum of 67 utterances per speaker (46 utterances per speaker in average). The CallHome data tend to be harder to recognize, mainly because of a greater prevalence of foreign-accented speech.

A single iteration of optimizing the MAP objective function in Eq. (12) with a fixed learning rate of 0.02 was employed during unsupervised adaptation in a batch fashion. On the other hand, unsupervised online adaptation was performed with a fixed learning rate of 0.01, and the prior was evolved as indicated in Eq. (14). The starting value of τ in Eqs. (14) and (15) was set to 1 in all experiments.

TABLE I
WERS (IN %) ON THE SWBD DATA FOR SEVERAL SPEAKER INDEPENDENT (SI) CD-DNN-HMM SYSTEMS USING NON-ADAPTIVE FEATURES. THE TOP PART OF THE TABLE SHOWS RESULTS OBTAINED IN OUR LABORATORY; WHEREAS, WERS AVAILABLE IN THE LITERATURE ON THE SAME TASK AND IN SIMILAR EXPERIMENTAL CONDITIONS ARE REPORTED IN THE LOWER PART OF THE TABLE. IN PARENTHESES, THE PUBLICATION YEAR IS REPORTED

	SYSTEM	WER (in %)
IN-LAB	CE CD-DNN-HMM	16.2%
	sMBR CD-DNN-HMM	15.1%
LITERATURE	CE CD-DNN-HMM (2011) [60]	16.1%
	CE CD-DNN-HMM (2014) [40]	16.2%
	CE CD-DNN-HMM (2016) [10]	15.2%

A fundamental step in our investigation on speaker adaptation of CD-DNN-HMM systems is to ensure that our baseline ASR systems are reliable. In recent years, many independent researchers worked on the NIST 2000 Hub 5 date set, yet most results are reported on the SWBD part only. To ease the comparison with what's available in the literature, we start our study by reporting the percentage of word error rate (WER) on the SWBD portion of the test data for different speaker independent (SI) CD-DNN-HMM systems built in our laboratory with non-adaptive features (i.e., filter-bank features) in the upper part of Table I. By visually inspecting the results, it is easy to verify that sMBR sequence-level training indeed boosts the SI ASR performance, and error rate is meaningfully reduced from an initial 16.2% to 15.1% moving from (frame-level) CE to sMBR training. Therefore, we will use the sMBR-trained CD-DNN-HMM systems to carry out speaker adaptation.

The bottom part of Table I shows the best ASR performance, retrieved from the literature, attained on the SWBD dataset using non-recurrent deep models [10], [40], [60] in experimental conditions similar to ours. The comparison between the WERs obtained in our laboratory, and those available in the literature allows us to duly confirm the reliability of the results to be reported in this study. Assessing the capability of our proposed MAP adaptation techniques is our primary goal, we therefore analyze ASR performances in different adaptation scenarios in the next few sections.

B. Unsupervised Batch Adaptation: Regular Speech Features

First, we evaluate the proposed technique focusing on unsupervised batch adaptation using non-adaptive speech features (e.g., filter bank features). The SWBD dataset was used for both adaptation and testing purposes. MAP adaptation of the activation function parameters (i.e., slope and bias terms), **MAP-AF**, which is accomplished as indicated in Eq. (12), is reported in the fifth row of Table II. We can observe that an already high-accuracy SI sMBR CD-DNN-HMM ASR engine has been further improved from an initial WER of 15.1% down to 14.0%, indicating a relative WER reduction (WERR) of 7.3% after adaptation.

TABLE II
UNSUPERVISED SPEAKER ADAPTATION RESULTS ON THE SWBD TASK FOR SEVERAL TECHNIQUES ARE REPORTED, IN TERMS OF PERCENTAGE OF THE WER. FOR EASY OF COMPARISON, THE FIRST AND THE SECOND COLUMN SHOW THE PERFORMANCE FOR SPEAKER INDEPENDENT (SI) AND SPEAKER ADAPTED (SA) CD-DNN-HMM SYSTEMS, RESPECTIVELY. NON-ADAPTIVE FEATURES HAVE BEEN USED, NAMELY FILTER-BANK BASED SPEECH FEATURES. CONSEQUENTLY, WERS ACROSS THESE DATA SETS ARE EXPECTED TO BE DIFFERENT

SYSTEM	SI WER (in %)	SA WER (in %)
LIN	15.1%	14.9%
KLD-LIN	15.1%	14.8%
AF	15.1%	14.4%
L2-AF	15.1%	14.3%
MAP-AF	15.1%	14.0%
LHUC [10, Table XII]	15.2%	14.7%
SAT-LHUC [10, Table XII]	15.2%	14.6%

The strength of our proposed MAP scheme can be better appreciated by comparing it with other recently-published adaptation solutions. Specifically, the WER attained by applying unsupervised linear hidden network (LIN) speaker adaptation, **LIN**, to the SI ASR systems is reported in the first row, and second column in Table II. In LIN, an affine transformation network is added to the input of the connectionist component. This transformation is estimated during adaptation while keeping all other DNN parameters frozen. LIN delivers a final WER of 14.9%, which corresponded to an relative WERR as small as 1.3%.

It may be argued that the KLD adaptation technique [6] could be combined with unsupervised LIN, **KLD-LIN**, to alleviate the catastrophic forgetting problem and boost the adaptation performance. KLD-LIN reduces the WER from the 15.1% down to 14.8% (see second row, and second column in Table II), and a small improvement over plain LIN was attained, as expected.[1] This result demonstrates that the proposed approach always attains a superior performance to the conventional LIN technique yet uses a more compact representation. The number of parameters to be stored with LIN is 576,840, i.e., 23 times more parameters than those needed with the proposed solution - that uses only 24,576 parameters. The latter outcome makes the proposed approach very appealing in deploying large-scale speech recognition service to the general public.

Next, the regularization capability of the proposed MAP solution can be better understood by performing unsupervised adaptation of the activation function parameters, **AF**, i.e. cross-entropy based training of bias and slope terms holding all remaining DNN parameters unchanged. AF can deliver a WER of 14.4%, as shown in the third row, and second column of Table II. In the fourth row, a WER equal to 14.3% is attained in the degenerate MAP approach, that is, we introduce a L2 regularization doing unsupervised AF adaptation (please refer to Section IV-B2). We can see that L2 regularization slightly enhances AF performance, but it delivers a worse recognition accuracy than the proposed MAP approach. We can therefore conclude that the ASR performance boost gained using the proposed approach is not negligible, and it is not simply due to a regularization effect.

[1]The relative error reduction attained with KLD-based adaptation is 2%, which is in the same range of improvement reported in [6]. The latter apparently confirms the correct implementation of our KLD-based adaptation scheme.

TABLE III
WERS (IN %) W/O ADAPTIVE FEATURES ON THE SWBD TASK. THE TOP PART OF THE TABLE SHOWS RESULTS OBTAINED IN OUR LABORATORY; WHEREAS, LHUC RESULTS IN SIMILAR EXPERIMENTAL CONDITIONS [10] ARE REPORTED IN THE MIDDLE PART OF THE TABLE. IN THE LOWER PART OF THE TABLE, THE RESULTS WITH LSTM, BLST, SFFMN, AND VFFMN ARE SHOWN. THE PLUS +FMLLR INDICATES THAT SPEAKER ADAPTED FEATURES HAVE BEEN USED TO ACCOMPLISH SEQUENTIAL SMBR TRAINING OF THE CONNECTIONIST COMPONENT. FMLLR CAN BE REGARDED AS A FEATURE-SPACE ADAPTIVE TRAINING. THE (++) SYMBOL INDICATES THAT THE FOLLOWING UNSUPERVISED ADAPTATION TECHNIQUES HAS BEEN PERFORMED OVER THE FMLLR-BASED SPEAKER INDEPENDENT HYBRID SYSTEM

	SYSTEM	WER
IN-LAB	SMBR CD-DNN-HMM	15.1%
	+FMLLR	13.8%
	++MAP-UF	13.2%
LHUC (2016) [10]	CE CD-DNN-HMM	15.2%
	+FMLLR (Table XII)	14.2%
	++LHUC (Table XII)	14.2%
	++SAT-LHUC (Table XII)	14.1%
LITERATURE (2016)	LSTM [31]	14.2%
	BLSTM [31]	13.5%
	SFFMN [31]	14.2%
	VFFMN [31]	13.2%

We have already mentioned that other authors have proposed different solutions to adapt the activation function shape after Siniscalchi *et al.* [13]. Among these methods, LHUC [10] is the most promising. We therefore find it instructive to report the LHUC performance attained in the similar experimental conditions, and on the same ASR task in the lower part of Table II. The WER attained with LHUC, and its improved speaker adaptive training (SAT) version, SAT+LHUC, are given in the fifth and sixth row, respectively. By comparing results shown in the fourth, fifth, and sixth rows, and first and second columns, we can confirm that the proposed MAP-AF approach outperforms both LHUC and SAT+LHUC.

C. Unsupervised Batch Adaptation: Speaker Adaptive Features

The results shown in Table II demonstrate the feasibility of the proposed MAP solution to the unsupervised, batch speaker adaptation problem of connectionist ASR system employing deep architectures. Nonetheless, it could be argued that baseline SI ASR engine with better seed models can be designed with either recurrent deep models, e.g., [34], [35], or feed-forward deep models equipped with memory, FFMN, [31]. In the lower part of Table III, the experimental results reported in [31, Table 2] of four systems utilizing long-term dependency of the speech signal, namely LSTM [35], bidirectional LSTM (BLSTM) [34], and scalar feed-forward memory networks (sFFMN) [31], and vectorized feed forward memory networks (vFFMN) [31] are shown. All systems were trained using filter-bank based features. WERs between 13.2% and 14.2% were attained using those systems.

An obvious questions that may arise is whether the proposed MAP-AF can further improve the state-of-the-art system

TABLE IV
UNSUPERVISED BATCH SPEAKER ADAPTATION RESULTS ON HUB5

	WER (in %)		
SYSTEM	SWBD	CallHome	TOTAL
sMBR CD-DNN-HMM	15.1%	28.7%	21.9%
+MAP-AF	14.0%	26.1%	20.1%
+fMLLR	13.8%	25.1%	19.5%
++MAP-AF	13.2%	24.0%	18.6%
CE CD-DNN-HMM [10]	15.2%	28.2%	21.7%
+fMLLR [10]	14.2%	26.2%	20.2%
++SAT-LHUC [10]	14.1%	25.6%	19.9%

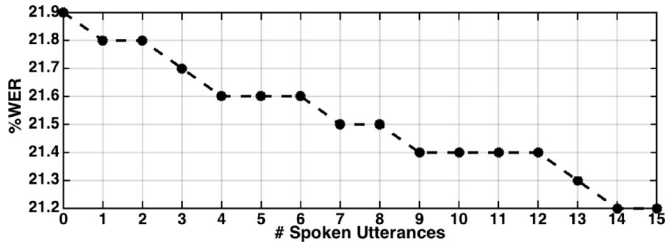

Fig. 1. Self-adaptation results, in terms of the %WER, as the number of adaptation utterances increases from 1 to 15. WERs are always given to the whole 15 utterances to make clear the effect of self-adaptation on the system performance. The SI ASR performance corresponds to that obtained with 0 adaptation utterances.

performances. In this section, we provide a sound attempt to address such a challenging question. Unfortunately, we do not have access to the ASR architectures utilizing either recurrent or memory blocks, thereby we cannot directly address such a question. Nevertheless, we are mainly interested in verifying whether a performance improvement can be observed using better seed models. To this end, we have built a speaker independent CD-DNN-HMM system using speaker adaptive features, namely fMLLR, and sequential sMBR training. The speaker compensated features allows us to reduce the WER from the initial 15.1% down to 13.8%, as indicated in the second row of Table III. By applying MAP-AF to this new system, the WER is further brought down to 13.2%, as shown in the third row of Table III. We can therefore conclude that MAP adaptation of the hidden activation function parameters could indeed effectively enhance state-of-the-art ASR systems. That is, our solution has the potential to boost recurrent and memory equipped deep models, since it works with a fMLLR-based ASR system that reported a recognition error between the BLSTM-based and vFFMN-based systems.

As a final remark, it may be instructive to remind that fMLLR adaptive features make the experimental setup more challenging because the proposed technique has to be applied to an already speaker adapted CD-DNN-HMM system - fMLLR can be thought of a feature-space speaker adaptation technique as seen in Section II-A. On top of that, the experimental setup is further exacerbated by the intrinsic characteristics of the data: (i) the SWBD data are narrow-band, containing less information for discrimination between speakers, as discussed in [10], especially when estimating relevant statistics from small amounts of unsupervised adaptation data, and (ii) the SWBD data set exhibits a large overlap between training and test speakers - 36 out of 40 test speakers are observed in training, which allows learning more accurate speaker characteristics during supervised as opposed to unsupervised speaker adaptive training, as argued in [10]. Therefore, there is a very small room for improvement, and the fact that MAP-AF already brings the WER down to 13.2% from the initial 13.8%, which corresponds to a relative WERR equal to 4.4%, indeed confirms the viability of the proposed technique. For the sake of comparison, LHUC results are reported in the lower part of Table III.

The set of results reported in Table IV completes our investigation into unsupervised, batch MAP speaker adaptation. In particular, the results on the CallHome subset, and on the full 2000 Hub5 benchmark are shown. The proposed MAP speaker adaptation technique performs relatively better under more mismatched conditions, namely the CallHome subset of the Hub5 2000 benchmark. MAP-AF also gives a consistent reduction from fMLLR, and the overall WER was reduced from 19.5% to 18.6%, as shown in the 3rd and 4th rows, respectively - a 4.6% relative WERR. In summary, we can attain up to a 15% relative WERR from the SI systems by combining fMLLR and MAP-AF.

D. Unsupervised Self-Adaptation

Next we investigate the *self-adaptation* properties of our proposed unsupervised MAP adaptation approach. To this end, we select for each speaker 15 spoken utterances for a total of 1200 utterances. The SI CD-DNN-HMM system, trained on filter bank features, attains an overall WER of 21.9% on the entire test set, as shown in Fig. 1 at 0 spoken utterances. For each speaker, we perform self-adaptation using a single utterance per time, as follows: after seeing one spoken utterance, we adapt the acoustic models with MAP and test on all utterances in order to properly compare results and isolate the effect of self-adaptation. This procedure repeats until we have seen all of the 15 utterances from each speaker. In Fig. 1, we display self-adaptation performance in terms of the number of adaptation utterances. As expected, adaptation consistently improves the acoustic models as the number of available adaptation utterances increases, which is a desirable property of any speaker adaptation scheme. Next, we can see that negative transfer learning is avoided, since a drop in the performance was avoided even with only a few adaptation sentences, e.g., 1, or 2.

E. Unsupervised Online Iterative Adaptation

Any learning framework becomes more useful for practical situations if it is performed in a sequential manner. The set of experiments reported below are meant to demonstrate that, leveraging upon the intrinsic recursive nature of Bayesian learning, online speaker adaptation can be successfully accomplished.

In real application scenarios, we can often acquire only one or a few utterances per speaker, and then the corresponding transcriptions have to be delivered to the end-user. A sequential iterative algorithm suits such a scenario better than adaptation with a large-sized batch. A key advantage of the proposed Bayesian

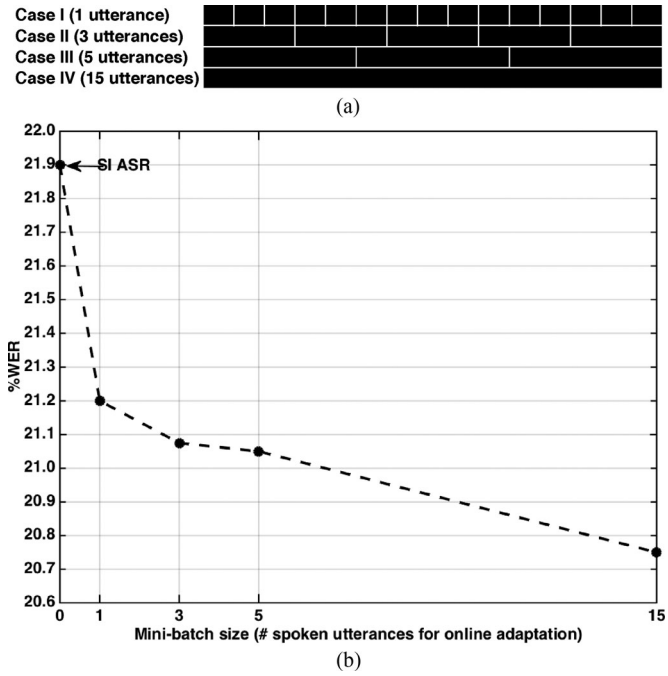

Fig. 2. Online adaptation with a varying mini-batch size. Results are averaged over all NIST 2000 Hub 5 speakers and spoken utterances.

framework is that it is truly adaptive in nature and suitable for performing iterative learning. This unique adaptive characteristic is achieved by incrementally evolving the hyper-parameters, as shown in Eqs. (14) and (15), while adapting the activation function parameters. Before delving into the experimental results, we should point out that experiments with fMLLR features would require to generate an affine transformation for each single iteration, and for each single speaker. That would make the experimental setup very cumbersome; therefore, we report results using only non-adaptive speech feature, i.e., filter bank features.

Fig. 2(b) shows that online incremental MAP adaptation brings a meaningful improvement even considering a single utterance at a time, and the WER is reduced from the initial 21.9% down to 21.2%. Next, it can be meaningful to characterize the online MAP adaptation scheme in terms of the mini-batch size, i.e., number of spoken utterances processed sequentially. Fig. 2(a) shows that four different mini-batch sizes have been considered, namely $\{1, 3, 5, 15\}$; furthermore, these mini-batch have been designed so that the data are used evenly across all min-batch sizes, i.e., there is an overlap among mini-batches, and that guarantees that changes in the ASR performance are imputable solely to the mini-batch size. In Fig. 2(b), we report WERs averaged over all NIST 2000 Hub 5 speakers and spoken segments for online incremental adaptation with a varying mini-batch size. Mini-batch size 0 implies no adaptation, and the SI ASR performance is given. We can observe that online MAP adaptation attains a better recognition accuracy as the mini-batch size increases, and a final WER of 20.7% is delivered with a mini-batch size of only 15 spoken utterances. The latter result confirms the viability of the proposed approach.

VI. Conclusion

In this paper, we have presented a theoretical framework of maximum *a posteriori* adaptation of the hidden activation function parameters in CD-DNN-HMMs. In the experimental part of this study, adaptation is performed in an unsupervised manner, i.e., the true transcriptions of the adaptation data are assumed unknown, since that scenario is more realistic in real applications. To examine the viability of the proposed Bayesian framework, MAP adaptation is applied to a batch speaker adaptation application using the NIST 2000 Hub5 benchmark. In a series of comparative experiments, we study the the effects of different speech features, namely non-adaptive, and speaker adaptive, reporting improvements in all tested scenarios. Moreover, the effectiveness of the proposed approach was demonstrated using high-accuracy speaker independent deep models built with discriminative sequential training. The experimental results also confirmed the feasibility of the proposed techniques. Leveraging on the intrinsic recursive Bayesian nature of the proposed technique, we have also proposed an online incremental approach to unsupervised, sequential speaker adaptation by simultaneously updating the hyperparameters of the approximate posterior densities and DNN parameters on a per utterance basis. Finally, self-adaptation properties of our proposed solution have also been successfully tested. In conclusion, our experimental results indeed confirm the viability of the proposed MAP adaptation framework of hidden activation function parameters in deep models. In future studies, we intend to expand the parameter set to include other deep model parameters.

References

[1] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[2] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. Norwell, MA, USA: Kluwer, 1994.

[3] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[4] C.-H. Lee, "On stochastic feature and model compensation approaches to robust speech recognition," *Speech Commun.*, vol. 25, no. 1–3, pp. 29–47, 1998.

[5] S. M. Siniscalchi, D.-C. Lyu, T. Svendsen, and C.-H. Lee, "Experiments on cross-language attribute detection and phone recognition with minimal target-specific training data," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 875–887, Mar. 2012.

[6] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7893–7897.

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

[8] R. Price, K.-I. Iso, and K. Shinoda, "Speaker adaptation of deep neural networks using a hierarchy of output layers," in *Proc. Spoken Lang. Technol. Workshop*, 2014, pp. 153–158.

[9] Z. Huang, J. Li, S. M. Siniscalchi, I.-F. Chen, J. Wu, and C.-H. Lee, "Rapid adaptation for deep neural networks through multi-task learning," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 3625–3629.

[10] P. Swietojanski, J. Li, and S. Renals, "Learning hidden unit contributions for unsupervised acoustic model adaptation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 8, pp. 1450–1463, Aug. 2016.

[11] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. IEEE Workshop Automat. Speech Recogn. Understanding*, 2011, pp. 24–29.

[12] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proc. IEEE Workshop Autom. Speech Recogn. Understanding*, 2013, pp. 55–59.

[13] S. M. Siniscalchi, J. Li, and C.-H. Lee, "Hermitian polynomial for speaker adaptation of connectionist speech recognition systems," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2152–2161, Oct. 2013.

[14] J. Li, J.-T. Huang, and Y. Gong, "Factorized adaptation for deep neural network," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 5537–5541.

[15] C. Wu and M. Gales, "Multi-basis adaptive neural network for rapid adaptation in speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 4315–4319.

[16] U. Remes, A. R. López, and D. Palomäki, "Bounded conditional mean imputation with observation uncertainties and acoustic model adaptation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 7, pp. 1198–1208, 2015.

[17] L. Samarakoon and K. C. Sim, "Factorized hidden layer adaptation for deep neural network based acoustic modeling," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 12, pp. 2241–2250, Dec. 2016.

[18] P. Swietojanski and S. Renals, "Differentiable pooling for unsupervised acoustic model adaptation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 10, pp. 1773–1784, Oct. 2016.

[19] M.-Y. Hwang and X. Huang, "Shared-distribution hidden Markov models for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 4, pp. 414–420, Oct. 1993.

[20] R. M. French, "Catastrophic forgetting in connectionist networks: Causes, consequences and solutions," *Trends Cogn. Sci.*, vol. 3, 1994, pp. 128–135.

[21] R. Gemello, F. Mana, S. Scanzio, P. Laface, and R. D. Mori, "Linear hidden transformations for adaptation of hybrid ANN/HMM models," *Speech Commun.*, vol. 49, no. 10, pp. 827–835, 2007.

[22] J.-L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[23] K. Shinoda and C.-H. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 276–287, Mar. 2001.

[24] Z. Huang, S. M. Siniscalchi, I.-F. Chen, J. Li, J. Wu, and C.-H. Lee, "Maximum *a posteriori* adaptation of network parameters in deep models," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 1076–1080.

[25] C.-H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," *Proc. IEEE*, vol. 88, no. 8, pp. 1241–1269, Aug. 2000.

[26] L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 5, pp. 1060–1089, May 2013.

[27] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proc. 33rd Annu. Meeting Assoc. Comput. Linguistics*, 1995, pp. 198–196.

[28] J. J. Godfrey and E. Holliman, "Switchboard-1 release 2," Philadelphia, PA, USA: Linguistic Data Consortium, 1997.

[29] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2009, pp. 3761–3764.

[30] K. Veseỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2013, pp. 2345–2349.

[31] S. Zhang, H. Jiang, S. Xiong, S. Wei and L. Dai, "Compact Feedforward Sequential Memory Networks for Large Vocabulary Continuous Speech Recognition," in Proc. Interspeech, Sep. 2016, pp. 3389–3393.

[32] S. Hochreiter and J. Schmidhuber, "Long short- term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, 1997.

[33] A. Graves, A. Mohamed, and G. E. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 6645–6649.

[34] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 338–342.

[35] T. N. Sainath, A. M. B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 8614–8618.

[36] V. V. Digalakis, D. Rtischev, and L. G. Neumeye, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 357–366, Sep. 1995.

[37] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Comput., Speech, Lang.*, vol. 12, pp. 75–98, 1998.

[38] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[39] Y. Miao, H. Zhang, and F. Metze, "Speaker adaptive training of deep neural network acoustic models using i-vectors," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 11, pp. 1938–1949, Nov. 2015.

[40] 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 Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1713–1725, Dec. 2014.

[41] A. Mohamed, T. Sainath, G. D. B. Ramabhadran, G. E. Hinton, and M. Picheny, "Deep belief networks using discriminative features for phone recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 5060–5063.

[42] J. Neto *et al.*, "Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system," in *Proc. Eur. Conf. Speech Commun. Technol.*, 1995, pp. 2171–2174.

[43] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *Proc. Spoken Lang. Technol. Workshop*, 2012, pp. 366–369.

[44] B. Li and K. C. Sim, "Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 526–529.

[45] S. Xue, H. Jiang, and L. Dai, "Speaker adaptation of hybrid NN/HMM model for speech recognition based on singular value decomposition," in *Proc. 9th Int. Symp. Chin. Spoken Lang. Process.*, 2014, pp. 1–5.

[46] J. Xue, J. Li, D. Yu, M. Seltzer, and Y. Gong, "Singular value decomposition based low-footprint speaker adaptation and personalization for deep neural network," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 6359–6363.

[47] J. Xue, J. Li, and Y. Gong, "Restructuring of deep neural network acoustic models with singular value decomposition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2013, pp. 2365–2369.

[48] Y. Zhao, J. Li, J. Xue, and Y. Gong, "Investigating online low-footprint speaker adaptation using generalized linear regression and click-through data," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 4310–4314.

[49] C. Zhang and P. C. Woodland, "DNN speaker adaptation using parameterised sigmoid and ReLU hidden activation functions," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 5300–5304.

[50] M. Karafiát, L. Burget, P. Matějka, O. Glembek, and J. Cernocky, "iVector-based discriminative adaptation for automatic speech recognition," in *Proc. IEEE Workshop Automat. Speech Recogn. Understanding*, 2011, pp. 152–157.

[51] P. Karanasou, Y. Wang, M. Gales, and P. Woodland, "Adaptation of deep neural network acoustic models using factorised i-vectors," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 2180–2184.

[52] S. Haykin, *Neural Networks: A Comprehensive Foundation*. New York, NY, USA: Macmillan, 1994.

[53] J. Kaiser, B. Horvat, and Z. Kačič, "A novel loss function for the overall risk criterion based discriminative training of HMM models," in *Proc. 6th Int. Conf. Spoken Lang. Process.*, 2000, pp. 887–890.

[54] M. Gibson and T. Hain, "Hypothesis spaces for minimum Bayes risk training in large vocabulary speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2006, pp. 2406–2409.

[55] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504–507, 2006.

[56] H. Robbins, "The empirical Bayes approach to statistical decision problems," *Ann. Math. Statist.*, vol. 35, no. 1, 1964.

[57] M. H. DeGroot, *Optimal Statistical Decisions*. New York, NY, USA: McGraw-Hill, 1970.

[58] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *Proc. IEEE Workshop Automat. Speech Recogn. Understanding*, 2011. [Online]. Availabel: http://kaldi-asr.org/doc/about.html

[59] J. Fiscus, W. M. Fisher, A. F. Martin, M. A. Przybocki, and D. S. Pallett, "2000 NIST evaluation of conversational speech recognition over the telephone: English and Mandarin performance results," in *Proc. Speech Transcription Workshop*, 2000, pp. 1–5.

[60] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 437–440.

Zhen Huang received the B.S. degree from Southeast University, Nanjing, China, in 2009, and the dual M.S. degree from Shanghai JiaoTong University, Shanghai, China, and Georgia Institute of Technology, Atlanta, GA, USA, in 2012, all in electrical and computer engineering. He is currently working toward the Ph.D. degree at Georgia Institute of Technology under guidance of Prof. Chin-Hui Lee. His research interests include the areas of speech recognition, deep learning, general machine learning, multimedia information retrieval, and image processing. His current research interests include deep learning based speech recognition and adaptation.

Sabato Marco Siniscalchi received the Laurea and Doctorate degrees in computer engineering from the University of Palermo, Palermo, Italy, in 2001 and 2006, respectively. In 2001, he was with STMicroelectronics, where he designed optimization algorithms for processing digital image sequences on very long instruction word architectures. In 2002, he was an Adjunct Professor with the University of Palermo and taught several undergraduate courses for computer and telecommunication engineering. In 2006, he was a Postdoctoral Fellow at the Center for Signal and Image Processing, Georgia Institute of Technology, Atlanta, GA, USA, under the guidance of Prof. C.-H. Lee. From 2007 to 2009, he joined the Norwegian University of Science and Technology, Trondheim, Norway, and as a Research Scientist in the Department of Electronics and Telecommunications under the guidance of Prof. T. Svendsen. From 2010 to 2015, he was an Assistant Professor with the University of Enna "Kore," Enna, Italy. He is currently an Associate Professor with the University of Enna "Kore" and affiliated with the Georgia Institute of Technology. His main research interests include speech processing, in particular automatic speech and speaker recognition, and language identification. He is currently an Associate Editor in the IEEE/ACMTRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING.

Chin-Hui Lee (F'97) is currently a Professor in the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. Before joining academia in 2001, he had 20 years of industrial experience ending in Bell Laboratories, Murray Hill, NJ, USA, as a distinguished member of Technical Staff, and the Director of the Dialogue Systems Research Department. He has published more than 450 papers, and 30 patents, and was highly cited close to 30 000 times for his original contributions with an h-index of 65 on Google Scholar. He received numerous awards, including the Bell Labs President's Gold Award in 1998, the IEEE Signal Processing Society's 2006 Technical Achievement Award for "Exceptional Contributions to the Field of Automatic Speech Recognition," and the International Speech Communication Association Medal in 2012 in scientific achievement for "pioneering and seminal contributions to the principles and practice of automatic speech and speaker recognition." In 2012, he was invited by International Conference on Acoustics, Speech, and Signal Processing to give a plenary talk on the future of speech recognition. He is a Fellow of International Science Congress Association.