Embedding-Based Speaker Adaptive Training of Deep Neural Networks

Xiaodong Cui, Vaibhava Goel, George Saon

IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA
{cuix,vgoel,gsaon}@us.ibm.com

Abstract
An embedding-based speaker adaptive training (SAT) approach is proposed and investigated in this paper for deep neural network acoustic modeling. In this approach, speaker embedding vectors, which are a constant given a particular speaker, are mapped through a control network to layer-dependent element-wise affine transformations to canonicalize the internal feature representations at the output of hidden layers of a main network. The control network for generating the speaker-dependent mappings are jointly estimated with the main network for the overall speaker adaptive acoustic modeling. Experiments on large vocabulary continuous speech recognition (LVCSR) tasks show that the proposed SAT scheme can yield superior performance over the widely-used speaker-aware training using i-vectors with speaker-adapted input features.

Index Terms: automatic speech recognition, deep neural networks, adaptive training, speaker embedding, LVCSR

1. Introduction
Pattern variation caused by speaker variability is one of the fundamental issues in acoustic modeling for automatic speech recognition (ASR). Such variation can give rise to covariate shift [1] that degrades the recognition performance. Speaker adaptation [2][3] and speaker adaptive training (SAT) [4][5] have been a common practice in ASR to reduce the covariate shift incurred by speaker variability.

In deep neural network (DNN) acoustic modeling, given the large number of parameters in the network and its lack of a generative structure for parameter tying, speaker adaption is typically carried out on a selected subset of parameters under the constraint of data sparsity. For instance, a common way of adaptation is to retrain or fine-tune the input or output layer of the network for a new speaker at the test time. Or, alternatively, the input or output layer is adapted by an affine transformation estimated using the adaptation data [6][7][8]. To reduce the number of parameters to be adapted given the sparse adaptation data, the output layer is further factorized in [9]. In another effort under the name of learning hidden unit contributions (LHUC), only the amplitudes of the outputs of hidden units are re-scaled with the adaptation data [10].

Speaker-aware training is also widely used for adaptation of DNN acoustic models, notably with i-vectors. In [11], an i-vector is estimated for each speaker and appended to all the input features from the same speaker, which is performed for both training and test speakers. DNNs learned from the features with such speaker indicators are supposed to be more speaker invariant.

SAT as another treatment of speaker variability, which is distinct from the above-mentioned adaptation techniques, aims at creating speaker invariance by explicitly introducing a speaker-dependent transformation in the feature space for canonicalization. SAT features have been a popular choice for both GMM-HMM and DNN-HMM acoustic modeling in some of the state-of-the-art LVCSR systems [12]. In [13], a SAT-LHUC setup is introduced where the speaker-dependent amplitude functions are directly integrated into the objective function and jointly trained with speaker independent representations.

Given the increasing complexity of DNNs, which imposes a severe problem of data sparsity, and also intrinsically lack of a generative structure to explore, canonicalizing the feature space to remove the speaker variability during training appears to be appealing. In this paper we focus on SAT of DNNs. Particularly, we propose an embedding-based SAT scheme for speaker invariance in DNN acoustic modeling. Speaker embedding maps a speaker with his/her speech data to a low-dimensional continuous vector with a fixed length. In that sense, i-vector [14] is one way of embedding speakers. It renders a compact representation of the speaker’s acoustic characteristics. We map i-vectors to element-wise affine transformations through a control network. Since each speaker is embedded into a constant i-vector, the resulting affine transformations are speaker dependent. The affine transformations are applied to the outputs of the hidden units of selected layers of a main network for speaker normalization. The control and main networks are then jointly optimized to create canonical internal feature spaces. Experiments conducted on Babel and Switchboard LVCSR tasks show that this approach can improve over i-vector-based speaker-aware training with speaker-adapted features.

The remainder of the paper is organized as follows. Section 2 gives the formulation of the proposed SAT approach. Section 3 provides the experimental results on Babel and Switchboard data sets followed by a discussion in Section 4. Section 5 concludes the paper with a summary and speculation on the future work.

2. Embedding-based SAT
SAT has been a very effective technique in ASR, especially in GMM-HMM acoustic modeling [4][5]. It is also commonly used in DNN-HMM acoustic modeling for speaker-adapted features [12] via FMLLR[15]. It creates a canonical feature space in terms of speaker variability. Since neural network acoustic models with a deep structure render a hierarchical feature representation, it is intriguing to see if such canonical internal feature spaces between layers are also helpful.

Suppose there are $n_{l}$ hidden units in layer $l$ whose outputs form an $n_{l}$-dimensional internal feature space. Suppose features $x^{(s)}$ are features from speaker $s$ in this space, $x^{(s)} \in \mathbb{R}^{n_{l}}$. We define a speaker-dependent affine transformation on the features from speaker $s$:

$$x^{(s)} = A_{l}^{(s)} x^{(s)} + b_{l}^{(s)}$$

(1)

where $A_{l}^{(s)}$ is a diagonal matrix

$$A_{l}^{(s)} = \text{diag}(a_{l,1}^{(s)}, \cdots, a_{l,n_{l}}^{(s)})$$

(2)
In other words, the affine transformation is element-wise.
Define the following two vectors
\[ a_l^{(s)} = [a_{l,1}^{(s)}, \ldots, a_{l,n_l}^{(s)}] \] (3)
\[ b_l^{(s)} = [b_{l,1}^{(s)}, \ldots, b_{l,n_l}^{(s)}] \] (4)
The affine transformation can be written as
\[ \hat{x}^{(s)} = a_l^{(s)} \odot x^{(s)} + b_l^{(s)} \] (5)
where \( \odot \) stands for element-wise multiplication and \( \oplus \) for element-wise addition.
Furthermore, we define a mapping between the embedding vector of speaker \( s \), \( e^{(s)} \), and the corresponding affine transformation
\[ a_l^{(s)} = f_{la}(e^{(s)}) \] (6)
\[ b_l^{(s)} = f_{lb}(e^{(s)}) \] (7)
where the mappings \( f_{la} \) and \( f_{lb} \) are realized by neural networks with \( e^{(s)} \) as input.
Fig.1 illustrates a realization of the proposed SAT scheme. The speaker embedding vectors are used as input to the network on the right which is referred to as the control network. The control network has multiple outputs which are the scaling vector \( a_l \) and bias vector \( b_l \) of the element-wise affine transformation applied to a selected number of layers \( l \) of a generic network on the left referred to as the main network. The resulting multiple transformations or the scaling and bias vectors of the same transformation share the bottom layers of the control network and only split up at the last layer, denoted scaling and bias weight layers in the figure. The scaling and bias weight layers have the same number of hidden units as that of the layer to be normalized in the main network. Let \( \hat{e} \) denote the output of the topmost layer of the shared bottom sub-network before splitting, the outputs of the scaling and bias layers for layer \( l \) are given by
\[ a_l = \sigma(W_{la} \hat{e} + b_{la}) \] (8)
\[ b_l = \tanh(W_{lb} \hat{e} + b_{lb}) \] (9)
where \( W_{la}, b_{la} \) and \( W_{lb}, b_{lb} \) are the weights and biases in the scaling and bias layers respectively and the nonlinearity for the scaling layer is sigmoid while hyperbolic tangent for the bias layer. The sigmoid nonlinearity enforces a positive scaling while the hyperbolic tangent nonlinearity allows the bias to be both positive and negative. The configuration of the control network has been empirically evaluated and determined.
The main network represents a feedforward neural network (FNN), convolutional neural network (CNN) or recurrent neural network (RNN) used for acoustic modeling in a general sense. A number of its layers are selected to get normalized by speaker-dependent transformations for canonical internal feature spaces, which is equivalent to inserting a SAT layer right after one selected layer. In training the control and main networks are treated as one single network. The SAT transformations and the acoustic model with canonical internal feature spaces are therefore jointly optimized.

3. Experimental results
Experiments are carried out on Babel and Switchboard data sets to evaluate the proposed SAT scheme.

![Diagram of SAT scheme](image)

**Figure 1**: An illustration of the realization of the embedding-based speaker adaptive training with a control network (right) and a main network (left). The control network maps a speaker embedding vector to layer-dependent element-wise affine transformations. The main and control networks are jointed trained to simultaneously estimate the speaker-specific transformations and acoustic models with canonical internal feature spaces. In the figure, \( \odot \) stands for element-wise multiplication while \( \oplus \) for element-wise addition.

### 3.1 Babel
The conversational speech data in Swahili and Georgian full language pack (FLP) under the IARPA Babel program [16] are used for training. They are the surprise languages in the Babel option period 2 and 3 evaluations respectively. The training set for each of the two languages consists of 40 hours of speech data. There are 525 training speakers in Swahili and 518 in Georgian. Word error rates (WERs) are measured on the development sets of the two languages. There are 20 hours of audio data from 142 speakers in Swahili and 20 hours from 124 speaker in Georgian.

Table 1 shows a variety of DNN models and their performance on the two languages. The first row presents a DNN model whose input is composed of 9 frames of 40-dimensional speaker-adapted FMLLR features, denoted “fmllr” in the table. The dimensionality of input is 360. The DNN consists of 5 hidden layers. Each hidden layer has 1,024 hidden units. The bottom three hidden layers use ReLU activation functions while the top two hidden lay-
ers use sigmoid activation functions. The softmax output layer has 3,000 units corresponding to context-dependent (CD) HMM states.

The DNN in the second row has the same network configuration except that a 100-dimensional i-vector is appended to the previous 360-dimensional input features, denoted “fmllr+ivec” in the table. So its input dimensionality of 460.

The embedding-based SAT training improves WERs on top of speaker-aware training with FMLLR features in Georgian. The embedding-based SAT training improves WERs on top of speaker-aware training with FMLLR features (denoted fmllr). The model in the second row is speaker-aware training by appending i-vectors to the fmllr features. The models in the last two rows are SAT models using fmllr+ivec features.

As can be observed from the table, the speaker-aware training by appending i-vectors to FMLLR features (fmllr+ivec) gives 0.3% absolute improvement on Swahili but does not help in Georgian. The embedding-based SAT training improves WERs on top of speaker-aware training with FMLLR features on both languages (50.4% → 49.9% in Swahili and 53.8% → 52.7% in Georgian).

In terms of SAT training, gating with only scaling the hidden unit outputs in the main network gives 0.2% absolute improvement on Swahili and 0.7% on Georgian. Furthermore, a full affine transformation with both scaling and bias compensation appears to be more helpful in this case. Specifically, a full SAT transformation obtains another 0.3% absolute improvement over gating on Swahili and 0.4% on Georgian.

3.2. Switchboard 300

We also evaluated the embedding-based SAT on the 300-hour Switchboard data set. The training set consists of 262 hours of Switchboard 1 audio with transcripts provided by the Mississippi State University. The test set is the Hub5 2000 evaluation set composed of two parts: One is the 2.1-hour switchboard (SWB) data from 40 speakers and the other is the 1.6-hour call-

home (CH) data from 40 speakers.

Acoustic models are long short-term memory (LSTM) networks. We evaluated LSTMs up to 5 bi-directional layers in the main network. Each layer contains 1,024 cells with 512 on each direction. On top of the LSTM layers, there is a linear bottleneck layer with 256 hidden units followed by a softmax output layer with 32K units corresponding to CD HMM states. The LSTMs are unrolled 21 frames and trained with non-overlapping feature subsequences of that length. In SAT LSTM, the scaling and bias weight layers at the end of the control network have 1,024 hidden units which are equal in number to that of the internal state of the bi-directional layer. The numbers of hidden units in the shared layers are configured to be [100, 512, 1024] with 100 being the input i-vector dimensionality. The FMLLR features have a dimensionality of 40.

Tables 2 shows LSTMs trained under the CE criterion with randomly initialized weights for 25 epochs. Again the control and main networks are jointly optimized in the SAT training. Mini-batch based SGD is used for optimizing with a batch size of 128. Each batch contains subsequences of 21 frames from different utterances. A 4-gram LM is used for decoding with a vocabulary size of 31K.

Table 2 shows the performance of LSTMs with different settings on SWB and CH respectively. As can be observed from the table, increasing the number of LSTM layers results in better performance until plateaued after 4 layers for LSTMs without SAT. Speaker-aware training by appending i-vectors to FMLLR features can significantly improve the performance. In a 4-layer LSTM, it reduces the WER by 0.7% (11.4% → 10.7%) absolute for SWB and 1.0% for CH (19.7% → 18.7%). The embedding-based SAT training can further improve the performance on top of the i-vector-based speaker-aware training. It can also improve performance for LSTMs with 5 layers. In this case, inserting two SAT layers in a 5-layer LSTM obtained 10.3% WER in SWB and 18.4% WER in CH compared to the speaker-aware training’s 10.7% and 18.8% in SWB and CH, respectively.

Table 1: WERs(%) on Swahili and Georgian with and without embedding-based SAT under the CE criterion. The model in the first row is a baseline DNN with speaker-adapted FMLLR features (denoted fmllr). The model in the second row is speaker-aware training by appending i-vectors to the fmllr features (denoted fmllr+ivec). The models in the last two rows are SAT models using fmllr+ivec features.

<table>
<thead>
<tr>
<th></th>
<th>Swahili</th>
<th>Georgian</th>
</tr>
</thead>
<tbody>
<tr>
<td>fmllr</td>
<td>50.7</td>
<td>53.7</td>
</tr>
<tr>
<td>fmllr+ivec</td>
<td>50.4</td>
<td>53.8</td>
</tr>
<tr>
<td>fmllr+ivec+gating</td>
<td>50.2</td>
<td>53.1</td>
</tr>
<tr>
<td>fmllr+ivec+sat</td>
<td>49.9</td>
<td>52.7</td>
</tr>
</tbody>
</table>
In this paper we proposed an embedding-based speaker adaptive training approach for acoustic modeling using deep neural networks. In this approach, i-vectors are mapped to layer-dependent element-wise affine transformations for internal feature space canonicalization. The mapping is realized through a control network and it is jointly trained with a main network the outputs of whose selected hidden layers are normalized. The approach is evaluated on Babel and Switchboard data sets. It shows that this approach can improve performance over DNNs with speaker-adapted input features and also speaker-aware training using i-vectors.

Looking forward, we will investigate the optimization of the SAT models under the sequence training criterion. Besides, alternative ways of speaker embedding other than i-vectors will be examined.

6. References


4. Related Work and Discussion

The embedding-based SAT training maps speaker embedding vectors (i-vectors) to speaker-dependent affine transformations to normalize internal feature spaces of a DNN acoustic model. In [17], i-vectors are mapped to speaker-dependent biases which are added to input features to compensate the speaker variability. The speaker normalization is only conducted in the input feature space. In that sense, it still belongs to the family of feature space normalization at input of DNNs. In [18], factorized hidden layers (FHL) are constructed as a linear combination of speaker dependent bases for neural network adaptation. The bases are initialized with i-vectors and optimized during training. The FHL strategy is conceptually different from the embedding-based SAT investigated in this work.

The embedding-based SAT introduces full affine transformations with scaling and bias. If the bias is null, it amounts to a gating mechanism that performs element-wise positive scaling of the outputs of the hidden units, which is equivalent to a SAT-LHUC form with the scaling vectors constructed from speaker embedding vectors.

The embedding vectors provide a flexible way of generating transformations with various granularity. For instance, the embedding can be performed on utterances, conversations or speakers and the resulting transformations will have the corresponding granularity.

Speech signals are known to be affected by a variety of variabilities, most notably speaker, environment and channel. Creating canonical representation spaces by removing these speech variabilities has always been pursued in acoustic modeling. In the embedding-based adaptive training framework, if one can estimate the relevant embedding of a known variability, the generated transformations will canonicalize the internal feature spaces of a DNN acoustic model accordingly. It is fairly straightforward to extend this framework for noise or channel normalization if the necessary embedding can be accomplished.

Table 2: WERs(%) on Hub5 2000 Switchboard (SWB) and Callhome (CH) of LSTMs with and without embedding-based SAT under the CE criterion. The models in the first row uses speaker-adapted FMLLRs feature (denoted fmllr). The models in the second row are speaker-aware training by appending i-vectors to the FMLLR features (denoted fmllr+ivec). The models in the third row are SAT models using fmllr+ivec features. The SAT normalization is applied after each of the bottom two LSTM layers for the 5-layer LSTM case, otherwise after the bottom LSTM layer only.

<table>
<thead>
<tr>
<th></th>
<th>WER (SWB)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1L</td>
<td>2L</td>
<td>3L</td>
<td>4L</td>
</tr>
<tr>
<td>fmllr</td>
<td>13.3</td>
<td>11.9</td>
<td>11.5</td>
<td>11.4</td>
</tr>
<tr>
<td>fmllr+ivec</td>
<td>12.1</td>
<td>11.2</td>
<td>10.8</td>
<td>10.7</td>
</tr>
<tr>
<td>fmllr+ivec+sat</td>
<td>12.0</td>
<td>10.9</td>
<td>10.5</td>
<td>10.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>WER (CH)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1L</td>
<td>2L</td>
<td>3L</td>
<td>4L</td>
</tr>
<tr>
<td>fmllr</td>
<td>22.6</td>
<td>20.7</td>
<td>19.9</td>
<td>19.7</td>
</tr>
<tr>
<td>fmllr+ivec</td>
<td>21.3</td>
<td>19.6</td>
<td>18.9</td>
<td>18.7</td>
</tr>
<tr>
<td>fmllr+ivec+sat</td>
<td>21.0</td>
<td>19.3</td>
<td>18.3</td>
<td>18.4</td>
</tr>
</tbody>
</table>


