Regularized Feature-space Discriminative Adaptation for Robust ASR

Takashi Fukuda\textsuperscript{1}, Osamu Ichikawa\textsuperscript{1}, Masafumi Nishimura\textsuperscript{1}, Steven J. Rennie\textsuperscript{2}, Vaibhava Goel\textsuperscript{2}

\textsuperscript{1}IBM Research – Tokyo, 6-52, Toyosu 5-chome, Koto-ku, 135-8511, JAPAN
\textsuperscript{2}IBM T. J. Watson Research Center, Yorktown Heights, NY10598
\{fukudal,ichikaw,nisimur}@jp.ibm.com, \{ajrennie,vgoel\}@us.ibm.com

Abstract

Model-space adaptation techniques such as MLLR and MAP are often used for porting old acoustic models into new domains. Discriminative schemes for model adaptation based on MMI and MPE objective functions are also utilized. For feature-space adaptations, one extension to the well-known feature-space discriminative training (fMPE) algorithm, feature-space discriminative adaptation, was recently proposed to adapt fMPE transforms. Feature-space discriminative adaptation was shown to work well for some situations when sufficient adaptation data is available. This paper improves the feature-space discriminative adaptation by introducing a regularization term for an indirect differential computation of the fMPE objective function, and also by updating the acoustic models with MAP instead of ML criterion during the adaptation. The proposed method performed favorably for the adaptation conditions from general-purpose LVCSR to automotive environments with small amounts of adaptation data, and yielded 4.4\% relative improvement as compared with MAP-adapted system without using the fMPE adaptation.

Index Terms: LVCSR, discriminative adaptation, regularization, MAP criterion

1. Introduction

A state-of-the-art Large Vocabulary Continuous Speech Recognition (LVCSR) system is usually trained with at least several hundred speakers in a target domain to provide robustness for many speakers and various environmental factors [1, 2]. Since Automatic Speech Recognition (ASR) performance is highly dependent on the acoustic environment in the target domain, an acoustic model in such a system is normally built with a large amount of target domain data. However, creating a large speech corpus for each ASR application involves enormous expenses and constructing an acoustic model for a new application from scratch also takes a lot of time. Therefore, model adaptation techniques are often used to convert one of the existing models into the target domain by using a small amount of target domain data.

Typical adaptation techniques such as Maximum Likelihood Linear Regression (MLLR) [3] and Maximum a Posteriori (MAP) [4] only adapt the acoustic model parameters. MPE-MAP and MMI-MAP also incorporate discriminative criterion into the MAP adaptations to vary only the model space [5]. In contrast, a frontend pipeline in many cutting-edge ASR systems includes a discriminative feature-space transform (IMPE transform) that is statistically trained with a large speech corpus to map the cepstrum-based or LDA features into a more canonicalized feature space [6]. This means that the transform depends on the acoustic conditions of the training data and should also be adapted to target a domain. Recently, an extension of feature-space discriminative training was proposed for such feature-space adaptation and was shown to work for the adaptation task from Conversational Telephone Speech (CTS) and Broadcast News (BN) to an ASR system for meetings [7]. In the paper, 200 hours of adaptation data from meetings was used to port both the fMPE transforms and acoustic models trained with 800 to 1400 hours of CTS and BN data for the meeting domain. However, the method needed comparatively a large amount of adaptation data to perform reliably.

This paper addresses an improved method for feature-space discriminative adaptation (the adaptation of the fMPE transform) that functions with a smaller amount of adaptation data by incorporating a regularization term into the indirect differential computation in the fMPE adaptation procedure. In [7], the MAP criterion was used for the acoustic model updates instead of the Maximum Likelihood (ML) criterion. Our proposed method also interleaves the acoustic model updates with MAP criterion during the fMPE adaptation. Experiments show that performing some additional iterations of fMPE training with small adaptation data that starts from the fMPE transform as an initial transform which was trained with a large data set in other domains causes over-training and performance degradation due to the shortage of adaptation data. Next it is shown that the combined usage of the regularization term and acoustic model updates based on MAP criterion effectively reduces the over-training of the feature-space discriminative adaptation, and provided a 4.4\% relative improvement over the ASR system adapted by the conventional model-space MAP adaptation.

The rest of this paper is organized as follows. Section 2 reviews a conventional fMPE algorithm. Section 3 describes the proposed feature-space discriminative adaptation procedure, and Sections 4 and 5 cover the experimental results. Finally, Section 6 presents our conclusions.

2. Brief Summary of fMPE

This section briefly summarizes a standard fMPE training algorithm focusing on the key points of comparison related to our proposed method. Most of the notations in this section are based on the original paper [6].

In the fMPE algorithm, a linear transform $M$ is applied to a posterior probability vector $h_t$ obtained by $g \in \mathcal{G}$ given $x_t$ to move the original feature vector $x_t$ at time $t$ to a more discriminative vector

$$y_t = x_t + Mh_t,$$  \hspace{1cm} (1)

where $y_t$ refers to the transformed feature. The set $\mathcal{G}$, of size $G$, is created by clustering the Gaussians of the acoustic model trained with the ML criterion. The transform $M$ is estimated by
maximizing MPE objective function

\[ M^* = \arg \max_M \mathcal{F}_{MPE}(y, \lambda), \]

(2)

where \( \lambda \) is a set of new model parameters. During the fMPE training, first order gradient descent optimization is used to update each matrix element \( M_{ij} \)

\[ M_{ij} := M_{ij} + \nu_{ij} \frac{\partial \mathcal{F}}{\partial M_{ij}}, \]

(3)

where \( \nu_{ij} \) is the parameter-specific learning rate. The term of \( \frac{\partial \mathcal{F}}{\partial M_{ij}} \) is decomposed into two parts consisting of a direct differential and an indirect differential

\[ \frac{\partial \mathcal{F}}{\partial M_{ij}} = \sum_{t=1}^{T} \frac{\partial \mathcal{F}}{\partial y_{ij}} h_{ij} \]

\[ = \sum_{t=1}^{T} \left[ \frac{\partial F_{\text{direct}}}{\partial y_{ij}} + \frac{\partial F_{\text{indirect}}}{\partial y_{ij}} \right] h_{ij}. \]

(4)

The direct differential is associated with the change of \( F \) directly caused by the variation of \( y_{ij} \), and the indirect differential is related to the changes of \( F \) caused by variations of the Gaussian means and variances that are indirectly related to the \( y_{ij} \) variation. In Equation (4), the term of \( \frac{\partial F_{\text{indirect}}}{\partial y_{ij}} \) can be further described as

\[ \frac{\partial F_{\text{indirect}}}{\partial y_{ij}} = \sum_{s=1}^{S} \sum_{m=1}^{M_s} \xi(s, m) \]

\[ \xi(s, m) = \frac{\gamma_{sm}(t)}{\gamma_{sm}} \left\{ \frac{\partial F}{\partial \mu_{smi}} + 2 \frac{\partial F}{\partial \sigma_{smi}} (y_{ij} - \mu_{smi}) \right\}, \]

(5)

where \( S \) and \( M_s \) are the numbers of states and Gaussians corresponding to the state \( s \) in the ML-trained acoustic model, respectively. The \( \gamma_{sm} \) is the ML occupation probability as used in the standard forward-backward training, and the \( \gamma_{sm} \) is the same probability summed over all of the training data. During the fMPE training, not only the fMPE transforms but also acoustic model parameters are updated with the estimated fMPE features following the standard ML training. See the original paper [6] for more details, including the direct differential computation.

3. Regularization of fMPE adaptation

3.1. Over-training

The most straightforward way to port the fMPE transform \( M \) trained on a large data set in the original domain into the target domain is to additionally train the transform \( M \) only with a target domain’s data (or both of the data sets) by applying the conventional fMPE training procedure as is and simply replacing the original domain training data with the target domain data. Table 1 compares the fMPE transform in the original domain with one that was additionally trained only with the replacement of the data. In the comparison, the original-domain acoustic model was ML and was discriminatively trained with a large amount of general-purpose LVCSR data. There were 400K Gaussians and 20K quinphone context-dependent states. The adaptation data here is 60 hours of automotive data. The recognition task and the vocabulary variation of the automotive ASR is the same as general-purpose LVCSR. Only the acoustic environment is different between the two (See Section 4 for details).

As shown in the table, the fMPE transform that receives additional training by replacing the original training data with the small adaptation data (automotive data) can lose accuracy because of the data sparseness. Since the parameter size of the acoustic model (the numbers of Gaussians and states) is very large as compared to the adaptation data size, the transform is easily over-trained. This is because, as shown in Equation (5), the objective function of fMPE is based on the statistics from all of the states and Gaussians in the acoustic model on indirect differential computation. The statistics for the direct differential computation also depend on the size of the ML-trained acoustic model. In other words, the small adaptation data set is insufficient to cover all of the states. Thus, the statistics have limited reliability for many of the states.

3.2. Overview of the proposed method

In this paper, a regularization parameter is first introduced to the fMPE adaptation to give a penalty for the unreliable statistics due to low counts. The parameter is set to use only the reliable statistics when the fMPE adaptation is performed as

\[ \frac{\partial F_{\text{indirect}}}{\partial y_{ij}} = \sum_{s=1}^{S} \sum_{m=1}^{M_s} \rho(\gamma_{sm}^{\text{num}}, \gamma_{sm}^{\text{den}}) \xi(s, m), \]

(6)

where \( \rho(\gamma_{sm}^{\text{num}}, \gamma_{sm}^{\text{den}}) \) is the regularization term, and \( \gamma_{sm}^{\text{num}} \) and \( \gamma_{sm}^{\text{den}} \) respectively denote the MPE statistics on the numerator and denominator lattices. The \( \gamma_{sm}^{\text{num}} \) and \( \gamma_{sm}^{\text{den}} \) are used for computing \( \frac{\partial F}{\partial \mu_{smi}} \) and \( \frac{\partial F}{\partial \sigma_{smi}} \) as described in Equation (5) in the original fMPE algorithm as

\[ \frac{\partial F}{\partial \mu_{smi}} = \frac{\kappa}{\sigma_{smi}^2} \left\{ \theta_{smi}^{\text{num}} (\mu_{smi} - \mu_{smi}) - \theta_{smi}^{\text{den}} (\mu_{smi} - \mu_{smi}) \left\{ \gamma_{sm}^{\text{num}} - \gamma_{sm}^{\text{den}} \right\} \right\}, \]

(7)

\[ \frac{\partial F}{\partial \sigma_{smi}} = \frac{\kappa \gamma_{sm}^{\text{num}}}{2} \left( \sigma_{smi}^4 - \sigma_{smi}^2 \right) - \frac{\kappa \gamma_{sm}^{\text{den}}}{2} \left( \sigma_{smi}^4 - \sigma_{smi}^2 \right), \]

(8)

where \( \theta_{smi}^{\text{num}} \) and \( \theta_{smi}^{\text{den}} \) are the sums of the data for the numerator and denominator lattices, and \( \kappa \) indicates the probability scale. The variances \( \gamma_{smi}^{\text{num}} \) and \( \gamma_{smi}^{\text{den}} \) are for the numerator and denominator statistics around the current mean.

For the regularization included in Equation (6), any functions can be used as \( \rho(\gamma_{sm}^{\text{num}}, \gamma_{sm}^{\text{den}}) \) as long as they range from 0 to 1 and take a value close to 0 when \( \gamma_{sm}^{\text{num}} \) and \( \gamma_{sm}^{\text{den}} \) are

<table>
<thead>
<tr>
<th>Transforms</th>
<th>%WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original fMPE</td>
<td>17.66</td>
</tr>
<tr>
<td>Additionally-trained fMPE</td>
<td>23.20</td>
</tr>
</tbody>
</table>
small. This paper uses a Sigmoid function as the regularization term

\[ \rho(\gamma_{\text{num}}^{\text{sm}}, \gamma_{\text{den}}^{\text{sm}}) = \frac{1}{1 + \exp(-\alpha\beta_{\text{sm}} + \epsilon)} \]  

(9)

where

\[ \beta_{\text{sm}} = \gamma_{\text{num}}^{\text{sm}} \text{ if } \gamma_{\text{num}}^{\text{sm}} < \gamma_{\text{den}}^{\text{sm}} \]

\[ \beta_{\text{sm}} = \gamma_{\text{den}}^{\text{sm}} \text{ otherwise} \]

In this equation, \( \alpha \) is the gain control coefficient and \( \epsilon \) is the shift coefficient.

A simpler implementation of the regularization parameter \( \rho \) pertains to the case of a step function with a threshold \( \eta \) as follows

\[ \rho(\gamma_{\text{num}}^{\text{sm}}, \gamma_{\text{den}}^{\text{sm}}) = u(\beta_{\text{sm}}) \]

\[ u(\beta_{\text{sm}}) = 1 \text{ if } \beta_{\text{sm}} > \eta \]

\[ u(\beta_{\text{sm}}) = 0 \text{ if } \beta_{\text{sm}} \leq \eta, \]

where

\[ \beta_{\text{sm}} = \gamma_{\text{num}}^{\text{sm}} \text{ if } \gamma_{\text{num}}^{\text{sm}} < \gamma_{\text{den}}^{\text{sm}} \]

\[ \beta_{\text{sm}} = \gamma_{\text{den}}^{\text{sm}} \text{ otherwise} \]

We apply the regularization term only to the indirect differential computations, and the direct differential computations remain the same because it assumes the model parameters are being held constant. When values of \( \rho \) for all of the states and Gaussians are all 1, the indirect differential computation for the adaptation is regarded as the original one.

In addition to the regularization term, we next introduce the MAP criterion to update the acoustic model between each iteration of fMPE adaptation, because the adapted IMPE feature critically differs from the original IMPE feature although ML criterion is used for the model updates in the original IMPE algorithm. In our proposed method, MAP updates of the acoustic model also has a simple regularization process that does not update the Gaussians below a certain count. After the feature-space discriminative adaptation, the model-space discriminative training is continued with the adapted features. The prior study addressed the use of \( \tau \) for the indirect computation to control the proportion of new statistics from the adaptation data [7]. Our method can be combined with the method described in [7] to suppress the over-training.

4. Experiments

The experiments presented in this paper are all based on speaker-independent models that are discriminatively trained on a general-purpose LVCSR task. We present the results from an in-house test set including an abundant English LVCSR data set and automatic data. In the LVCSR data set, the recording environments were quite diverse, including offices, homes, restaurants, and trains, but there is no location information tagged to each utterance. The acoustic models are built on data from several hundred speakers with the data ranging from a few seconds to few hours per speaker.

Figure 1: Schematic diagram of adaptation experiments.

4.1. Acoustic Feature

The front-end acoustic features are 13-dimensional MFCC features that are computed from 25-ms frames with a 10-frame shift. The derivative features were extracted from the static MFCC, and were then combined with each other to form 48-dimensional features. The sampling frequency was 16kHz. Utterance-level mean normalization for the static features, where the statistics were calculated only on the speech regions of the data, was used throughout the ML and discriminative training steps. In the ML training, LDA+MLLT transform was generated to project the 48-dimensional MFCC-based features onto the 24-dimensional features [8]. The LDA+MLLT transform updates were interleaved with standard HMM updates during the ML training.

4.2. Baseline Acoustic Modeling

Words in the recognition lexicon are represented as sequences of phones, and the phones are modeled with 3-state left-to-right HMMs that do not permit state skipping. Acoustically distinct variants of the HMM states are identified using decision trees that ask questions about the phonetic context in which a state occurs, and the leaves of the decision trees are the basic acoustic
units that we model.

We used a large amount of the general-purpose LVCSR data to build a source (original) domain acoustic model as the baseline. The LVCSR data represented several hundreds of thousands of speech. The baseline ML model contains roughly 400K Gaussians with 20K quinphone context-dependent states. After the ML training, the feature-space discriminative training was done using the procedure described in Section 2 to map the acoustic features to a more canonicalized space, and then the acoustic models are further discriminatively trained using the MPE criterion on top of the fMPE training. For the feature-space discriminative training, a 512-dimension posterior probability vector \( h_t \) was used, and the inner and outer context transforms were set to 8 and 4, respectively.

4.3. Adaptation Experiments

Figure 1 shows a block diagram of the experiments. In the experiments presented here, the fMPE transform and the acoustic model in the baseline system constructed with the general-purpose LVCSR data were ported to the automotive domain by using several adaptation techniques including MAP adaptation, feature-space discriminative adaptation, and the proposed method. The adaptation data consists of 60 hours of automotive data. The recognition task for the automotive environment was also LVCSR, which is the same as the baseline, and thus there is no difference between the general-purpose LVCSR and the automotive ASR in terms of the language modeling. The results are presented on an automotive test set consisting of 8,976 utterances spoken by 44 speakers (22 males and 22 females) that were recorded with a far-field microphone in moving and parked cars.

5. Results

The experimental results are given in Table 2. In the table, “Baseline” is the result with the acoustic model trained by the LVCSR data, meaning no adaptation. “MAP” indicates the result by MAP adaptation limited to the model space. The results on the feature-space discriminative adaptation without using the regularization are shown as “fMPE-Adapt. ML” and “fMPE-Adapt. MAP”. The “IMPE-Adapt. ML” updated the acoustic model by the ML criterion, and the “IMPE-Adapt. MAP” updated it by the MAP criterion. For the “fMPE-Adapt. MAP” results, the smoothing parameter \( \tau \) was used also for the indirect differential computation [7]. Lastly, the effects of the regularization on the feature-space discriminative adaptation were examined in the “Regl. fMPE-Adapt. ML” and “Regl. fMPE-Adapt. MAP” nouns.

From the table, we can see that MAP adaptation worked very well, as already proved in the literature. The relative improvement obtained by the conventional MAP adaptation was 9.17%. The reason why “fMPE-Adapt. ML” and “fMPE-Adapt. MAP” degraded the performance is that the data sparseness against the baseline model size was very large, so the IMPE transform was overly trained as discussed in Section 3.1. Considering “Regl. fMPE-Adapt. ML” and “Regl. fMPE-Adapt. MAP”, both of them reduced the WER as compared to the baseline system, but the improvement of “Regl. fMPE-Adapt. ML” was marginal and was worse than MAP adaptation alone. In contrast, the proposed method “Regl. fMPE-Adapt. MAP” achieved significant improvements over both the baseline system and the MAP-adated model, and in addition, the model-space discriminative training on top of feature-space discriminative training increased the relative improvement by 13.19% as compared to the baseline system.

6. Conclusion

This paper addressed an improved method of feature-space discriminative adaptation with a focus on over-training due to data sparseness. The regularization term was applied to the indirect differential computation during the fMPE adaptation, and also, the MAP criterion was utilized for the model updates instead of ML criterion. Adapting the fMPE transforms by changing the training data between the ML and feature-space discriminative training stage is not novel, but those methods usually train the fMPE transform from scratch, meaning that the initial transform is generated based on certain rules, while our method regards the fMPE transform trained with a large corpus as the initial transform. The proposed method combining the regularization and MAP update of the acoustic model showed gains of up to 4.4% relative to the model-space MAP-adapted model.

In the future, we will investigate leveraging other regularization techniques such as L1 and L2 regularizations for updating the transform, and apply this scheme to sequential training for a DNN-based system.

7. References