ABSTRACT

In the past decade, methods to extract long-term acoustic features for speech recognition using Multi-Layer Perceptrons have been proposed. These features have been proved to be good complementary features in some feature augmentations and/or through system combination. Usually, conventional linear dimension reduction algorithms, e.g. Linear Discriminative Analysis, are not applied on the combined features. In this paper, Region Dependent Transform is applied to jointly optimize the feature combination under a discriminative training criterion. When compared to a conventional augmentation, 3% to 6% relative character error rate reduction for Mandarin speech recognition has been achieved using Region Dependent Transform.

Index Terms— Multi-Layer Perceptrons, Region Dependent Transform, discriminative training, Mandarin speech recognition

1. INTRODUCTION

In the past decade, a number of methods, e.g. TempoRAI Patterns (TRAP), have been proposed to capture long-term features (0.5 to 1 second) for speech recognition using Multi-Layer Perceptrons (MLP) [1, 2]. It has been shown that these MLP features do not outperform the conventional ones, for instance, Perceptual Linear Predictive (PLP) [3] features, but are good complementary features in feature augmentation and/or through system combination [4, 5, 6]. Usually, acoustic feature dimension needs to be reduced in a recognition system for efficiency in acoustic modeling. However, the conventional linear dimension reduction algorithms, e.g. Linear Discriminative Analysis (LDA), do not directly minimize speech recognition errors. The resulting transform could be sub-optimal. To avoid undermining the discriminative power of the MLP features, such methods are not applied on the PLP together with the MLP features [6] though the two kinds of features might share some information.

In this work, Region Dependent Transform (RDT) [7, 8] is used to reduce the dimensionality of a combined feature of PLP and MLP. Since RDT is a discriminatively trained piecewise linear transformation, the discriminative power from both sets of features could be kept and the redundant information excluded in dimension reduction. In addition to using PLP and MLP features as input to RDT, we also use the MLP phone posterior probabilities to define the regions. This results in a fusion of the two algorithms. Experimental results show that 3% to 6% relative reduction in Character Error Rate (CER) is achieved for Mandarin speech recognition using RDT with MLP when compared to a conventional feature augmentation approach.

The paper is organized as follows. We describe the conventional way of feature augmentation in our system in Section 2. In Section 3, the theory of RDT is reviewed and its application in this work is depicted. The experimental results are given in Section 4. Section 5 concludes the paper.

2. CONVENTIONAL USE OF MLP FEATURES

In the work, MLP bottle-neck features derived from the time-warped Linear Predictive TRAP (wLP-TRAP) [2] are used\(^1\). The settings in wLP-TRAP feature extraction and configurations of the MLP reported in [6] are adopted. The MLP features were incorporated into our Mandarin Speech-to-Text (STT) system for the GALE project. There are 76 phones (including silence) in our Mandarin STT system. Hence, the total number of nodes in the input, hidden, bottle-neck and output layer of the MLP are 475, 3500, 26 and 76, respectively.

In our standard speaker independent (SI) acoustic model (AM) training [9], the basic features (15 PLP and 1 pitch ) of 9 frames are concatenated and LDA is then applied to reduce the dimension to a designated number, e.g. 60. Maximum Likelihood Linear Transform (MLLT) [10] is then applied on the dimension-reduced features to diagonalize the covariance matrices in the AM. The resulting features are denoted as Long Span Features (LSF). To incorporate MLP features in SI training, we append the 26 MLP bottle-neck features to the

\(^1\)We would like to thank Petr Fousek for sharing the software to extract wLP-TRAP features with us.
and 144 features derived from PLP and pitch. Instead of using posterior probability produced by GMM, the phone

It is because speaker variability that is irrelevant to the resulting features are then appended to the 45-dimensional features. MLLT is then applied on the 71-dimensional features.

In a general speaker adaptive training (SAT), the process of deriving the LSF is similar to the SI training [9]. To remove inter-speaker variabilities, the LSF extraction procedure described in the SI training is placed between two speaker dependent transforms (SDT). The SDT which is applied on the basic features before frame concatenation is denoted as pre-SDT while the one at the end post-SDT. Both transforms are obtained using constrained Maximum Likelihood Linear Regression (MLLR) [11]. The final features are denoted as SAT LSF. The SAT models are then trained from the SAT LSF.

There are many possible ways of combining the MLP features with the basic features in the SAT training. The feature combination shown in Figure 1 provided better results in our experiments. MLLT is applied on the 26 MLP features and the resulting features are then appended to the 45-dimensional SAT LSF. No SDT is needed to be applied on the MLP features. It is because speaker variability that is irrelevant to phone classification can be reduced in the MLP training.

3. JOINT OPTIMIZATION USING RDT

Region Dependent Transform (RDT) was successfully used as a method for discriminative feature extraction in large vocabulary speech recognition [7, 8]. A global Gaussian mixture model (GMM) is usually used to divide the acoustic space into a number of regions. Based on the application, the regions can represent different acoustic categories, e.g. phone or phone state clusters. The transformed feature of RDT is a weighted sum of the region-specific transformed features, defined as

\[ x_t = \sum_{i=1}^{N} p(i|\alpha_t) A_i \ast f_t \]

where \( N \) is the total number of regions, \( p(i|\alpha_t) \) the posterior probability of region \( i \) given a feature vector \( \alpha_t \) at time \( t \), \( A_i \) the affine transformation for region \( i \), \( f_t \) the input feature vector at time \( t \). In general, \( \alpha_t \) and \( f_t \) can be different acoustic observations derived from the same speech frame with different time span using different algorithms.

The affine transforms, \( A_i \), can be estimated discriminatively under the Minimum Phone Frame Error (MPFE) criterion [12]. For \( R \) training utterances with transformed features \( X = \{X_1, ..., X_R\} \) and reference transcriptions \( \{W_1^{ref}, ..., W_R^{ref}\} \), the MPFE objective function is defined as

\[ H_{MPFE}(X, \lambda) = \sum_{r=1}^{R} \sum_{t} p(X_r | W_r, \lambda)^{\beta} p(W_r) \alpha(W_r, W_r^{ref}) \]

where \( X_r = \{x_1^r, ..., x_N^r\} \) is the sequence of transformed feature vectors of utterance \( r \) having \( k \) frames long, \( W_r \) is a hypothesis word sequence of the utterance, \( \alpha(W_r, W_r^{ref}) \) is the phoneme frame accuracy score of that hypothesis with respect to the reference, the exponent \( \beta \) is used to reduce the dynamic range of the acoustic scores, and \( \lambda = \{\mu_1, ..., \mu_S, \Sigma_1, ..., \Sigma_S\} \) is the set of Gaussian means and covariances of an HMM that is trained from \( X \). \( \lambda \) is trained using the Maximum Likelihood (ML) criterion given current features. The use of ML criterion for \( \lambda \) is to force the training to focus on improving the feature transforms [7]. The resulting transforms can also be utilized to train other ML acoustic models.

Explicitly, \( A_i \) can be optimized by gradient decent using derivative

\[ \frac{\partial H_{MPFE}(X, \lambda)}{\partial A_i} = \sum_{r,t} p(i|\alpha_t^r) \frac{\partial H_{MPFE}(X_r, \lambda)}{\partial x_t^r} f_t^T \]

The derivative of MPFE with respect to \( x_t^r \) can be expressed as a summation of two terms:

\[ \frac{\partial H_{MPFE}(X, \lambda)}{\partial x_t^r} = \left( \frac{\partial H_{MPFE}(X, \lambda)}{\partial x_t^r} \right)_\lambda + \frac{\partial H_{MPFE}(X, \lambda)}{\partial \lambda} \frac{\partial \lambda}{\partial x_t^r} \]

The first term is the derivative of \( H_{MPFE}(X, \lambda) \) with respect to \( x_t^r \) holding \( \lambda \) as a constant. The transforms and HMM are updated iteratively.

To exploit the information extracted from MLP, a framework of SAT RDT training is used. The details of the general SAT RDT training can be found in [8]. In this work, we define a region as a phone in our system so that they are more relevant to the MPFE and MLP training criteria. As shown in Figure 2, the 170-dimensional input feature, \( f_t \) to RDT is the concatenation of 26 bottle-neck features from the MLP and 144 features derived from PLP and pitch. Instead of using posterior probability produced by GMM, the phone
Fig. 2. Block diagram for applying RDT on PLP together with MLP features

The region specific transforms, $A_t$, are initialized as

$$A_t^0 = [L \ 0], \ \forall t \in [1, N]$$

where $L$ is the projection of the 144-dimensional features in Figure 2. It is estimated using LDA and MLLT. $0$ is a 60-by-26 zero matrix. Thus there is no contribution from the MLP features in the initial transformations. The initial RDT-transformed features are actually the SAT LSF without using post-SDT. As a result, the initial HMM can be trained without the augmentation of MLP features. Based on this design, RDT performs a joint optimization on feature transformation to the combined feature of the LSF and MLP. In addition to making no assumption on the characteristics of the MLP features, it is possible for redundant information to be excluded and complementary features be kept under the MPFE criterion.

The RDT-transformed features are further transformed by a set of post-SDT which is estimated based on the HMM obtained from the RDT training using constrained MLLR. The SAT-RDT HMM can now be trained from the resulting features using the ML criterion. Finally, the ML SAT-RDT models are updated using MPFE training.

4. EXPERIMENTS

4.1. System Setup

The baseline acoustic models were built on 60-dimensional LSF derived from the basic features (15 PLP and 1 pitch) of 9 frame concatenation. The details of the basic feature extraction can be found in [9]. All of the acoustic models were trained from the 1900 hours of audio data released by LDC for the GALE Phase 4 evaluation. The language models (LM) used in the ML experiments were for the GALE P3 Evaluation and they were trained on 4.6 billion characters text, while those in the MPFE experiments were trained on 6.8 billion characters text for GALE P4 Evaluation. Both of them were estimated on a vocabulary size of 64K. Details of the decoding system setup used in this work are similar to [9].

The systems were evaluated on four test sets: dev07 (2.6 hours), dev08 (1.0 hour), eval08 (1.5 hours) and dev09sub (3.0 hours). Where eval08 is the non-sequestered data of the GALE Phase 3 evaluation set. dev07 and dev08 were released by LDC as the development sets for GALE Phase 2 and 3 evaluations, respectively. dev09sub is chosen by BBN and shared in the community for the development for GALE Phase 4 evaluation. dev07 served as a cross-validation set in the MLP training and a tuning set for the LM and decoding optimization.

The MLP was trained on the 1900 hours audio data. The frame-level phone labels were from the Viterbi alignments produced by the baseline ML models. The ICSI public software, QuickNet [13], was used for the neural network training. The best frame accuracy rate on dev07 is 65.8% while 66.7% for the training data.

### 4.2. Experimental Results

The MLP unadapted results in Table 1 show that, by incorporating the 26 MLP features in the SI training using the conventional feature augmentation, 6% relative reduction in CER is observed for dev07 (from 12.8% to 12.0%) and 4% for both eval08 (17.0% to 16.4%) and dev09sub (13.4% to 12.9%) while no improvement is obtained for dev08.

<table>
<thead>
<tr>
<th>Systems</th>
<th>dev07</th>
<th>dev08</th>
<th>eval08</th>
<th>dev09sub</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>12.8</td>
<td>11.1</td>
<td>17.0</td>
<td>13.4</td>
</tr>
<tr>
<td>+MLP</td>
<td>12.0</td>
<td>11.1</td>
<td>16.4</td>
<td>12.9</td>
</tr>
</tbody>
</table>

Table 1. Character error rate for ML speaker independent systems with and without MLP

The ML speaker-adapted results in Table 2 show that most of the improvement from augmenting MLP features in the conventional way vanishes after SAT training and/or speaker adaptation. We obtained 0.2% and 0.3% absolute CER reduction for dev07 and dev09sub, respectively. Nevertheless, 0.3% and 0.1% absolute degradation are also observed for dev08 and eval08, respectively. These results were produced by the system which was optimized for the augmented features. Instead of using a full matrix in the MLLR adaptation, a block diagonal matrix has to be used to keep the LSF and
MLP features separated, otherwise, the results would be degraded significantly. This could be because feature dimension increases and there is not enough data for estimating a full matrix in adaptation.

<table>
<thead>
<tr>
<th>Systems</th>
<th>dev07</th>
<th>dev08</th>
<th>eval08</th>
<th>dev09sub</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>10.5</td>
<td>9.2</td>
<td>13.9</td>
<td>10.9</td>
</tr>
<tr>
<td>+MLP</td>
<td>10.3</td>
<td>9.5</td>
<td>14.0</td>
<td>10.6</td>
</tr>
<tr>
<td>+RDT</td>
<td>9.7</td>
<td>8.7</td>
<td>12.7</td>
<td>10.1</td>
</tr>
</tbody>
</table>

Table 2. Character error rate for ML speaker adapted systems with MLP before and after RDT

The results for applying RDT on the augmented features in the SAT training are shown in the third row of Table 2. By using RDT, we observed 5% to 9% relative CER reduction as compared to the baseline system. The unadapted hypotheses from the baseline system were used as the supervision for model adaptation for the RDT system. Since the region-dependent transforms were trained under a discriminative criterion, it might not be entirely fair to compare it with the purely ML-trained baseline and conventional MLP augmented system. The improvement might be shrunk when all of the systems are updated using the MPFE training.

After the MPFE model training and speaker adaptation, as shown in Table 3, there is no significant improvement by using the conventional concatenation of MLP and PLP features, compared to the baseline system. By contrast, we obtained 3% to 6% relative reduction in CER using the SAT RDT training, when compared to the conventional feature augmentation and the baseline system.

<table>
<thead>
<tr>
<th>Systems</th>
<th>dev07</th>
<th>dev08</th>
<th>eval08</th>
<th>dev09sub</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>9.4</td>
<td>8.3</td>
<td>12.3</td>
<td>9.2</td>
</tr>
<tr>
<td>+MLP</td>
<td>9.3</td>
<td>8.3</td>
<td>12.5</td>
<td>9.2</td>
</tr>
<tr>
<td>+RDT</td>
<td>9.0</td>
<td>7.8</td>
<td>11.7</td>
<td>8.9</td>
</tr>
</tbody>
</table>

Table 3. Character error rate for MPFE speaker adapted systems with/without MLP and RDT

5. CONCLUSION

This paper has presented a method of using Region Dependent Transform to jointly optimize the feature augmentation of the Perceptual Linear Predictive and Multi Layer Perceptron features under a discriminative training criterion. As compared to a conventional augmentation, we achieved 3% to 6% relative reduction in character error rate for Mandarin speech recognition after speaker adaptation using a discriminatively trained system. The method is currently evaluated on a Mandarin speech recognition system and, it is planned to be applied on other languages, e.g. Arabic and English.

6. REFERENCES