ADAPTIVE MODEL COMBINATION FOR DYNAMIC SPEAKER SELECTION TRAINING

Chao Huang, Tao Chen* and Eric Chang

Microsoft Research Asia
5F, Beijing Sigma Center, No. 49, Zhichun Road, Beijing 100080, China
Department of Automation, Tsinghua University, Beijing 100084, China
{chaoh, echang}@microsoft.com; chentao@proc.au.tsinghua.edu.cn

ABSTRACT
Acoustic variability across speakers is one of the challenges of speaker independent speech recognition systems. In this paper we propose a two-stage speaker selection training method for speaker adaptation. After cohort speakers are selected for test speaker, an adaptive model combination method is developed to replace the formerly used retraining process. In addition, impacts of number of selected cohort speakers and number of utterances from test speaker are investigated. Preliminary experiments on dynamic speaker selection are shown. Relative error rate reduction of 12.27% is achieved when only 10 utterances are available. Finally, further extensions of model combination scheme and dynamic selection are discussed.

1. INTRODUCTION

Acoustic variability across speakers is one of the challenges of speaker independent (SI) speech recognition system. In state-of-the-art speaker adaptation (SA) methods dealing with this problem, there are three categories: linear transformation family (such as MLLR [8]), Bayesian learning family (such as MAP [7]) and speaker space family (such as eigenvoice [6]). However, the effectiveness of these methods depends on the amount of data available from the test speaker. When given corpus is very limited, even dominant adaptation methods do not work well and efficiently for LVCSR.

In recent years, a promising speaker adaptation method, speaker selection training (SST), has emerged in the literature. The idea of SST [10] is to select a subset of cohort speakers from training speakers, and to build SA model based on these cohorts, either by parameter re-estimation or by model combination. In general, SST is a two-stage method: cohort speaker selection and model building.

It is hard to train a transformation matrix for MLLR with one adaptation utterance, and even more difficult to estimate the parameters with MAP. However, the data may be a good index to select acoustically similar speakers from a pool of training speakers. One motivation is that in speaker recognition field, one utterance with three seconds has achieved good accuracy on speaker identification. Then we can make full use of the statistics from the selected speaker subsets.

There are various implementations of SST in practice. In the first stage of selecting cohorts, the key issue is to define the similarity measure. Padmanabhan [10] and Wu et al [11] made use of SA models (HMM) to obtain a likelihood score. In their systems, adaptation data from test speaker are fed to SA models of training speakers to calculate the corresponding likelihood. Instead, Yoshizawa [12] and Matrouf [9] utilized likelihood score from Gaussian mixture model (GMM). In the second stage, there are also different algorithms, such as retraining in [11], MAP adaptation in [9], feature transformation in [10] and model combination in [12]. However, retraining the SI model by data of selected cohorts is very time-consuming. Model combination is the fastest, though only pre-calculated statistics are used in [12].

In our experiments [5], it is found that incorporating some learning mechanism, such as maximum likelihood (ML) or maximum a posterior (MAP) based parameter estimation, may enhance the SST method. There are two schemes we can use: 1) Feature normalization, that is, transforming all the features from selected cohorts closer to the test speaker and retraining SI model based on normalized features; 2) Model combination, that is, interpolating the Gaussian mean vectors of cohort models to obtain a new model. The concept of model combination is similar to reference speaker weighting [4], cluster adaptive training [3] and cluster weighting [2], though these likelihood do not include specific speaker subset selection. Our method can be viewed as an extension to static model combination [12] while incorporating learning mechanism.

In this paper we focus on developing SST based adaptive model combination algorithm for speaker adaptation. We propose to estimate interpolating weight vector(s) under MAP framework. The algorithm is fast as models for training speakers can be obtained offline and only a small number of parameters need to be estimated. Experiments show 10.10% relative error reduction over baseline SI system when 10 utterances are used both for selecting cohorts and for estimating the weight vector. In addition we evaluate the effect of the amount of enrollment data on adaptation performance. Even with just one utterance, 5.42% relative error reduction can be achieved. At last we present our preliminary experiment on dynamic speaker selection. It is shown that selecting different number of cohorts for test speaker can bring 12.27% relative error reduction over baseline system.

This paper is organized as follows. Section 2 briefly...
describes the basic idea of speaker selection training. Section 3 introduces our adaptive model combination algorithm. Section 4 presents our dynamic speaker selection scheme. Speaker adaptation experiments are shown in Section 5. Conclusion and discussions are provided in Section 6.

2. SPEAKER SELECTION TRAINING

Given limited data from test speaker and a certain training corpus, SST considers the following issues:

- Efficient speaker representations.
- Reliable similarity measurement between speakers.
- Number of close speakers (cohorts) to be picked out.
- Number of utterances from test speaker when considering the performance and enrollment efforts.

In [5], we tried two representations of speaker (transformation matrix and GMM/HMM) and two kinds of similarity measurement (Euclidean distance and likelihood score). Experiments showed that likelihood score from GMM is the most efficient strategy. The main procedure for SST is as follows:

- Train one GMM for each training speaker ready for selection.
- Calculate the likelihood of adaptation data of test speaker in each GMM.
- Select training speakers with the $K$ largest likelihood as cohorts.

After cohorts are selected, the information they bring can be used for speaker adaptation. There are several methods such as retraining the SI model with the data of cohorts [11] and combining speaker dependent (SD) models of cohorts with pre-calculated statistics by SI model [12]. We propose an adaptive model combination method for adaptation.

3. ADAPTIVE MODEL COMBINATION

3.1. The Proposed Algorithm

The first step of adaptive model combination is to obtain one SA model for each training speaker by MLLR adaptation from SI model. Then we linearly combine cohorts' SA models and the combination weight coefficients can be learned through the MAP framework, which is described as follows:

Assume that there are $R$ cohorts selected. For simplicity, a global weight vector is learned for all phone classes of test speakers. In practice, we can estimate different weight vectors for different phone classes determined by regression tree. For a particular Gaussian component, $m$, the mean vector for test speaker, $\mu^n$, is given by

$$\mu^n = M_m \lambda,$$

(1)

where $M_m$ is the matrix of $R$ cohort mean vectors for component $m$,

$$M_m = [\mu^n, \cdots, \mu^n],$$

(2)

where $\mu^n$ is the mean vector of Gaussian component $m$ associated with cohort $r$. The weight vector, $\lambda$, is given by

$$\lambda = [\hat{\lambda}_1, \cdots, \hat{\lambda}_R]^T.$$

(3)

The standard ML learning scheme for model combination is similar to other adaptive training schemes [3]. The goal is to find the value of $\lambda$ which maximizes the likelihood of adaptation data from test speaker. Let $X$ be observation sequence of adaptation data

$$X = [x_1, \cdots, x_N].$$

(4)

Here $N$ is the total number of adaptation observations. The objective function of ML estimation is expressed as

$$\arg \max_{\lambda} \log p(X | \Lambda).$$

(5)

As the model $\Lambda$ is determined by weight vector $\lambda$, and assuming all observations are independent, (5) can be expressed as

$$\arg \max_{\lambda} \sum_{n=1}^{N} \sum_{r=1}^{R} \log p(x_n | \lambda_r),$$

(6)

where $M$ is the number of all Gaussian mixture components associated with observation $x_n$.

In SST, a certain number of cohort speakers are selected in the first stage. We view these cohorts as prior information and will incorporate it into ML framework. A straightforward way is to use MAP estimation. We assume the prior probability distribution function (pdf) of the weight vector as Gaussian function,

$$P(\lambda) \sim N(\tilde{\lambda}, \Sigma^{-1}).$$

(7)

The ML estimation of (6) becomes the following MAP estimation,

$$\arg \max_{\lambda} \sum_{n=1}^{N} \sum_{r=1}^{R} \log p(x_n | \lambda_r)+\log p(\lambda).$$

(8)

The solution of (8) can reduce to

$$\arg \max_{\lambda} \{2v^T\lambda - \tilde{\lambda}^T U \lambda\}.$$

(9)

$U$ and $v$ are defined as follows,

$$U = \sum_{n=1}^{N} \sum_{r=1}^{R} \gamma_m(n) M_m S_m^{-1} M_m + \Sigma^{-1},$$

(10)

$$v = \sum_{n=1}^{N} \sum_{r=1}^{R} \gamma_m(n) x_n + \Sigma^{-1} \tilde{\lambda}.$$

(11)

$\gamma_m(n)$ is the posterior probability of Gaussian $m$ at time $n$, usually determined by SI model. $S_m$ is the covariance matrix.
matrix of Gaussian \( m \). By differentiating with respect to \( \lambda \) and equating to zero, we obtain the weight vector,
\[
\lambda = U^{-1}v.
\] (12)

3.2. Prior Selection

A crucial issue in the above scheme is to select prior parameters \( \lambda \) and \( \Sigma_i \) for weight vector. In our system each cohort model is weighed by its occupation probability given adaptation data,
\[
X_r = \frac{\sum_{m=1}^{M} \sum_{n=1}^{N} \gamma_m(n, r)}{\sum_{m=1}^{M} \sum_{n=1}^{N} \gamma_m(n, r)}, \quad r = 1, \ldots, R \tag{13}
\]

Here \( \gamma_m(n, r) \) is the posterior probability of Gaussian \( m \) in cohort model \( r \) at time \( n \). A diagonal covariance matrix \( \Sigma_n \) is used for simplicity. We can control the impact of the prior information by adjusting \( \Sigma_i \). The less adaptation data, the greater impact of prior weight as we have to rely more on prior information. On the other hand, when enough adaptation data are available, covariance should be set relatively large.

3.3. Incorporating Bias Item

Like MLLR adaptation and cluster weighting [2], bias item can be introduced to consider the effect of channel condition,
\[
\mu^m = M_m \lambda + b = M_m \tilde{\lambda}.
\] (14)

where \( M_m = [M_m, I] \) and \( \tilde{\lambda} = [\lambda^T, b^T]^T \). \( \tilde{\lambda} \) can still be estimated by equation (10),(11),(12). The prior for \( \tilde{\lambda} \) is set according to (13), except that entries corresponding to bias item are set to be zero.

3.4. Comments

The learning process can iterate for more than once. Before learning, the sufficient statistics of adaptation data against SI model must be accumulated. After one iteration, statistics can be more accurately estimated with new SA model. Then we can re-estimate the weight vector.

In reference speaker weighting [4], two constraints are added to maximize the goal function (6):
\[
\forall r, \quad \lambda_r \geq 0 \quad \text{and} \quad \sum_{r=1}^{k} \lambda_r = 1 \tag{15}
\]

Cluster adaptive training proposed by Gales [3] does not apply these constraints. However, the SI model is considered as a cluster and the corresponding weight is always assigned to 1. Both the constraints and incorporating SI model are aimed to achieve robust estimation. In our scheme, as acoustically close speakers are pre-selected, the prior information contained in \( \lambda \) and \( \Sigma_i \) is used to guarantee the reliability of learning weight vector(s).

4. DYNAMIC COHORT SELECTION

The fact that the optimal number of cohorts for each test speaker is different [5][12] motivates us to conduct dynamic cohort selection. After some candidate models (e.g., from 10, 20 and 30 cohorts) are obtained, a multi-model selection scheme is utilized to determine the optimal model for test speaker as follows,
\[
\Lambda_{opt} = \arg \max_{\Lambda} \{ \text{score}(\Lambda) - \alpha \times N(\text{cohorts}) \} \tag{16}
\]

Here \( \text{score}(\Lambda) \) is some kind of score of candidate model given adaptation data. We will evaluate two kinds of score: 1) likelihood (normalized to [0, 1]) from forced alignment of adaptation data against corresponding decoded transcriptions; 2) recognition accuracy of adaptation data. Since the more training speakers selected, the more risk to bring large mismatch between cohorts and test speaker, factor \( \alpha \) is used to penalize the number of cohorts, \( N(\text{cohorts}) \). Equation (16) is derived according to Bayesian information criterion, which is widely used in model selection and hypothesis test.

5. EXPERIMENT

5.1. Experiment Setup

About 180 hours of speech data from Chinese male speakers are used to train a gender-dependent SI model for the Microsoft Mandarin Whisper system [1]. In addition to such mature technologies as decision tree based state tying, context dependent modeling and trigram language modeling, tone related information has also been integrated into our system through pitch information and tone modeling. Compared with the baseline system in [5], more data are used for training and better accuracy is obtained. In one word, all improvements and results show here are achieved based on a solid and powerful baseline system.

The speakers ready for selection consist of 250 male speakers, with 200 utterances each. Typically one utterance, both in training and test set, lasts 3–5 seconds. Before model combination, 250 SA models are obtained by MLLR with 200 utterances per speaker.

Test set consists of 25 male speakers from the same accent with training set, 20 utterances each. 10 of them are used for selecting cohorts and estimating weight vector. The other 10 are used for testing. Character error rate (CER) is calculated for evaluation. In all the following experiments, one weight vector is estimated for each phone class. 65 broad phone classes, according to Mandarin phonetic structure, are used.

5.2. Results

Table 1 shows CER of model combination based SST, varying the number of cohort speakers. To illustrate the effectiveness of adaptive learning scheme, results of using prior weight only are also listed. The proposed algorithm consistently achieves lower CER over baseline system in our experiment condition. The lowest error rate is obtained when only 10 cohorts are selected. Too many cohorts will bring in some training speakers acoustically far away from the test speaker and then degrade recognition accuracy.

\[^2\] Assuming supervised adaptation.
Table 1: Comparison of different number of cohort speakers. (CER of baseline SI model is 9.56%)

<table>
<thead>
<tr>
<th># Cohorts</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learn</td>
<td>9.06</td>
<td>8.60</td>
<td>8.94</td>
<td>8.88</td>
<td>9.19</td>
<td>8.96</td>
</tr>
<tr>
<td>Rel. Err. Reduc.</td>
<td>5.23</td>
<td>10.10</td>
<td>6.53</td>
<td>7.18</td>
<td>3.86</td>
<td>6.27</td>
</tr>
</tbody>
</table>

Table 2 shows the performance of the proposed scheme varying the number of adaptation utterances. According to Table 1, 10 cohort speakers are selected for each test speaker. From Table 2 we can see that our method is quite robust against data variations. Only one utterance is sufficient to obtain 5.42% relative error rate reduction. It is suggestive when deciding the proper adaptation data to make a trade-off between enrollment effort and final performance gain.

<table>
<thead>
<tr>
<th>%</th>
<th>SI</th>
<th>1</th>
<th>3</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>CER</td>
<td>9.56</td>
<td>9.04</td>
<td>8.94</td>
<td>8.60</td>
</tr>
<tr>
<td>Rel. Err. Reduction</td>
<td>--</td>
<td>5.42</td>
<td>6.51</td>
<td>10.10</td>
</tr>
</tbody>
</table>

Table 3 illustrates the effectiveness of dynamic speaker selection. For simplicity, candidate models for selection are obtained from 10, 20 and 30 cohorts. 10 adaptation utterances per speaker are used. The Recognizer Output Voting Error Reduction (ROVER) criterion gives us the upper bound for dynamic SST. While likelihood based dynamic SST brings no further improvement, accuracy based one leads to 12.27% relative error reduction over SI model. However, the performance gap between our dynamic scheme and ROVER reminds us that much more can be done to improve recognition accuracy.

<table>
<thead>
<tr>
<th>%</th>
<th>CER</th>
<th>Rel. Err. Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic (α = 0.001) Accuracy</td>
<td>8.39</td>
<td>12.27</td>
</tr>
<tr>
<td>Likelihood</td>
<td>8.60</td>
<td>10.10</td>
</tr>
<tr>
<td>ROVER</td>
<td>7.93</td>
<td>17.04</td>
</tr>
</tbody>
</table>

Table 3: Preliminary results on dynamic speaker selection.

6. CONCLUSION AND DISCUSSIONS

In this paper, we propose a speaker selection training strategy for speaker adaptation given very limited data. Instead of time-consuming process of retraining, we develop adaptive model combination method to obtain the SA model. The scheme is efficient: 1) models for training speakers can be obtained offline; 2) only a small number of parameters are needed to be estimated; 3) explicit solution of equation (12) avoids iterative optimization. Experiments illustrate the effectiveness and reliability of the proposed scheme. In addition, we have investigated the impact of the number of cohort speakers and the number of adaptation utterances on recognition accuracy. Furthermore we present our preliminary experiment on dynamic speaker selection, which is shown to bring promising performance gain over static selection scheme.

There can be some further extensions to SST adaptation method. First, in applications where more adaptation data are available, SST may provide a good prior for other adaptation method such as MAP and MAPLR [2]. Second, dynamic SST based on recognition accuracy of adaptation data is a rather crude option, and is limited when extending to unsupervised applications. We are investigating some new criteria to efficiently assess the quality of a specific model for test speaker. In addition, more flexible strategy on the number of cohorts is undergoing in order to customize each test speaker. Furthermore, work on adaptive feature normalization after SST is also ongoing.

7. REFERENCES