Transformation and Combination of Hidden Markov Models for Speaker Selection Training

Chao Huang, Tao Chen* and Eric Chang

Microsoft Research Asia
5F, Sigma Center, No. 49, Zhichun Road, Beijing 100080, P.R. China
chaoh@microsoft.com

Abstract
This paper presents a 3-stage adaptation framework based on speaker selection training. First a subset of cohort speakers is selected for test speaker using Gaussian mixture model, which is more reliable given very limited adaptation data. Then cohort models are linearly transformed closer to each test speaker. Finally the adapted model for the test speaker is obtained by combining these transformed models. Combination weights as well as bias items are adaptively learned from adaptation data. Experiments showed that model transformation before combination would improve the robustness of the scheme. With only 30s of adaptation data, about 14.9% relative error rate reduction is achieved on a large vocabulary continuous speech recognition task.

1. Introduction
Acoustic variance across speakers is one of the main challenges of speaker independent (SI) speech recognition systems. To deal with this problem, a great amount of research has been conducted on adapting a system to a particular speaker. Conventional hidden Markov model (HMM) adaptation methods can be divided into three families: linear transformation family, such as maximum likelihood linear regression (MLLR) [7]; Bayesian learning family, such as maximum a posterior (MAP) [8]; and speaker space family, such as Eigenvoices [6]. Depending on available data from a test speaker, adaptation algorithms usually estimate a limited number of parameters to obtain precise description of the speaker.

In recent years, a promising speaker-adaptation method, speaker selection training (SST), has emerged in the literature [4]-[5], [9]-[11]. SST selects a subset of cohort speakers from training speakers, and builds a speaker-adapted (SA) model based on these cohorts. In general, SST is a two-stage process: cohort speaker selection and model generation.

SST can make efficient use of very limited adaptation data. For example, given one adaptation utterance of several seconds, MLLR or MAP can hardly achieve improvements. However, the data may be a good enough index to select acoustically similar speakers from a pool of training speakers (recall the fact that very few data can achieve excellent accuracy on speaker recognition [12]). This motivates to make full use of the statistics from a selected speaker subset.

There are various practical implementations of SST. In the first stage, selecting cohorts, the key issue is to define similarity measure. In [10], adaptation data from the test speaker are fed to the SA HMMs of all training speakers to calculate likelihood score so as to similarity measure. In [9] and [11], likelihood scores from Gaussian mixture model (GMM) are used instead. In the second stage, there are various options, such as HMM retraining, MAP adaptation in [9], data transformation in [10] and model combination in [11]. However, retraining an SI model by data of selected cohorts is very time-consuming. Model combination is much faster, because only pre-calculated statistics are used in [11].

In this paper we propose a 3-stage adaptation framework based on speaker selection as proposed in [4]: speakers selection at the first step, then model transformation, and finally model combination. Experiments show that compared with the baseline, over 14% relative error rate reduction can be achieved when only 10 utterances (about 30s) are available for adaptation.

Compared with our former work [4] [5], where only speaker selection and/or model combination strategies are used, the main difference in this paper is the introduction of model transformation before model combination. The motivation is illustrated in Fig. 1. When all selected cohorts are close enough to the test speaker, this transformation step may not be necessary (e.g. speaker A in Fig. 1). However when there is a larger difference between training speakers and the test speaker, e.g. due to accents, transform cohorts’ models closer to the test speaker can make the combination step more robust (like speaker B).

* Work carried out as visiting student at MSR Asia.
This paper is organized as follows. Section 2 describes the general 3-step adaptation framework first and then describes them in detail: GMM based speaker selection, model transformation, and model combination. Speaker-adaptation experiments are shown in Section 3. Conclusions and discussion are provided in Section 4.

2. 3-stage adaptation strategy

In this paper we propose a 3-stage adaptation framework: cohorts speakers selection, selected speaker model transformation, and finally model combination. The whole procedure is illustrated in Fig. 2. Before adaptation, a GMM is created for each training speaker. The training speakers’ HMMs are adapted from a SI system by MLLR.

The first stage is to select a subset of the speakers who are acoustically close to test speaker according to likelihood scores from GMM. Next the selected speakers’ models are transformed to match the test speaker better. Finally these transformed models are linearly combined (interpolating Gaussian mean vectors) to form the SA model for the test speaker. The interpolation weights are estimated from the adaptation data. The concept of model combination is similar to reference-speaker weighting [3], cluster-adaptive training [2] and cluster weighting [1], though these methods do not include speaker-specific subset selection. Our method can also be viewed as an extension of static model combination [11] by learning weights from data. Experiments shows 14.9% relative improvement over the baseline SI system was achieved when only 10 utterances (about 30s) are available for adaptation. Details are discussed for each step as follows.

2.1. Speaker selection

The first stage of the procedure is to find a cohort of training speakers who are closest to the test speaker. The HMM based likelihood score is consistent with recognition stage and widely used in some systems. However it has been observed in [4] that an HMM adapted with 200 utterances may not be reliable in selecting cohort speakers. The reason may be that HMM focuses more on modeling probability distribution of phonemes instead of speakers.

On the other hand, Gaussian Mixture Models (GMMs) are successfully used in state-of-the-art speaker recognition systems. In other words, GMMs may be more efficient to represent a speaker’s characteristics. Based on this, we use a GMM to model each training speaker as follows

\[ b(O) = \sum_{i=1}^{M} c_i G(O, \mu_i, \Sigma_i) \]  

(1)

where \( b(O) \) is the output probability of observation sequence \( O \), \( c_i \) is the weight for \( i \)-th mixture component, \( G \) is a Gaussian function with mean vector \( \mu_i \) and covariance matrix \( \Sigma_i \). In practice, diagonal covariance matrix is usually used and referred as vector \( \sigma_i^2 \).

Given a certain amount of data, the posterior probability for mixture component \( k \) is given by

\[ p(k | o(t), \Lambda_n) = \frac{c_k G(o(t), \mu_k, \sigma_k^2)}{\sum_{i=1}^{M} c_i G(o(t), \mu_i, \sigma_i^2)} \]  

(2)

The EM re-estimation formulae for GMM can be found in [12].

A test speaker’s adaptation data is fed to all GMMs to calculate the likelihoods. The training speakers with the \( K \) largest likelihood scores are selected as cohorts.

2.2. Transformed models for training speakers

After cohorts are selected, the information they carry can be used for speaker adaptation in several ways, such as retraining the SI model with the data of cohorts [4] and combining speaker dependent (SD) models of cohorts with pre-calculated statistics.

In this paper we make use of the cohorts’ statistics at model level. Training speakers’ models are adapted from the SI model in advance. Specifically, each training speaker’s initial model is adapted from SI system using the MLLR algorithm [8]. This step is accomplished offline beforehand. Given selected cohorts we can combine their models to obtain an SA model for the test speaker [5]. However the adaptation results depend on the relative distribution between training

---

1 The following notation is used: capital bold letters refer to matrices, e.g., \( M \), bold letters refer to vectors, e.g., \( \mu \), and scalars are not bold, e.g., \( m \).
and test speakers. If the test speaker is “far” from all training speakers, the characteristics of selected cohorts do not match that of the test speaker. Directly combining these not-so-similar models may not be optimal.

Consequently before model combination we transform the cohorts’ models to the test speaker in order to reduce the difference between the cohort speakers and the test speaker. For simplicity, the transform is also based on MLLR. It is expected that a combination procedure based on more compact models would obtain better performance. As illustrated in Fig.1, we can not change the mean of the final model; however, combination based on transformation model will help to reduce the covariance of the final model, similar to speaker adaptation training.

2.3. Model combination

After we select the cohort speaker and transform them according to the test speaker, we get more and more close model to the test speaker. However, since they don’t match the target equally well, we need to combine them with different weights to measure their relative contribution to the test speaker’s model.

Assume there are $R$ selected cohorts with their transformed models. For simplicity, a global weight vector is learned for all phoneme 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 the test speaker, $\mu^m$, is given by

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

where $M_m$ is the matrix formed by $R$ cohort mean vectors for component $m$. And $\lambda$ is the weight vector given by

$$\lambda = [\lambda_1, \ldots, \lambda_R]^T$$

Here $\lambda$ can be set according to the prior information

$$\lambda_r = \frac{\sum_{m=1}^M \sum_{r=1}^R \gamma_{mr}(t,r)}{\sum_{r=1}^R \sum_{m=1}^M \sum_{r=1}^R \gamma_{mr}(t,r)}, \quad r = 1, \ldots, R$$

where $\gamma_{mr}(t,r)$ is the posterior probability of Gaussian $m$ in cohort model $r$ at time $t$.

Weight vector $\lambda$ can also be learned through ML or MAP by maximizing the likelihood or posterior probability of adaptation data from the test speaker given the combined model.

$$\text{argmax}_\lambda \sum_{m=1}^M \sum_{r=1}^R \log p(o(t)|\lambda_r) + \log p(\lambda)$$

Like MLLR adaptation and cluster weighting [2], a bias item can be introduced to consider the effect of channel condition,

$$\mu^m = M_m \lambda + b = M_m \tilde{\lambda}.$$  

The detailed deduction and solution for the weight vector can be found in [7].

2.4. Implementation issues

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

In reference-speaker weighting [4], two constraints are added to maximize the goal function (6):

$$\forall r, \quad \lambda_r \geq 0 \quad \text{and} \quad \sum_{r=1}^R \lambda_r = 1$$

Cluster-adaptive training as proposed by Gales 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 $\Sigma_k$ is used to guarantee the reliability of learning weight vector(s).

3. Results

The results presented in this section are to evaluate the proposed scheme on a speaker adaptation task of large vocabulary continuous speech recognition.

3.1. Experiment setup

An internal Microsoft dictation task is used to examine the proposed scheme. Although the experiments are based on Chinese Mandarin, the proposed adaptation scheme definitely can be applicable to other languages. About 70 hours of speech data from 250 Chinese male speakers (200 utterances each) are used to train a gender-dependent SI model. A state clustered decision tree system is used throughout with 6,000 tied states. There are 8 Gaussian components for each state. The front end consists of 36 cepstral coefficients, e.g. 12MFCC with their delta and accelerations. Tones of Mandarin are modeled together through pitch feature and a specially designed phoneme set. A trigram language model is used in all experiments with a 54k word vocabulary.

<table>
<thead>
<tr>
<th>%</th>
<th>baseline MLLR</th>
<th>N=5</th>
<th>N=10</th>
<th>N=20</th>
<th>N=30</th>
<th>N=40</th>
<th>N=50</th>
</tr>
</thead>
<tbody>
<tr>
<td>CER</td>
<td>10.4</td>
<td>9.6</td>
<td>10.1</td>
<td>10.1</td>
<td>10.0</td>
<td>9.8</td>
<td>9.9</td>
</tr>
<tr>
<td>Rel imp</td>
<td>-</td>
<td>6.8</td>
<td>2.7</td>
<td>2.5</td>
<td>3.1</td>
<td>4.8</td>
<td>4.4</td>
</tr>
</tbody>
</table>

Table 2: Model combination by learned weights thr MAP

<table>
<thead>
<tr>
<th>%</th>
<th>baseline MLLR</th>
<th>N=5</th>
<th>N=10</th>
<th>N=20</th>
<th>N=30</th>
<th>N=40</th>
<th>N=50</th>
</tr>
</thead>
<tbody>
<tr>
<td>CER</td>
<td>10.4</td>
<td>9.6</td>
<td>9.9</td>
<td>9.1</td>
<td>9.7</td>
<td>9.3</td>
<td>9.5</td>
</tr>
<tr>
<td>Rel imp</td>
<td>-</td>
<td>6.8</td>
<td>4.4</td>
<td>12.1</td>
<td>6.1</td>
<td>10.4</td>
<td>8.2</td>
</tr>
</tbody>
</table>

Table 3: Model transformation and combination
The same 250 training speakers are used for selection, typically one utterance, lasts 3 to 5 seconds, both for training and test. Before adaptation, a GMM with 64 mixture components is trained for training speaker with 200 utterances. Training speakers’ HMMs are adapted by MLLR with the same 200 utterances from the SI system.

The test set consists of 25 male speakers, 20 utterances each. 10 of them are used for selecting cohorts, transforming and combining cohort models. The other 10 are used for testing. The performance is evaluated in term of Character Error Rate (CER), which is equivalent to word error rate (WER) for e.g. English. In all the following experiments, one weight vector is estimated for each phone class, and pre-defined 65 phone classes according to Mandarin phonetic structure are used.

The gender-dependent SI model is chosen as our baseline.

3.2. Cohort selection and model combination

This experiment examines the effect of picking a subset of \( N \) close training speakers, and directly combining their models. No model transformation before combination is done here. Table 3 and Table 4 show the results of combining cohorts’ models by prior weights from (5) and learned weights through (6), respectively. It can be seen that recognition results benefit from selection stage. Fig. 3. Summary and comparisons among different kinds of speaker adaptation strategies vs. baseline. In addition combining through MAP learning significantly decreases error rates, providing a relative improvement of 12.1%, at best.

It is noted that 2-stage adaptation performance varies greatly (4% -12%) with number of cohort speakers.

3.3. Transforming the cohorts’ models before combination

This experiment examines the effectiveness of the model transformation stage. Table 3 shows the results with various numbers of cohorts. Compared with Table 2, we can see that transformation stage provides further improvement. The best results are achieved with 10 cohorts, giving 14.9% relative improvement over the baseline and 8.6% over MLLR. More important is that this step makes the adaptation procedure much more robust against the number of selected speakers. Varying the number of cohorts from 5 to 50, the adaptation improvements are consistently above 12%, much better than MLLR. Therefore, model transformation before combination is a very useful step since it is usually hard and heuristic to determinate the optimal number of cohorts beforehand.

4. Conclusions

In this paper, we have proposed a speaker adaptation scheme consisting of 3 stages: cohort speakers selection, speaker transformation and model combination. Likelihood scores from GMMs are adopted for selecting close speakers. Linear model interpolation is deployed to measure the relative contributions of cohorts to a test speaker in terms of phone classes. Cohorts’ models are then transformed closer to test speaker before combination to alleviate further the mismatch between cohorts and the test speaker. Our experiments have shown the effectiveness and reliability of the proposed scheme. The 3-stage adaptation scheme, especially by introducing the additional transformation step, can significantly improve the performance and robustness of adaptation, given that we currently have no good solution to select the optimal number of cohorts.

5. References