COHORTS BASED CUSTOM MODELS FOR RAPID SPEAKER AND DIALECT ADAPTATION

Jian Wu* and Eric Chang

Microsoft Research China
5F, Beijing Sigma Center, No. 49, Zhichun Road
Haidian District, Beijing 100080, China
jwu@csis.hku.hk echang@microsoft.com

Abstract

It is well known that speaker dependent acoustic models can achieve an error rate that is up to a factor of two smaller compared to well trained speaker independent acoustic models [1]. Thus, for improved accuracy, many modern dictation systems require the user to perform enrollment sessions to adapt the acoustic model of the system. In this paper, we present an approach that uses as few as three sentences from the test speaker to select N closest speakers (cohorts) from both the original training set and newly available training speakers to construct customized models. By using such an approach, our adaptation scheme can be updated online without re-configuring anything that has been determined or calculated before, i.e. speaker independent model. When applying this approach to address dialectal differences, the cohort based user specific models constructed with 3 user sentences can obtain a lower error rate even when compared to user-adapted models based on 170 user sentences. On average, a relative error rate reduction of 22% was achieved.

1. Introduction

Ways of dealing with speaker variability have been one of the most important research areas in speech recognition. Speaker differences can result from the configuration of the vocal cord and the vocal tract, dialectal differences, and speaking style. Previous approaches in addressing the speaker variability problem have included speaker adaptation, which modifies the parameters in the acoustic model according to some adaptation data [2], speaker normalization, which attempts to map all speakers in the training set to one canonical speaker [3], and speaker data boosting, which attempts to artificially increase the amount of speaker variability in the training database [4].

In order to meet the increasing requirement of practical system, one of the current research issues on adaptation is to use only a small amount of data from the test speaker while moving the parameter of the speaker-independent model toward speaker-dependent values.

* Work carried out as visiting student at MSR China. Currently at University of Hong Kong and can be reached at jwu@csis.hku.hk.

According to [8] and [9], the adaptation based on the technique of speaker clustering has been shown to be a successful solution for the sparseness of enrollment data. However, they need to determine the initial cluster definition or structure in advance, and the initial cluster definition needs to be reconstructed if more training data become available. This is not an easy process for real applications when training data can be collected gradually. Another class of solutions are based on the selection of reference speakers, i.e. Reference Speaker Weighting (RSW) [10] and eigenvoice approach [11]. They choose a few individual speakers as the reference speakers, and then use only few statistics, i.e. mean vectors or eigenvoices, to represent the reference speakers and construct the model for adapted speaker by a weighted combination scheme. Of course they are very efficient for implementation. But their successes will greatly depend on whether these statistics are sufficient for describing the distribution of the reference speakers. In other words, the results are very sensitive to both the choice of reference speakers and the accuracy of the estimation of the statistics.

The common idea of above approaches comes from the assumption that some speakers or speaker clusters of the training set are acoustically close to the test speaker. In this paper, agreeing with the assumption, we try to find a set of cohorts and collect their data to construct a new model, which is a custom model for the adapted speaker. But in order to improve the robustness of such approaches, the construction of the so-call custom model uses the training data itself instead of only few statistics. It is similar to that described in [5], but we simplify the estimation procedure to make it more feasible for real application. Another important difference lies in that the cohorts may not come from the training speakers for speaker independent (SI) models, which makes our adapted scheme updateable online with the collection of the data without reconstruct anything, even the SI model. To demonstrate the improvements from using this approach, we apply the proposed method to address the problem of dialect variation, where we allow the cohorts to come from speakers whose dialects are different from that used in the training of SI model.

This paper is organized as follows. Section 2 describes the algorithms for selecting the closest training speakers to the test speaker. In Section 3, we describe the algorithm of using data from the selected speakers for
improved recognition accuracy. Section 4 describes the experimental conditions and the experimental results. Lastly, we present our conclusion and future work in Section 5.

2. Methods for selecting cohorts

As mentioned in the first section, we should select a set of cohorts who are acoustically similar to the target speaker from a pool of candidates. Before performing the selection procedure, the parameters of the speaker adapted model for each “on-call” speakers are estimated by using the maximum likelihood linear regression (MLLR) technique of [2]. These adapted models can be considered as the approximation of the speaker dependent models for each of the “on-call” speakers.

Definitions on distances between the acoustic spaces of different speakers, used in speaker clustering, can also be used in the procedure of cohort selection. In our approach, since the enrollment data is very limited, we adopt only two criteria that can directly reflect the improvement in system performance to select cohorts. The first one is the accuracy of the enrollment data in a syllable recognition task. The second criterion is the likelihood of the adaptation data after forced alignment against the true transcriptions. In order to apply such criteria, the true transcript text corresponding to the enrollment data should be prepared as the reference. In practical applications, the adaptation sentences are usually predefined and a rejection mechanism is used to ensure that the new user speaks the enrollment sentences. Therefore, in this study, we assume that the correct transcriptions of the selection sentences are available. In cases where the content of enrollment is not available, “approximately true” transcriptions can also be obtained by decoding the enrollment data with the speaker independent model.

2.1 Syllable recognition task

In syllable recognition task approach, a recognition system using only the syllable tri-gram information and the acoustic model is developed to decode the enrollment data in order to get a high quality syllable transcription without too much influence by the lexicon. The system would run several times using specific dependent models of different speakers to decode the enrollment data. Then the syllable accuracy can be retrieved by applying a DP-based alignment procedure to compare each recognized result with the true transcript text. Finally, all of the “on-call” speakers are ranked in the order of syllable accuracy and then the top N speakers with the highest syllable accuracy are chosen as the cohorts to the test speaker.

2.2 Likelihood from forced alignment

With the likelihood criteria, the enrollment data is aligned against the true transcription text using each speaker dependent model to compute the acoustic likelihood. Once again, the speakers are sorted according to their likelihood and the top N speakers with the highest likelihoods are picked as the cohorts. It should be noted here that the enrollment data are not aligned against the transcription using speaker independent model to get the detail boundaries of phones or even syllables, as was done in [5]. Accordingly, more computation is necessary but the results are fairer to each speaker. Furthermore, since the likelihood can be calculated as a byproduct of the syllable recognition task, the final cohort list can be easily tuned according to both the syllable accuracy and the likelihood.

3. Building cohort based custom models

3.1 Constraints in practical system

After the cohorts who are similar to the test speaker are found, the data from them can be used to enhance the model of the test speaker in many ways (i.e. to perform the common adaptation technique, MAP or MLLR, on these data directly, or to adopt an adaptive training procedure). However, a practical system requires a quick but efficient algorithm to get the usable custom model. Therefore, a single-pass re-estimation procedure, conditioned on the speaker-independent model, is adopted in our approach. Such a procedure has several advantages. First of all, during the re-estimation, different weights can be easily added on the feature vectors of the different speakers according to their degrees of similarity to the test speaker. Secondly, the process of re-estimation would update the value of each parameter instead of only means as in most adaptation scheme. Thirdly, since the posteriori probability of occupying the $m$th mixture component, conditioned on the SI model, at time $t$ for the $r$th observation of the $i$th cohort, denoted by $L_{i}^{r}(t)$, has been computed and thus can be stored in advance, the one-pass re-estimation procedure would not consume much computation resources. The modified estimation formula may now be expressed as follows:

$$
\bar{\mu}_{i} = \frac{1}{N} \sum_{n=1}^{N} \sum_{r=1}^{R} \sum_{t=1}^{T} (L_{i}^{r}(t) \cdot O_{i}^{r}(t)) \cdot \frac{1}{R} \sum_{r=1}^{R} L_{i}^{r}(t) - \frac{1}{R} \sum_{r=1}^{R} Q_{i}^{r}
$$

(1)

where $L_{i}^{r}(t) = \sum_{n=1}^{N} \sum_{r=1}^{R} L_{i}^{r}(t)$, $Q_{i}^{r} = \frac{1}{N} \sum_{n=1}^{N} \sum_{r=1}^{R} L_{i}^{r}(t) \cdot O_{i}^{r}(t)$, and they can be stored in advance. $O_{i}^{r}(t)$ is the observation vector of the $r$th observation of the $i$th speaker at time $t$, $\bar{\mu}_{i}$ is the estimated mean vector of the $m$th mixture component of the target speaker. The variance matrix and the mixture weight of the $m$th mixture component can also be estimated in a similar way.

3.2 Algorithm description

The algorithm for constructing the cohort based custom model is summarized in Figure 1, and consists of the following steps. Firstly, construct speaker dependent
models for each speaker in the on-call list by applying the MLLR technique. Those on-call speakers may not be included in the training set for the general independent model and thus the on-call list can be adjusted according to the requirement of target speaker. This also has the benefit of incrementally updating the pool of “on-call” speakers as the user population increases without having to release new speaker independent models. Secondly, cohorts will be selected after ranking all of the on-call speakers in the order of syllable accuracy or likelihood based on a few sentences of the test speaker. Finally, the custom model is obtained by re-estimating the parameter of SI model using data from cohort speakers.

4. Experimental results

4.1 System description
Microsoft’s Mandarin Whisper system, described in more detail in [7], was used as the baseline system. The baseline system uses a large phone set with syllable initial and tone dependent syllable final models. Since the goal of this paper was to study possible improvements to acoustic models through cohort selection, a tonal syllable recognition task was used for all experiments.

4.2 Database
The baseline acoustic model was trained using a training set consisting of read sentences collected from 250 male speakers, most of whom have a Beijing dialect. Each speaker spoke 200 sentences and some poorly recorded waveforms were not used, resulting in approximately 50,000 waveforms used during training. For cohort selection, a training set consisting of 105 of the Beijing dialect training speakers and 145 Shanghai dialect training speakers were used. For evaluation, two testing sets were created. One was a male testing set consisting of 80 sentences for evaluation and 3 sentences for enrollment, each from 6 Shanghai dialect speakers. For comparison purposes, recordings from 6 male Beijing dialect speakers were also used.

4.3 Results

The experimental results using the custom model for speakers with Shanghai dialect based on different criteria for cohort selection are shown in Table 1. N is the number of selected cohorts. The symbol Acc. indicates that the cohorts are ranked and selected according to the accuracy of recognized syllables. The symbol LL. indicates that the cohorts are ranked and selected according to the likelihood. The error rate obtained with the SI model is around 24% while the relative reduction in error rate by the adapted model using 170 adaptation sentences for each speaker is about 19%, which indicates that the mismatch between the test speakers and the SI model is very large. Comparing the results of different criteria for cohort selection, we can conclude that the likelihood approach is more stable for finding the closest speakers though both of them work well. The results also show that the cohorts based custom model can achieve better performance than SA model in the case where the great mismatch exists between the training and test speakers. Figure 2 provides the individual performance.

Table 1. Experimental results on six speakers with Shanghai dialect (%).

<table>
<thead>
<tr>
<th>SI (Err.)</th>
<th>Selection Criteria</th>
<th>N=10</th>
<th>N=20</th>
<th>N=50</th>
<th>N=100</th>
<th>SA</th>
</tr>
</thead>
<tbody>
<tr>
<td>23.85</td>
<td>Err.</td>
<td>Acc.</td>
<td>23.63</td>
<td>20.91</td>
<td>19.87</td>
<td>19.05</td>
</tr>
<tr>
<td></td>
<td>Rate</td>
<td>LL.</td>
<td>22.77</td>
<td>20.52</td>
<td>19.05</td>
<td>18.65</td>
</tr>
<tr>
<td></td>
<td>Rel.</td>
<td>Acc.</td>
<td>0.9</td>
<td>12.33</td>
<td>16.69</td>
<td>20.13</td>
</tr>
<tr>
<td></td>
<td>Impro.</td>
<td>LL.</td>
<td>4.53</td>
<td>13.96</td>
<td>20.13</td>
<td>21.80</td>
</tr>
</tbody>
</table>

Table 2. Experimental results on six speakers with Beijing dialect (%).

<table>
<thead>
<tr>
<th>SI (Err.)</th>
<th>Selection Criteria</th>
<th>N=10</th>
<th>N=20</th>
<th>N=50</th>
<th>N=100</th>
<th>SA</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.65</td>
<td>Err.</td>
<td>Acc.</td>
<td>11.68</td>
<td>10.89</td>
<td>9.84</td>
<td>9.51</td>
</tr>
<tr>
<td></td>
<td>Rate</td>
<td>LL.</td>
<td>10.95</td>
<td>10.45</td>
<td>9.50</td>
<td>9.25</td>
</tr>
<tr>
<td></td>
<td>Rel.</td>
<td>Acc.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.45</td>
</tr>
<tr>
<td></td>
<td>Impro.</td>
<td>LL.</td>
<td>-</td>
<td>-</td>
<td>1.55</td>
<td>4.15</td>
</tr>
</tbody>
</table>

Figure 2. Comparison between Custom Model and SA Model. s1,...,s6 are six Shanghai dialect speakers.
of the six speakers. It shows that the improvements are significant for all the test speakers.

Another trend observed from Table 1 is that better performance can be achieved by using more cohorts. However, it’s clear that with a limited training set, performance should not improve continuously with N since when N equals to the number of training speakers, then the custom model is essentially the SI model (assuming that test speaker specific weights had not been applied in the re-estimation.) This implies that an optimal point for the value of N must exist and the appropriate value of N for a given pool of “on-call” candidates is an open problem that will be focused on in future work.

Table 2 shows the result of the same experiment performed on six speakers with Beijing dialect, when the dialect of the test speakers are much closer to the dialect of the training set for the speaker independent model. The average error rates produced by the SI model is only 9.65%, which indicates that the mismatch between the test set and the training set is small.

Compared with Table 1, the relative improvements are not so significant. This result is reasonable because the SI model has described the acoustic space occupied by the test speakers very well. We should also note that the proposed approach can still produce comparable relative error reduction with SA model even in such cases when the cohort size is sufficiently large.

5. Conclusion

It is well known that speaker dependent acoustic models can achieve an error rate that is up to a factor of two smaller compared to well trained speaker independent acoustic models [1]. Thus, for improved accuracy, many modern dictation systems require the user to carry out enrollment sessions to adapt the acoustic model of the system. However, the large number of sentences required during enrollment discourages the use of dictation systems. In this paper, we present an approach that uses as few as three sentences from the test speaker to select N closest speakers (cohorts) from the training set to construct customized models. When applying this approach to address dialectal differences, the cohort based user specific models can obtain a lower error rate even when compared to user adapted models based on 170 sentences. On average, a relative error rate reduction of 22% was obtained.

By using this approach, it becomes possible to provide customized acoustic models for each new user as a service. Since the potential cohort database can be continuously enlarged, it is possible to selectively collect more training speakers for a subset of users who are having higher error rates without shipping new speaker independent models. In the future, we plan to study the scalability of the current approach as the number of cohort candidates is increased. For example, if thousands of cohort candidates are available and hundreds of cohorts are selected, the amount of improvement in recognition accuracy may continue to improve. However, cohort specific weighting may be necessary to maximize the benefit of the cohort database.

6. Acknowledgement

The authors gratefully acknowledge assistance provided by Chao Huang, Jianlai Zhou, and others in the group for their support and fruitful discussions.

7. References


