SPEAKER SELECTION TRAINING FOR LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

Chao Huang, Tao Chen [+] and Eric Chang

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

ABSTRACT

Acoustic variability across speakers is one of the challenges of speaker independent (SI) speech recognition systems. As a powerful solution, dominant speaker adaptation technologies such as MLLR and MAP may become inefficient because of the lack of enough enrollment data. In this paper, we propose an adaptation method based on speaker selection training, which makes full use of statistics of training corpus. Relative error rate reduction of 5.31% is achieved when only one utterance is available. We compare different speaker selection strategies, namely, PCA, HMM and GMM based methods. In addition, impacts of number of selected cohort speakers and number of utterances from target speaker are investigated. Furthermore, comparison and integration with MLLR adaptation are also shown. Finally, some ongoing work such as dynamically varying number of selected speakers, measuring the relative contribution among the selected speakers and speeding up the computationally expensive procedure of re-estimation with model synthesis are also discussed.

1. INTRODUCTION

Acoustic variability across speakers due to physiological differences is still one of the challenges in state of the art SI speech recognition system. Speaker adaptation technology is a powerful approach by adjusting the parameters of SI model to maximize the likelihood of given observations of the target speaker. However, its effectiveness depends on the amount of data available from the speaker. When the given corpus is very limited, e.g. one utterance of three seconds, even the dominant adaptation methods such as MAP and MLLR [1] do not work well. They may degrade the performance because there is bias in parameter estimation due to too little data. In this paper, we propose a speaker adaptation method based on speaker selection training strategy. It is hard to train a transformation matrix for MLLR with one adaptation utterance, and it is even more difficult to estimate the parameters with MAP. However, the data may be a good index to select the reference speakers from many training speakers. One reason is that in speaker recognition field, 3 seconds of utterance has achieved very good accuracy on speaker identification. Then we can make full use of the statistics from the selected speaker subsets.

The main procedures behind the method consist of the following steps when given limited data from target speaker and a certain amount of training corpora.

1. Efficient speaker representations corresponding to feature selection in pattern classification.
2. Reliable similarity measure criteria between speakers.
3. Number of close speakers (cohorts) to be picked out.
4. Number of utterances from target speaker when considering the performance and enrollment efforts.

According to [2], we have tried two representations of speaker, namely transformation matrix based and model based. The model based methods used either GMM or HMM. As to the distance measurement, we have adopted Euclidean distance and likelihood based one. When considering the representations and measurements, we have obtained three kinds of cohort speaker selection strategies:

1. Euclidean distance after projecting the MLLR transformation matrix into low dimension space using principal component analysis (PCA).
2. GMM plus likelihood measurement (GMM).
3. HMM plus likelihood measurement (HMM).

In addition, we have investigated the effects of varying the number of cohort speakers and the number of utterances of target speaker by experiments. It is shown that they are two important factors in practice.

There is some related work in speaker clustering based adaptation methods. Wu et al. [3] proposed HMM based cohorts selection to customize the speaker model. Both accuracy and likelihood are used as measurements. The main disadvantage is its huge computation cost and space to store the speaker dependent (SD) model in the selection procedure. In addition, measurement based on recognition accuracy was found to be less reliable because of fewer observations. Padmanabhan [4] also adopted the HMM based selection method and the HMMs are obtained by MAP from SI model, compared with MLLR used in [3]. The main improvement mentioned by [4] owes to transforming the data of selected speakers closer to the target speaker in addition to the selection itself. As an extension of [4], Gao [5] introduced pre-clustering procedure and reduced the model sets greatly by replacing the SD model with cluster dependent model. GMM based speaker selection was used by Matrouf [6]. Since a very small subset of cohort speakers was picked out, MAP instead of re-training was used to adjust the

[+] Work carried out as visiting student at MSR Asia.

0-7803-7402-9/02/$17.00 ©2002 IEEE

baseline model. One problem in [6] is that both the test corpus and enrollment corpus share the same set. Therefore, it is hard to evaluate the performance when both selection and standard MLLR adaptation procedure are unified. In [7] GMM based speaker selection strategy and sufficient statistics from SI model were combined for fast speaker adaptation.

This paper is organized as follows: in the next section, we will describe the speaker selection methods in detail. In Section 3, evaluation experiments are conducted to compare the efficiency of adaptation based on different speaker selection strategies. Effects of different number of cohort speakers and target speaker utterances are also presented. Comparison among the proposed scheme, MLLR and their integration is shown in this section too. Conclusions and detailed discussions, including some ongoing work, are given in Section 4.

2. SPEAKER SELECTION TRAINING METHODS

There are several key issues in building cohorts based speaker adaptation models. One is the appropriate speaker representation method. The other is an efficient and reliable measurement of similarity between target and training speakers. In practice, when considering the data available, there are another two important issues that will affect the final performance: 1) how many utterances from the target speaker are sufficient to be used to effectively select his/her cohorts; 2) how many cohorts should be picked out to create the speaker customized model.

2.1. Speaker Representation

2.1.1. Representation based on MLLR matrix

In [2] we suggested using MLLR matrix to represent a speaker, and then using PCA to investigate the inter-speaker variability. Here we give a short description. More detailed information can be found in [2].

After MLLR adaptation, for each speaker, we obtain several transformation matrices plus offsets; each pair corresponds to a phone (or phone class). Then a set of supporting phone classes which are most typical to characterize the speaker are selected. In our system, only the offset vectors of the supporting classes are concatenated to form a big vector to represent a certain speaker. These vectors may be of high dimension. We put these vectors of all training speakers to form a matrix and perform PCA to reduce the dimension. It is found that this method is quite efficient for describing the characteristics of a speaker. Some other dimension reduction methods, such as independent component analysis (ICA), can be also used in the procedure. Here only PCA is investigated in selecting cohort speakers.

2.1.2. Representation based on models

The idea of using models to represent speakers is somewhat easy to understand. In speech technologies, HMM and GMM are widely and successfully used. They are able to model the main characteristics of a speaker.

GMM based representations ignore the phone differences and the parameters that are needed to be estimated are much less than HMM based one. As a result it does not require the transcriptions of the utterance and the computation cost and storage space for GMM are much less. HMM based method may describe a speaker in detail when given enough data. When applying HMM model to force-align the adaptation data to select cohort speakers, we have to know the transcriptions in advance, though we may obtain it through SI system. Performance comparison of these methods will be described when they are used to select cohorts in the next section.

2.2. Speaker Similarity Measurement

According to the methods to represent speakers, we classify the schemes for selecting cohorts into two categories and evaluate their performance in speaker adaptation.

2.2.1. Distance measurement on MLLR matrix

While performing PCA, a transformation matrix is learned to project the original matrix to a lower dimensional one. We represent a test speaker with the same MLLR offset vector and project it to lower dimension space with the same transformation matrix learned from training speakers. Then weighed Euclidean distance (the eigenvalue corresponding to each dimension is used as weight) in the low dimension is used as the similarity measurement between the test and training speakers. Top M training speakers with the smallest distances are selected as cohorts. Then a simple re-training procedure is performed to obtain the speaker adapted model.

2.2.2. Likelihood measurement on models

In likelihood based measurement method, one model is built for each training speaker. Then the adaptation data of test speaker is fed to the model of each training speaker. The likelihood score of the observations is used as the similarity measure. While no transcriptions are needed when using GMM models, they are needed for HMM models. In our system, the likelihood from HMM is obtained through forced alignment of the adaptation data against each SD model.

2.3. Some Practical Issues

In practice there are some issues we should consider with the speaker selection training method. One is determining the proper number of cohorts for each test speaker. This problem is very critical when there are thousands of training speakers available. How to make trade-off between data coverage (the more data, the more reliable the estimation of model parameters) and similarity (the more cohorts selected, the more risk to incorporate acoustically more dissimilar speakers and finally enlarge the variance of the model and reduce the discriminative ability)? Another problem is how much adaptation data is enough to reliably select cohorts. We will discuss these two problems by experiments.

3. EXPERIMENTS AND ANALYSIS

The baseline SI system consists of two gender dependent models trained on about 120 hours of speech data using Microsoft's Mandarin Whisper system [8]. The training speakers ready for selection are the same ones used for SI model training, consisting of 500 speakers and 200 utterances per speaker. After the cohorts are selected, all utterances from them are fed to the corresponding baseline system to re-estimate the model parameters. Character error rate (CER) is used as evaluation in

all experiments.

For the test corpus, we have 50 speakers (25 female and 25 male) and 20 utterances per speaker. 10 utterances per speaker are used to select cohort speakers for experiments in Section 3.1, 3.2 and 3.4. Test speakers have the same accent as the training speakers. That is, relatively small variation exists between the SI model and the test speakers. This is rather a tough task, as it is known that adaptation method is less effective in this situation than where there is larger mismatch between test speaker and SI model [3].

First, we evaluate the performance of different methods to select cohorts. Then we investigate the impact of number of cohorts and number of enrollment utterances. Finally we provide comparison of our scheme with the standard MLLR adaptation.

It should be noted that training speakers ready for selection are not separated by gender. In other words, cohorts who are picked out may have different gender from the target speaker.

3.1. Comparison of Different Selection Schemes

When considering both speaker representations and similarity measurement, we have experimented with three kinds of speaker selection strategies: PCA, HMM and GMM based methods. The details of methods can be found in Section 1 and Section 2. Because there is not enough data to train a SD model for each speaker, HMM of each speaker is obtained through the MLLR adaptation based on the SI model. And 64 Gaussian mixtures are trained [10] for GMM. We can conclude from Table 1 that GMM based measure obtains the best recognition accuracy on average with 3.24% relative error reduction. Although PCA achieved best accuracy in female test speakers and HMM achieved best result in male ones, GMM is more reliable because of the consistent result on both genders. Now let us have a look at the computation and storage cost for these three methods. PCA based one is much faster since it needs only Euclidean distance computations of two vectors in low dimension (e.g. 6). HMM based one is much slower since it has to compute the likelihood of all pre-trained SD model on the enrollment corpus. Computation cost of GMM ranks in the middle. The storage of GMM based method is far less excessive than that of both PCA and HMM based ones (about 70% space reduce) because the latter two have to store the baseline SI model in order to get the MLLR matrix. Furthermore, GMM does not need the transcriptions in advance although both PCA and HMM may be realized this through SI system.

CER (%)	Female	Male	Average	Rel. Err. Reduction (%)
Baseline	9.86	9.87	9.87	--
PCA	9.45	10.03	9.74	1.32
HMM	9.92	9.45	9.69	1.82
GMM	9.49	9.61	9.55	3.24

Table 1: CER of adaptation model using HMM and GMM based likelihood criterion. 100 cohorts per target speaker are selected for adaptation.

3.2. Impact of the Number of Cohorts

In cohorts based adaptation method, it is difficult to determine the number of cohorts we should use. We investigated fluctuating performance when varying the number of selected cohort speakers. Since GMM based scheme is the most robust among the three, we will focus our study on GMM based selecting method in all the following experiments.

CER (%)	Baseline	50	100	150	200	250	ROVER
Female	9.86	10.26	9.49	9.62	9.76	9.57	8.73
Male	9.87	9.62	9.61	9.25	9.29	9.32	8.53
Average	9.87	9.94	9.55	9.44	9.53	9.45	8.63
Rel. Err. Reduction (%)	--	-0.71	3.24	4.36	3.44	4.26	12.56

Table 2: Impact of number of cohorts on recognition accuracy.

Table 2 shows when only 50 cohorts are selected, even worse recognition accuracy is obtained on female test speakers. The average lowest error rate is achieved when 150 cohorts are selected. Too many cohorts will bring in some training speakers who are acoustically far away from the test speaker, which degrades the accuracy. One interesting finding is that when increasing the number of cohorts further, the error rate decreases a little, probably due to a better coverage. Since GMM can classify gender with error rate less than 1%, selecting 250 cohorts is similar to one more iteration training on gender dependent model, which is hoped to achieve more accurate and reliable estimation of the HMM parameters because of improved coverage on phone context.

Experiments also show the optimal number of cohorts for each test speaker is different. Some fluctuating performance in Table 2 also confirms this observation. If applying Recognizer Output Voting Error Reduction (ROVER) criteria to the recognition results from models trained on different number of cohort speakers, we obtain relative error reduction of 12.56% on average. It is shown that dynamically choosing the optimal cohort speakers for each target speaker is one of the key concerns in order to keep the balance between good coverage of phone context and acoustic similarity to the target speaker.

3.3. Impact of the Amount of Adaptation Data

Typical adaptation methods, such as MLLR and MAP, achieve rather low improvement over baseline system when only very few adaptation data are available, especially when the variation between SI model and the test speaker is small. In this section, we will explore the impact of the amount of adaptation data that are used for selecting cohort speakers.

CER (%)	Baseline	1	3	10
Male	9.87	9.35	9.33	9.25
Rel. Err. Reduction (%)	--	5.31	5.54	6.31

Table 3: Impact of number of utterances on recognition accuracy (150 cohort speakers are selected for each test speaker).

We select N utterances per test speaker and average their likelihood in each GMM. The averaged likelihood is used as distance measure to select cohorts. From Table 3 we can see GMM based speaker selection is quite robust against utterance variations. Only 1 utterance can achieve 5.31% relative error rate reduction. It is suggestive when deciding the proper adaptation corpus in order to consider both enrollment efforts and final performance gains.

3.4. Comparison with the MLLR Adaptation

To compare with the well-known MLLR [1], we randomly picked out 10 utterances from each test speakers and use them as the enrollment corpus for both speaker selection based and MLLR based adaptation. Enrollment data are excluded from the test corpus. In other word, the rest 10 utterances per speaker, forming 500 test utterances of all 50 test speakers, are used as the evaluation sets. Baseline is still gender dependent SI models. We have tried three different adaptation schemes: adaptation based on speaker selection, MLLR, and a combination of them. 10 utterances instead of one are used for MLLR because the latter is far from enough to estimate a global transformation matrix for MLLR with diagonal form. However, we know from Section 3.3 that it has little impact on the proposed method. From Table 4, adaptation based on speaker selection has achieved a comparable result with MLLR. More importantly, MLLR adaptation based on adaptation result of speaker selection obtains more gains than either of them used alone. It is also interesting that for female test speakers, although speaker selection based adaptation gets little improvement itself, while combined with MLLR, it helps achieve more improvement than MLLR alone. Results demonstrate that these two adaptation methods, speaker selection based and MLLR are complementary to each other. It may be explained by the hypothesis that speaker selection based adaptation captures information that are more readily available but less specific to the test speaker while MLLR seeks the specific, accurate but insufficient knowledge about the test speaker from a small corpus.

CER (%)	Baseline	+Speaker Selection	+MLLR	+Speaker Selection + MLLR
Female	9.54	9.51	9.15	8.95
Male	10.35	9.51	9.64	9.21
Average	9.95	9.51	9.40	9.08
Rel. Err. Reduction (%)	--	4.42	5.53	8.74

Table 4: Comparison with the standard MLLR alone adaptation and integration of both speaker selection (150 cohorts) and MLLR adaptation. (10 utterances per speaker)

4. CONCLUSIONS AND DISCUSSIONS

In this paper, we propose a speaker selection training strategy in order to realize efficient speaker adaptation when given very limited data when standard adaptation is not efficient any more. Behind the algorithm, MLLR transformation based and model based (include both HMM and GMM) speaker representations and correspondingly, Euclidean distance after PCA and likelihood based measure have been investigated. Results show that GMM plus likelihood based speaker selection method is more efficient and reliable. Additionally, it may be implemented with unsupervised mode. Compared with HMM based method, it is much more computationally efficient.

In addition, we have investigated the practical issues of applying such methods, the impact of the number of cohort speakers and the number of utterances from target speaker used to select the cohort speakers. It is shown that GMM based selection is very efficient even when only one utterance of about 3s is available.

Further comparison with well-known MLLR tells us that the proposed adaptation method is not only efficient, but also can also improve MLLR adaptation performance further.

Keeping number of cohort speakers is a very trick problem in order to make a tradeoff between good coverage and small variance among observations across cohorts. It is still relied on the experiments. Dynamic instead of fixed number of close speaker selection seems to be a good alternative.

One problem we have not discussed in this paper is how to use the corpus of selected speakers. The simplest way is to use all the corpora from the cohorts to re-train the SI model as we used in the paper. However, the main disadvantages of such method are that computation cost of re-estimations procedure when cohort speakers are decided. We are investigating the fast model fusion method. That is: we can offline compute the speaker adaptation model for each speaker in the training corpus. Given some utterances from incoming speakers, we can online synthesize the adapted model which is the interpolation of cohorts' SD models and the weight can be learned through ML framework, which is similar to reference speaker weighting described in [9] where no close speaker selection is used. The weights can be phone-dependent or not. In such a way, we not only select close speakers, but also can measure the relative contribution of each phone model of each specific speaker to the final output model.

6. REFERENCES

[1] C. J. Leggetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", *Computer Speech and Language*, vol. 9, n2, pp. 171-185, 1995.

[2] C. Huang, T. Chen, S. Li, E. Chang and J. L. Zhou, "Analysis of Speaker Variability," in *Proc. Eurospeech2001*, vol.2, pp.1377-1380, 2001.

[3] J. Wu and E. Chang, "Cohorts Based Custom Models for Rapid Speaker and Dialect Adaptation," in *Proc. Eurospeech2001*, vol. 2, pp. 1261-1264, 2001.

[4] M. Padmanabhan, L. Bahl, D. Nahamoo and M. Picheny, "Speaker Clustering and Transformation for Speaker Adaptation in Speech Recognition Systems", *IEEE Trans. Speech and Audio Processing*, vol. 6, n1, pp. 71-77, 1998.

[5] Y. Q. Gao, M. Padmanabhan and M. Picheny, "Speaker Adaptation Based on Pre-clustering Training Speakers", in *Proc. Eurospeech1997*, vol. 4, pp. 2091-2094, 1997.

[6] D. Matrouf, O. Bellot, P. Nocera, et al. "A Posteriori and a Priori Transformations for Speaker Adaptation in Large Vocabulary Speech Recognition Systems," in *Proc. Eurospeech2001*, vol. 2, pp. 1245-1248, 2001.

[7] S.Yoshizawa, A. Baba, K. Matsunami, et al. "Unsupervised Speaker Adaptation Based on Sufficient HMM Statistics of Selected Speakers," in *Proc. ICASSP2001*, vol. 1, pp. 341-344, 2001.

[8] E. Chang, J. L. Zhou, C. Huang, S. Di, K. F. Lee, "Large Vocabulary Mandarin Speech Recognition with Different Approaches in Modeling Tones," in *Proc. of ICSLP2000*, vol. 2, pp. 983-986, 2000.

[9] T. J. Hazen, "A Comparison of Novel Techniques for Rapid Speaker Adaptation," *Speech Communication*, vol. 31, pp. 15-33, 2000.

[10] The HTK Toolkit: http://htk.eng.cam.ac.uk.