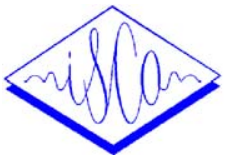

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

# REGRESSION CLASS SELECTION AND SPEAKER ADAPTATION WITH MLLR IN MANDARIN CONTINUOUS SPEECH RECOGNITION

*Chengrong Li Jingdong Chen Bo Xu*
National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing, 100080, P.R.China
Email: chengrong.li@nlpr.ia.ac.cn

10.21437/Eurospeech.1999-547

## ABSTRACT

Currently, CDHMM based continuous speech recognition has been widely extended to speaker-independent (SI) system. However, the performance of the SI system is highly dependent on the speakers, especially for Mandarin speech with accent, speaker adaptation becomes crucial important for real application. In this paper, MLLR approach is studied for speaker adaptation in mandarin continuous speech recognition and three approaches for defining regression classes are investigated: the first is based on Chinese phonetic classification, the second is based on statistical information of mixture distribution parameters and the third is based on state duration using segmental information. Other experiments like the effect of adaptation data and mixtures are presented also in the paper. The new variance-based regression class selecting scheme is proposed and has been proved to be effective.

Keywords: MLLR, regression class, speaker adaptation, Mandarin speech recognition

## 1. INTRODUCTION

The objective of speaker adaptation is to adapt the SI system to the SD system using as less adaptation data from unseen speaker as possible. There are two main categories of adaptation techniques to solve the mismatches between reference speakers and testing speakers that are respectively called speaker normalization and speaker adaptation. The former normalizes the testing speech and training speech in feature level to keep the consistence between them. The later is to adjust the parameters of the well trained models, usually of SI system, to fit for the new speaker. Though there are some model adaptation schemes like maximum a posteriori (MAP) that had been studied for many years, linear transformation has been shown to be a powerful tool for both speaker and environmental adaptation [3]. Among linear transformation approaches, the MLLR (Maximum Likelihood Linear Regression) has been proved to be the robust one and many research works have been done [2,3,4,5]. This method is superior to other methods in that it can adapt all models even if no model-specific data is available, while the methods like MAP only update those model parameters for which sufficient data is available. In this way, MLLR is more suitable for the cases that adaptation data is not enough or adaptation speed is specially of importance. In MLLR, regression class selection plays special parts in successful use of it. In the paper, we will investigate the regression class selection and speaker adaptation with MLLR for Mandarin continuous speech recognition based on CDHMM. A new method for selecting regression class based on statistical information of mixture distribution parameters is presented compared with basic methods such as global regression and one based on Chinese phonetic classification.

As we know, most of Chinese speech recognition systems are based on standard Mandarin and require speakers having no accent or very slight accent. High performance can be reached if a person speaks standard Mandarin and environmental condition is adequate, for example, 80% accuracy. of pure syllable decoding in our baseline system. But most of Chinese including overseas Chinese have accent or even very strong accent. In order to extend the use of speech recognition practically, current Chinese speech recognition systems have to be developed to adapt different speakers and wide range of environmental condition. Therefore, we collected a speech database where speakers have different accents. Some of them are used to test current recognizer and the results are shown in Table 4 which is not satisfied. The objective of the paper is to adapt speakers with accents with less adaptation data using MLLR speaker adaptation approach.

The rest of the paper is organized as follows. After giving an outline of the MLLR approach in section 2, the important issues for some regression class selection and MLLR configuration are discussed in section 3. The baseline system and experimental results of Mandarin continuous speech recognition using context-dependent acoustic models are presented in section 4. At last, in section 5, the conclusion is given.

## 2. MLLR APPROACH

The MLLR approach is based on linear transformation of well-trained CDHMM speaker-independent system. For the differences between speakers are mainly characterized in the estimates of the mixture component means, so in most MLLR based adaptation only means of mixture component are transformed and updated for the new speaker-specific models, all other parameters

such as the transition probabilities, mixture component weights, mixture component covariances, take their values from initial model set.

The estimates of the adapted mean for the adapting speaker are given by linear transformation

$$\mu = W\xi \quad (1)$$

$$\xi = [w \ \mu_1 \ \mu_2 \ \Lambda \ \mu_n]^T \quad (2)$$

where $W$ is a linear regression matrix ($n \times (n+1)$) which optimises a maximum likelihood objective function, $w$ is the offset term for the regressioin($w$ =1 for sdandard offset), $\xi$is the extended mean vector.

The transformation matrix $W$ is estimated by maximizing the likelihood of adapted models generating the adaptation data, that is, the objective function is as follows

$$\begin{aligned} F(O|\lambda) &= \sum_{q \in \Theta} F(O,q|\lambda) \\ &= \sum_{q \in \Theta} a_{q_T N} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(o_t) \end{aligned} \quad (3)$$

By defining an auxiliary function $Q(\lambda, \bar{\lambda})$

$$Q(\lambda, \bar{\lambda}) = \sum_{q \in \Theta} F(O,q|\lambda) \log(F(O,q|\bar{\lambda})) \quad (4)$$

and maximise $Q(\lambda, \bar{\lambda})$ with respect to $W_s$, the general form of the optimisation of $W_s$ can be got as equation (5).

$$\sum_{t=1}^{T} \gamma_s(t) C_s^{-1} o_t \hat{\xi}_s' = \sum_{t=1}^{T} \gamma_s(t) C_s^{-1} W_s \hat{\xi}_s \hat{\xi}_s' \quad (5)$$

When tying regression matrices, the summation should be performed over all tied states. Assuming regression matrix $W_s$ is shared by R states, the formula is as following

$$\sum_{t=1}^{T} \sum_{s=1}^{R} \gamma_s(t) C_s^{-1} o_t \hat{\xi}_s' = \sum_{t=1}^{T} \sum_{s=1}^{R} \gamma_s(t) C_s^{-1} W_s \hat{\xi}_s \hat{\xi}_s' \quad (6)$$

## 3. BASIC CONFIGURATION AND REGRESSION CLASS

### 3.1 Basic experiments configuration

There are many issues for MLLR implementation. Based on the preliminary theoretical and experimental analyses, following conditions are considered.

For the poor performance of adapting speakers with accents, unsupervised adaptation is not suitable for our first stage tasks. There are three adaptation schemes: static (or batch) adaptation, incremental adaptation and instantaneous adaptation. We use the first one which starts adaptation while all the adaptation data is available.

Because preliminary experiments have shown that further iterations had very little contributions to the adapted models, which indicates that the frame alignment for the adapted models is almost the same as that of original SI models, only one iteration of adaptation is performed in all given cases.

The single variable regression used for the diagonal matrix is clearly not powerful enough to capture many different variations within a class [2]. Much more classes are needed for diagonal transformation compared to full matrices. So adaptation using a small number of full matrices is superior to using many diagonal matrices. Therefore, full matrices will be used in later experiments.

### 3.2 Regression class

As above description, regression class selection becomes our key investigation.

In order to effectively estimate the regression transformation matrices, we take the clustering unit, which can be a model, a state, or a mixture component. In estimating regression matrix, all the clustering units that will be transformed with same matrix are tied into the same class. The problem is what clustering units should be put into same class, that is, the problem of choosing regression class. We make an assumption that clustering units with similar parameter values should be transformed in a similar manner. So one regression class will include all the clustering units representing similar acoustic phenomena.

Three approaches for choosing regression classes are presented. Besides of the basic regression class based on Chinese phonetic classification, two new methods are based on statistical information of mixture distribution parameters and state duration from Viterbi segmental information.

### 3.2 Implementation and evaluation

The approach is implemented using a Baum-Welch HMM training frame with forward-backward to determine the mixture component occupancy probabilities.

The data from silence model is not accumulated to evaluation of transformations and the silence model still takes its original parameters.

Seven adaptation speakers are used to evaluate the approach and proposed regression class. The speakers are arbitrarily chosen each having different accent. There are 250 adaptation sentences available and 60 sentences for testing.each speaker.

## 4. EXPERIMENTS

### 4.1 Baseline system

The speaker-independent Mandarin continuous speech recognition system is well trained before adaptation process using database DB863 with 116 male speakers each having about 500 sentences. The models take the

initials and the finals considering the co-articulation in the intra-syllable and inter-syllable.

The speech is coded into 25 ms frames. Each frame is represented by a 39 component vector consisting of 12 MFCCs plus energy, and their first and second time derivatives.

Syllable error rate, SER, is used to evaluate the effectiveness of adaptation methods which is got from a pure continuous syllable decoding based recognition system.

### 4.2 Regression class based on phonetic classification

Models are classified to the initials and finals according to Chinese phonetic class. So two regression matrices are estimated. This is the basic regression class choosing scheme which associates a set of models to a regression class not mixture components.

Compared with global transformation, this kind of regression class reduce the syllable error rate by 1.56 points in average(see Table 1).

Table 1 Reduction(%) of error rate by phonetic class

| | Speaker | Global | 2 classes |
|---|---|---|---|
| 1 | Gsh | 20.99 | 28.05 |
| 2 | Zsb | 7.90 | 3.41 |
| 3 | Yhui | 10.17 | 6.67 |
| 4 | Hfei | 16.92 | 14.26 |
| 5 | Chlei | 17.56 | 23.30 |
| 6 | Liumk | 16.33 | 18.89 |
| 7 | Licr | 4.40 | 10.59 |
| Average | | 13.47 | 15.03 |

### 4.3 Regression class on variance of mixture distribution means

Statistical information of model parameters –the means and variances for means and variances of mixture components is used to cluster the regression classes.

The classification is implemented by calculating the variance $V_m$ of the mixture means for all models except of the model silent.

Assuming that the variance of means for mixture i of model j is $v_{ij}$, then

$$V_{ij} = \begin{cases} class\,1 & \text{if } V_{ij} \le V_m \\ class\,2 & \text{others} \end{cases} \quad (7)$$

The results shown in Table 2 indicate that adding variance classifying after phonetic class clustering can present additional 2.91 points drop in average.

Table 2 Regression classes by variance of distribution means

| | Speaker | Baseline | Adapted | Error reduction |
|---|---|---|---|---|
| 1 | Gsh | 42.64 | 60.58 | 31.28 |
| 2 | Zsb | 78.62 | 80.07 | 6.78 |
| 3 | Yhui | 65.38 | 67.14 | 5.08 |
| 4 | Hfei | 39.19 | 50.75 | 19.01 |
| 5 | Chlei | 53.87 | 67.17 | 28.83 |
| 6 | Liumk | 54.68 | 65.09 | 22.97 |
| 7 | Licr | 55.26 | 60.46 | 11.62 |
| Average | | 55.66 | 64.47 | 17.94 |

### 4.4 State duration based on segmental information

Model or state duration is defined as the number of frames assigned to the model or state which is from Viterbi segmentation results. In our experiment, we use statistical information of state duration to choose regression classes. Two regression classes are gotten by the means of each state duration assuming that mean values which are close to each other present the similar transformation property. The results are shown in Table 3. The average error rate is a little lower than global transformation.

Table 3 State duration based on segmental information

| | Speaker | Baseline | Adapted | Error reduction |
|---|---|---|---|---|
| 1 | Gsh | 42.64 | 61.85 | 33.49 |
| 2 | Zsb | 78.62 | 77.66 | -4.49 |
| 3 | Yhui | 65.38 | 68.02 | 7.63 |
| 4 | Hfei | 39.19 | 45.66 | 10.64 |
| 5 | Chlei | 53.87 | 64.86 | 23.82 |
| 6 | Liumk | 54.68 | 63.24 | 18.89 |
| 7 | Licr | 55.26 | 58.27 | 6.73 |
| Average | | 55.66 | 62.79 | 13.82 |

### 4.5 Gaussian mixture components

To test the effectiveness of MLLR adaptation for different models, we experiment it on the models with 8 mixture distributions and 32 mixture distributions. The adaptation data is 8 sentences in general length. The result is shown at Table 4.

Different mixtures show consistent improvement of adaptation performance. The average reduction of syllable error rate with 32 mixture distributions is 18.02% which is more than with 8 mixture distributions. The probable reason is that the frame alignment for the models with 32 mixture distributions is more accurate.

### 4.6 The amount of adaptation data

To assess how the amount of adaptation data affects system performance, a global transformation (class) is used to adapt the baseline while varying the number of adaptation sentences from 3 to 40 (the average length of the sentences is about 12 seconds).

Table 4 MLLR with different Gaussian mixture components

| Speaker | Mixture=8 | | | Mixture=32 | | |
|---|---|---|---|---|---|---|
| | Baseline | adapted | Error reduce | Baseline | adapted | Error reduce |
| gsh | 42.64 | 58.73 | 28.05 | 49.65 | 58.96 | 18.49 |
| zsb | 78.62 | 79.35 | 3.41 | 82.38 | 85.50 | 17.71 |
| yhui | 65.38 | 67.69 | 6.67 | 69.30 | 73.63 | 14.10 |
| hfei | 39.19 | 47.86 | 14.26 | 43.87 | 54.34 | 18.65 |
| chlei | 53.87 | 64.62 | 23.30 | 55.84 | 67.40 | 26.18 |
| liumk | 54.68 | 63.24 | 18.89 | 60.32 | 67.63 | 18.42 |
| licr | 55.26 | 60.00 | 10.59 | 57.13 | 62.54 | 12.62 |
| Average | 55.66 | 63.07 | 15.03 | 59.78 | 67.14 | 18.02 |

The results in Fig 1 show that when regression class keeps unchanged the reduction (%) of syllable error rate seems to have several local maximuma. When the amount of adaptation data is less than certain value the reduction (%) of syllable error rate increases as the amount of adaptation data increases. But adding more adaptation data can not assure further improvement.

Practically 8 sentences are efficient for global transformation of Mandarin continuous speech recognition.

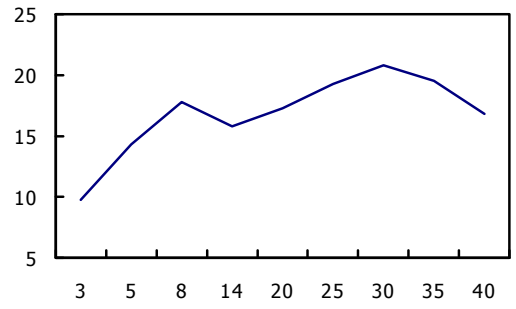


Fig. 1 Reduction sentences ıte varying with the amount of adaptation data

## 5. CONCLUSION AND FURTHER WORK

The speaker adaptation approach called maximum likelihood linear regression (MLLR) has been investigated for Mandarin continuous speech recognition. A set of SI models based on Chinese initials and finals are adapted to specific speaker by applying a set of linear transformation to the Gaussian mean vectors.

Three approaches of regression class selecting and other issues of MLLR are explored. The results show that regression class based on the variances of mixture distribution means is a effective classification method. Further work will investigate proposed methods with a large number of regression classes and adaptation data.

## 6. REFERENCES

[1] L. Lee, and R.C. Rose. Speaker normalization using efficient frequency warping procedures. Proc. ICAASP'96, 1:353-356, Atlanta.

[2] C.J. Leggetter, and P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech and Language, 9:171-185,1995.

[3] M.J.F. Gales. Maximum likelihood linear regression for HMM-based speech recognition. Computer Speech and Language, 12:75-98,1998.

[4] J.L. Gauvain, and C.H. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observation of Markov chains. IEEE Transactions on Speech and Audio Processing, 2: 291-298,1994.

[5] Q. Huo, and B. Ma. A new CDHMM adaptation method: being incremental, adaptive and more efficient. ISCSLP'98, 71-74, Singapore.

[6] M. Padmanabhan, L.R. Bahl et al. Speaker clustering and transformation for speaker adaptation in speech recognition systems. IEEE Transactions on Speech and Audio Processing, 6(1): 71-77,1998.

[7] L.E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. Inequalities, Vol.3:1-8,1972.

[8] B-H, Juang. Maximum likelihood estimation for mixture multivariate stochastic observation of Markov chains. AT&T Technical Journal, 64(6): 1235-1249,1985.