ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

TASK-SPECIFIC ADAPTATION IN CHINESE NAME RECOGNITION

Guo-Hong Ding# Bo Xu+ Xia Wang# Yang Cao# Feng Ding# and Yuezhong Tang#*

Nokia Research Center#, Beijing
High-Tech Innovation Center+, National Laboratory of Pattern Recognition*
Institute of Automation, Chinese Academy of Sciences, Beijing

ABSTRACT

In this paper, task-specific adaptation is proposed to improve Chinese name recognition performance. Since acoustic models are usually trained using large vocabulary continuous speech corpora, there exists distortion between modeling and decoding in name recognition. To compensate the mismatch, task-specific adaptation, which is performed in the MLLR framework with multi-regression classes, is proposed. Experimental results show that task-specific adaptation is very effective in Chinese name recognition to compensate the mismatch.

1. INTRODUCTION

Name recognition is a practical and interesting application of speech recognition on telephony systems and on mobile communication. Many corporations, such as IBM, Nokia etc., try their best to improve the performance of name recognition.

Name collection for modeling is an onerous burden and acoustic models are usually built using large vocabulary continuous speech corpora. When the models are used in Chinese name recognition, the mismatch between modeling and decoding may impact the recognition performance. This paper addresses the issue of Chinese name recognition and discusses how to compensate the mismatch.

In the literature, there exist many speaker adaptation algorithms, such as MAP and MLLR. The main aim of speaker adaptation is to compensate the mismatch between the specific speaker and the speaker-independent acoustic model. To compensate the distortion between modeling and decoding in Chinese name recognition with the acoustic model trained using large vocabulary continuous speech corpora, the usual speaker adaptation algorithms can also be utilized in this case. Thus task-specific adaptation, which is performed in the MLLR framework with multi-regression classes, is proposed and applied to improve name recognition performance.

The MLLR speaker adaptation algorithm is proposed by Leggetter and Woodland [3] and it performs compensation by transforming the mean vectors (and the covariance matrices) of outputs through an affine transformation. In [4], Gales extended the algorithm with multi-regression classes, thus it can deal with large amounts of enrollment data and let the acoustic model more matched with the tested data.

In this paper, extensive analysis on a large name database is performed and it is concluded that there exists triphones distribution distortion between the name corpus and the large vocabulary continuous speech corpora, thus task-specific adaptation is necessary to deal with the mismatch between modeling and decoding in Chinese name recognition. Then detailed implementation of the task-specific adaptation is described and the adaptation is performed in the MLLR framework with multi-regression classes. Finally experiments are designed to evaluate the proposed approach and the results show that task-specific adaptation is very effective in compensating the mismatch and can give a distinct performance improvement in Chinese name recognition.

2. DATABASE ANALYSIS AND NECESSITY OF TASK-SPECIFIC ADAPTATION

We have been researching Chinese name analysis and recognition and have collected a large name database, which includes 992424 Chinese names.

2.1 Definitions of some terms

There are many duplicate names in the database and we follow the definitions of some terms in [1]. A CDN (Character-Dependent Name) is defined as a name, with a unique character sequence, and different CDNs have different character sequences. A SDN (Sound-Dependent Name) is defined as a name with a unique pronunciation or sound, that is, different SDNs can be distinguished by syllable recognition combined with tone recognition. A SBN (Syllable-Based Name) represents a name with a unique

baseform syllable sequence regardless of tones, and different SBNs have different baseform syllables.

According to the definitions above, our name database contains 580969 CDNs, 503357 SDNs and 420454 SBNs.

2.2 Analysis on triphones

In Chinese speech recognition, triphones can be classified into initial triphones and final triphones, where the base phones are consonants and vowels, respectively [2]. For initial triphones, the left phone is silence or one of 9 head vowels and the right phone is one of 11 tail vowels. For final triphones, the left phone and the right phone are either silence or one of 21 consonants.

According to the definitions of triphones, our name database has 15298 triphones and 13851 triphones if *not considering silence*. While in large vocabulary continuous speech corpora, there exist 23745 triphones and 22295 triphones if *not considering silence* [2].

2.3 Necessity of task-specific adaptation

As it is addressed above, there are many large vocabulary continuous speech corpora, and collecting large vocabulary names to build acoustic models is a difficult job. Thus the acoustic model used in name recognition is usually the same as that in large vocabulary continuous speech recognition.

	Triphones	Triphones with silence	Triphones without silence
Name corpus	15298	1447	13851
Large vocabulary continuous speech corpora	23745	1450	22295

Table 1. Triphones distribution of large vocabulary continuous speech corpora and the name corpus

Table 1 lists the triphones distribution of large vocabulary continuous corpora and the name corpus according to the analysis on triphones in section 2.2. It is obvious that the name database has only 64.4% triphones of Chinese large vocabulary continuous speech corpora. However, the number of triphones with silence is almost the same in the two corpora, which indicates great mismatch in triphones distribution between the two corpora. In other words, names have their special characteristics and a more focused triphone set, while large vocabulary continuous corpora have different characteristics and far more triphones in number. This brings about two problems, on the one hand, the training data are sparse and acoustic models suffer from insufficient training as we could observe from large vocabulary continuous speech recognition task; on the other hand, there are many redundant triphones that are not so important in a name recognition task.

It is clear that there exists great mismatch between modeling and decoding in Chinese name recognition when large vocabulary continuous speech corpora are used to train the acoustic model. Thus it is necessary and effective to compensate the distortion. In this paper, task-specific adaptation is presented to deal with this problem.

3. IMPLEMETATION OF TASK-SPECIFIC ADAPTATION

This section depicts detailed the task-specific adaptation in the MLLR framework with multi-regression classes.

3.1 Derivation of task-specific adaptation from speaker adaptation

In the past years, there were many speaker adaptation algorithms proposed to compensate the distortion between speakers and speaker-independent acoustic models. MLLR, a famous speaker adaptation approach, is proposed by Leggetter and Woodland [3] and it performs compensation by transforming the mean vectors (and the covariance matrices) of outputs through an affine transformation.

In this paper, we propose a task-specific adaptation in the MLLR framework with multi-regression classes. Since the aim of the adaptation is to compensate the mismatch between name recognition and the acoustic models, the adaptation data come from many speakers and cover all the triphones appearing in the name database to ensure that relevant model parameters could be adjusted to characterize those triphones better.

3.2 Implementation of task-specific adaptation

The implementation of task-specific adaptation is illustrated in the following steps.

A. Segment observations to the outputs

The segmentation is performed in a supervised mode given adaptation utterances and the corresponding transcriptions. The adaptation features are assigned to the mixture components of the outputs with the probability of 0 or 1.

B. Build regression tree

The regression tree is built according to the distances of the outputs' means and a clustering procedure is applied. Figure 1 shows a simple binary regression tree, which consists of a hierarchy of regression classes and a set of base classes. A simple top-down scheme is performed to determine proper regression classes according to the number

of observations assigned to the classes to satisfy that each class has enough data to estimate the transformation matrix.

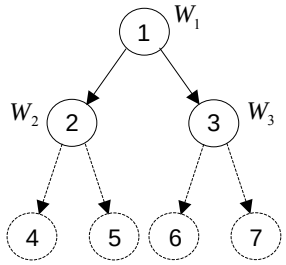

Figure 1. Regression tree for adaptation with multi-regression classes

C. Estimate transformation matrices

The transformation is defined as follows

$$\begin{cases} \hat{\alpha}_{im} = \cdot_k \alpha_{im} + b_k \\ \hat{\Sigma}_{im} = \Sigma_{im} \end{cases}, i \cdot \cdot_k, m = 1,...,M, k = 1,...,K \quad (1)$$

where K denotes the number of regression classes, the outputs are composed of M mixture Gaussian components, and $\cdot_k$ denotes the outputs assigned to regression class k. Transformation parameters $\{\cdot_k, b_k\}$ only modify the outputs belonging to regression class k. The estimation is implemented in the Maximum Likelihood criterion using the EM algorithm. The auxiliary function can be formulated as follows

$$Q(\cdot, \hat{\cdot}) = \sum_k \sum_t \sum_{i \cdot_k} \sum_m \cdot_t(i) \left[K_i - \frac{1}{2} (y_t - \cdot_k \alpha_{im} - b_k)^T \Sigma_{im} (y_t - \cdot_k \alpha_{im} - b_k) \right]$$

where $\hat{\cdot} = \{\cdot_k, b_k, k = 1,...,K\}$ is the transformation parameters needed to be estimated and $\cdot$ is the parameters estimated in the last iteration, $\cdot_t(i)$ denotes the probability of y_t belonging to output i. The estimates of the transformation parameters can be obtained by letting the auxiliary function maximum, which has been described detailed in [3,4] and this paper will not deal with it.

D. Adapt the acoustic model

For output $i(i \cdot \cdot_k)$, the transformation described in equation (1) is adopted to modify the acoustic model.

4. EXPERIMENTAL EVALUATION

This section presents the experiments on task-specific adaptation in Chinese name recognition.

4.1 Baseline system and the corpora

The baseline system is an isolated-word recognizer, and the acoustic model is trained using large vocabulary continuous speech corpora, including the 863 mandarin speech corpus and some other corpora. The model has 827 outputs, each of which has 8 mixture components. The 39-d feature consists of log-power and 12-d MFCC and their 1st- and 2nd-order derivatives. A one-pass decoder is used for recognition. The recognition engine has showed good performance in large vocabulary continuous speech recognition [5].

To testify the performance of task-specific adaptation, two corpora, including the adaptation corpus and the test corpus, were designed. All names were chosen from the 992424-name database. The *adaptation corpus* has 9997 names and they were chosen from 420454 SBNs to cover all the triphones appearing in the database with fewest names. The corpus was divided into 20 subsets, each of which has about 500 names. The adaptation corpus was recorded by 20 speakers and each read one subset. The *test corpus* contains 3000 names *stochastically* chosen from the SBNs in the name database and was divided into 10 subsets, each of which was read by a speaker different from those who read the adaptation corpus.

For the test corpus, there are 3000 names, each of which has a different syllable sequence, so we can design a 3000-name recognizer. The baseline recognition accuracy for the test corpus is 94.2%.

4.2 Task-specific adaptation experiment

For task-specific adaptation, the adaptation corpus is used to modify the acoustic model and the test corpus is used to test the performance.

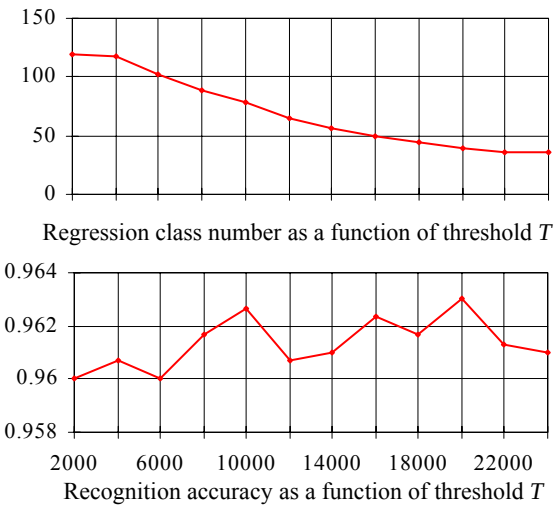

Figure 2. task-specific adaptation recognition results

Before adaptation, a regression tree is built according to the distances of 827 outputs of the acoustic model and then the features of 9997 adaptation names are segmented to the outputs in the supervised mode given the corresponding transcriptions. Then, the depth of the regression tree and the

number of regression classes are determined according to the number of features assigned to each node to ensure that each class has enough data to estimate robustly the transformation matrix (Here, "enough data" is determined by threshold *T*). Finally the transformation matrices are estimated and the corresponding output parameters are modified.

Figure 2 depicts the experimental results of task-specific adaptation for Chinese name recognition. The figure shows that when *T* is from 8000 to 24000, the recognition accuracy is almost above 96.1%. Since the baseline accuracy is only 94.2%, it is obvious that task-specific adaptation can introduce word error rate reduction about 33.8%. In other words, it is concluded that task-specific adaptation is very effective in name recognition to compensate the mismatch between modeling and decoding.

5. CONCLUSION

One important issue on Chinese name recognition, task-specific adaptation, which is proposed to compensate the mismatch between modeling and decoding in Chinese name recognition using acoustic model trained using large vocabulary continuous speech corpora, is discussed extensively in this paper. Experimental evaluation shows that the proposed approach is very effective to compensate the mismatch and can introduce a distinct performance improvement.

6. REFERENCES

[1]. Y. Zhang, A. Medievski, J. Lawrence, J. Song, A Study on Tone Statistics in Chinese Names, *Speech Communication*, 36: 267-275, 2002.

[2]. H. Wu, B. Xu and T. Huang, Automatic Corpus Selection Algorithm Based on Triphone Models, *Chinese Journal of Software*, 11(2): 271-276, 2000.

[3]. C. J. Leggetter, P. C. Woodland, Maximum Likelihood Linear Regressionfor Speaker Adaptation of Continuous Density HMMs, *Computer Speech and Language*, 9: 171-186, 1995.

[4]. M. J. F. Gales, The Generation and Use of Regression Class Trees for MLLR Adaptation, *Technical Report CUED/F-INFENG/TR263*, Cambridge University, 1996.

[5]. S. Gao, T. Lee, Y. W. Wang, B. Xu, P. C. Ching, T. Huang, Acoustic Modeling for Chinese Speech Recognition: A Comparative Study of Mandarin and Cantonese, *In Proc. ICASSP*, 2000.