International Journal of Information Technology, Vol. 11, No. 7, 2005

Comparison of Non-native Speaker Adaptations for Large Vocabulary Continuous Mandarin Speech Recognition

Jian Yang, Hong Wei, Yuanyuan Pu, and Zhengpeng Zhao

School of Information Science and Technology, Yunnan University
Kunming, 650091, Yunnan, P. R. China
jianyang@ynu.edu.cn, hwei1973@163.com
oldsan@21cn.com, zhpzhao@ynu.edu.cn

Abstract

Improving the performance of the state-of-the-art Mandarin speech recognition system for non-native speech remains a challenging task because of wide varieties of non-native accents. The recognition accuracy of the baseline models trained by Standard Mandarin Corpus was drastically low for the non-native speakers from Naxi and Lisu in Yunnan than for the native ones. To verify that the acoustic deviation is one of important factors affecting the performance of recognizer, maximum likelihood linear regression (MLLR) was adopted both alone and in combination with maximum a posteriori (MAP) with the Linguistic Minorities Accents Mandarin Speech Corpus which collected by our laboratory. It is shown that when MLLR and MLLR + MAP are used, the correct rates increase evidently.

Keyword: Mandarin speech recognition, non-native speaker, Naxi accent, Lisu accent, speaker adaptation, MLLR, MAP.

1 Introduction

Over the past decade, there have been tremendous efforts on large vocabulary continuous speech recognition for Chinese. Among the multifarious Chinese dialects, Mandarin (or Putonghua) has received the most research and commercial interests, given its huge speaker population and the unique role as the official standard of spoken Chinese. Nevertheless, there has been an obvious and ever increasing demand for speech recognition technology that can deal with Chinese dialects and non-native Mandarin, spoken by foreigner or the speakers from the minority areas in China. The reasons are at least two-fold. Firstly, more and more foreigners learn Chinese and speak Mandarin with foreign accent. Secondly, most of the national minorities in

This work was supported by the National Natural Science Foundations of China under grant No.60265001.

China, such as *Naxi*, *Dai*, *Zang* etc., have their languages, so they speak Mandarin with their native language accents. Non-native specific investigation is not only justifiable but also necessary for the advancement of Chinese speech recognition technology.

Although the current speech recognition systems work very well for native talkers, their performance degrades dramatically when recognition is performed on speech with heavy non-native accents [1]. One reason is because the non-native speakers' pronunciation differs from those native speakers' pronunciation observed during system training. A number of methods for handling non-native speech in speech recognition have been proposed. The most straightforward approach is to use the non-native speech from the target language spoken by the group of non-native speakers for recognizer training [2]. However the problem of this method is that the non-native speech data is only rarely available. Another approach is to apply general speaker adaptation techniques such as MLLR (Maximum Likelihood Linear Regression) and MAP (Maximum A-Posterior) on speaker-independent models to fit the characteristics of a non-native accent [3].

The model based adaptation schemes are divided into three families: parameter transformation based adaptation using maximum likelihood linear regression (MLLR) and similar schemes; the maximum a posteriori (MAP) adaptation family; and a family related to speaker clustering methods or speaker-space methods. The strengths and weaknesses of these methods are considered and a number of hybrid schemes have appeared which combine the above methods in various ways [4].

For this study, we implement a number of acoustic modeling techniques to compare their performance on non-native speech recognition. Here we restrict our study to non-native Mandarin speech spoken by speakers from *Naxi* and *Lisu* in *Yunnan*, China. In more detail, we explore how the acoustic models can be adapted to better handle the non-native speech. To verify that the acoustic deviation is one of important factors affecting the performance of recognizer, MLLR was adopted both alone and in combination with MAP.

This paper is organized as follows. The speech corpus is presented in section 2. In section 3, we describe the baseline HMM models of our experiments. Section 4 describes the MLLP and MAP algorithm. Detailed experiments and results on acoustic models adaptation are given in section 5. Section 6 concludes with summary of our work.

2 Speech Corpus

In this study, two speech corpora shown in Table 1, one native speech corpus and one non-native speech corpus, are used.

The native Mandarin speech data are extracted from the Mandarin Dictation Corpora supported by *China National* Hi-*Tech Project 863*. We used the utterances from 87 speakers (38 males and 49 females) to train the baseline HMM models. The non-native Mandarin speech data are extracted from the *Linguistic Minorities Accents Mandarin Speech Corpus (LMAMSC)*, which collected by our laboratory. The Chinese sentence prompts of the LMAMSC were the same sentences as the first corpus.

Recordings were made with a high-quality head-mounted microphone in a quiet laboratory environment. The data was digitized at 16 bits per sample and a sampling rate of 16 kHz. The all speakers are from minority areas in Yunnan and their native languages are not Chinese. The non-native accents are obvious when they speak Mandarin.

Table 1. Speech corpus overview

Corpus	Accent	Partition	Speakers	Sentences
Project 863	Native	Training	87	39800
		Testing	11	5700
LMAMSC	Naxi		6	3600
	Lisu		6	3600

3 Baseline System

3.1 Training

All recognition experiments described in this paper use the HTK Toolkit [5]. The acoustic models of the baseline system for native Mandarin speech are trained on the native corpora data. The whole training procedure closely follows the one outlined in the Microsoft Mandarin Speech Toolbox [6]. The feature used is a 39order feature vector, consisting of 12 MFCCs (Mel Frequency Cepstral Coefficient), energy, and their first and second order differences. The feature vector is calculated using a window size of 25ms and a step size of 10ms. The whole training procedure should be divided into two stages: monophone and triphone. In each stage, there are always two steps, which are repeated iteratively: estimation and realignment. The process begins with the training of the monophone models, followed by training of the triphone models. For predicting unseen triphone in recognition, the parameter of tied-state triphone should be estimated.

Table 2. Initial and tonal final units [6]

Initial	b, c, ch, d, f, g, ga, ge, ger, go, h, j, k, l, m, n, p, q, r, s, sh, t, w, x, y, z, zh
Tonal final	a(1-5), ai(1-4), an(1-4), ang(1-5), ao(1-4), e(1-5), ei(1-4), en(1-5), eng(1-4), er(2-4), i(1-5), ia(1-4), ib(1-4), ian(1-5), iang(1-4), iao(1-4), ie(1-4), if(1-4), in(1-4), ing(1-4), iong(1-3), iu(1-5), o(1-5), ong(1-4), ou(1-5), u(1-5), ua(1-4), uai(1-4), uan(1-4), uang(1-4), ui(1-4), un(1-4), uo(1-5), v(1-4), van(1-4), ve(1-4), vn(1-4)

In this study, we train the acoustic models based on syllables. The basic acoustic units used for recognition are shown in Table 2. The baseline acoustic model was designed to be tonal since tone is an important feature of the Chinese language. After the monophone models are trained, all possible triphone expansions based on the full syllable dictionary are performed. This results in a total of 270,998 triphones. Out of these triphones, 24,127 triphones actually occur in the training corpus. After performing several iterations of embedded reestimation, we use the decision tree based clustering capability of the HTK toolkit to tie similar states of triphones to each other. After clustering, the number of unique Gaussian mixtures is reduced to 16,112. We then use the HTK toolkit's Gaussian splitting capability to incrementally increase the number of Gaussians mixture to 8.

3.2 Testing

After the acoustic models of the baseline system for native Mandarin speech are trained, we perform a set of recognition experiments. The standard HTK decoder HVite was used for the experiment.

We first performed syllable decoding without language model based on a syllable loop word net. This recognition task puts the highest demand on the quality of the acoustic models. All 1677 syllables are listed in the network and any syllable can be followed by any other syllable, or they may be separated by short pause or silence. Secondly, we have included a syllable bigram language model that had been estimated from the training set syllable transcription with the tool HLstats. Since recognizing Chinese tones is a very difficult task, we have also calculated results that do not count tone misrecognitions as errors (shown in the following Tables as the Base syllable correct).

The baseline syllable recognition results on the test set of 5,700 sentences from 11 native speakers are shown in Table 3 and the results on the test set of two non-native accents speech, 600 sentences from 6 Naxi speakers and 600 sentences from 6 Lisu speakers, are shown in Table 4. As expected, the recognition accuracy of baseline models is drastically low for the non-native speakers.

Table 3. Recognition results on the native test set

	Base syllable correct (%)	Tonal syllable correct (%)
Without LM	82.4	63.8
With LM	93.0	91.3

Table 4. Recognition results on the non-native set without language model

Accent	Base syllable correct (%)	Tonal syllable correct (%)
Naxi	39.4	23.1
Lisu	33.8	19.4

4 Speaker Adaptation

In speaker adaptation, acoustic models that have been trained for general speech are adjusted so that they better model the speech characteristic of a specific condition. Those adaptation techniques do not have to be limited to speaker adaptation; general models can be specialized to compensate for differences in acoustic environment or the characteristic of a group of speakers.

Speaker adaptation techniques can be used in various different modes. If the true transcription of the adaptation data is known then it is termed supervised adaptation, whereas if the adaptation data is unlabelled then it is termed unsupervised adaptation. In the case where all the adaptation data is available in one block, e.g. from a speaker enrollment session, then this termed static adaptation. Alternatively adaptation can proceed incrementally as adaptation data becomes available, and this is termed incremental adaptation.

4.1 Model Adaptation Using MLLR

Maximum likelihood linear regression or MLLR computes a set of transformations that will reduce the mismatch between an initial model set and the adaptation data [5] [7]. More specifically MLLR is a model adaptation technique that estimates a set of linear transformations for the mean and variance parameters of a Gaussian mixture HMM system. The effect of these transformations is to shift the component means and alter the variances in the initial system so that each state in the HMM system is more likely to generate the adaptation data.

The transformation matrix used to give a new estimate of the adapted mean is given by

$$\hat{\mu} = W\xi \ . \qquad (1)$$

Where, W is the $n \times (n+1)$ transformation matrix (where n is the dimensionality of the data) and ξ is the extended mean vector,

$$\xi = [w \ \mu_1 \ \mu_2 \ \dots \ \mu_{\mathrm{n}}]^T . \qquad (2)$$

Where w represents a bias offset whose value is fixed (within HTK) at 1.

Hence W can be decomposed into

$$W = [b \ A] \ . \qquad (3)$$

Where A represents a $n \times n$ transformation matrix and b represents a bias vector.

The transformation matrix W is obtained by solving a maximization problem using the *Expectation-Maximization* (*EM*) technique. This technique is also used to compute the variance transformation matrix.

This adaptation method can be applied in a very flexible manner, depending on the amount of adaptation data that is available. If a small amount of data is available then a *global* adaptation transform can be generated. A global transform (as its name suggests) is applied to every Gaussian component in the model set. However, as more adaptation data becomes available, improved adaptation is possible by increasing the number of transformations. Each transformation is now more specific and applied to certain groupings of Gaussian components. For instance the Gaussian components could be grouped into the broad phone classes: silence, vowels, stops, glides, nasals, fricatives, etc. The adaptation data could now be used to construct more specific broad class transforms to apply to these groupings.

MLLR makes use of a *regression class tree* [7] to group the Gaussians in the model set, so that the set of transformations to be estimated can be chosen according to the amount and type of adaptation data that is available. The tying of each transformation across a number of mixture components makes it possible to adapt distributions for which there were no observations at all. With this process all models can be adapted and the adaptation process is dynamically refined when more adaptation data becomes available.

The regression class tree is constructed so as to cluster together components that are close in acoustic space, so that similar components can be transformed in a similar way. Note that the tree is built using the original speaker independent model set, and is thus independent of any new speaker. The tree is constructed with a centroid splitting algorithm, which uses a Euclidean distance measure. The terminal nodes or leaves of the tree specify the final component groupings, and are termed the *base (regression) classes*. Each Gaussian component of a model set belongs to one particular base class.

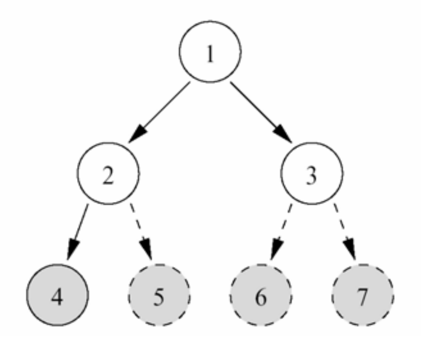

Fig. 1. A binary regression tree [5]

Figure 1 shows a simple example of a binary regression tree with four base classes, denoted as $\{C_4, C_5, C_6, C_7\}$. During “dynamic” adaptation, the occupation counts are accumulated for each of the regression base classes. The diagram shows a solid arrow and circle (or node), indicating that there is sufficient data for a transformation matrix to be generated using the data associated with that class. A dotted line and circle indicates that there is insufficient data. For example neither node 6 or 7 has sufficient data; however when pooled at node 3, there is sufficient adaptation data.

In the HTK, the amount of data that is “determined” as sufficient is set by the user as a command-line option to HEAdapt. HEAdapt uses a top-down approach to traverse the regression class tree. Here the search starts at the root node and progresses down the tree generating transforms only for those nodes which

1. have sufficient data and
2. are either terminal nodes (i.e. base classes) or have any children without sufficient data.

In the example shown in figure 1, transforms are constructed only for regression nodes 2, 3 and 4, which can be denoted as W_2, W_3 and W_4. Hence when the transformed model set is required, the transformation matrices (mean and variance) are applied in the following fashion to the Gaussian components in each base class:

$$\left\{\begin{matrix} W_2 \to \{C_5\} \\ W_3 \to \{C_6, C_7\} \\ W_4 \to \{C_4\} \end{matrix}\right\} \tag{4}$$

At this point it is interesting to note that the global adaptation case is the same as a tree with just a root node, and is in fact treated as such.

4.2 Model Adaptation Using MAP

Model adaptation can also be accomplished using a maximum a posteriori (MAP) approach [5] [8]. This adaptation process is sometimes referred to as Bayesian adaptation. MAP adaptation involves the use of prior knowledge about the model parameter distribution. Hence, if we know what the parameters of the model are likely to be (before observing any adaptation data) using the prior knowledge, we might well be able to make good use of the limited adaptation data, to obtain a decent MAP estimate. This type of prior is often termed an informative prior. Note that if the prior distribution indicates no preference as to what the model parameters are likely to be (a non-informative prior), then the MAP estimate obtained will be identical to that obtained using a maximum likelihood approach.

For MAP adaptation purposes, the informative priors that are generally used are the speaker independent model parameters. For mathematical tractability conjugate priors are used, which results in a simple adaptation formula. The update formula for a single stream system for state *j* and mixture component *m* is

$$\hat{\mu}_{jm} = \frac{N_{jm}}{N_{jm} + \tau} \bar{\mu}_{jm} + \frac{\tau}{N_{jm} + \tau} \mu_{jm} \ . \tag{5}$$

Where τ is a weighting of the a priori knowledge to the adaptation speech data and *N* is the occupation likelihood of the adaptation data, defined as,

$$N_{jm}=\sum_{r=1}^{R}\sum_{t=1}^{T_r}L_{jm}^{r}(t)\ . \tag{6}$$

Where, μ_{jm} is the speaker independent mean and $\overline{\mu}_{jm}$ is the mean of the observed adaptation data and is defined as,

$$\overline{\mu}_{jm}=\frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}L_{jm}^{r}(t)o_t^r}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}L_{jm}^{r}(t)} \tag{7}$$

Where, the following notation is used in above equations

R the number of the training observation sequences

T_r the number of observations of sequence *r*

O^r observation sequence *r*, $1\le r\le R$

o_t^r the observation of sequence *r* at time *t*, $1\le t\le T_r$

$L_{jm}^{r}(t)$ the occupancy probability for state *j* and mixture component *m* at time *t* of sequence *r*

As can be seen, if the occupation likelihood of a Gaussian component (N_{jm}) is small, then the mean MAP estimate will remain close to the speaker independent component mean. With MAP adaptation, every single mean component in the system is updated with a MAP estimate, based on the prior mean, the weighting and the adaptation data. Hence, MAP adaptation requires a new “speaker-dependent” model set to be saved.

One obvious drawback to MAP adaptation is that it requires more adaptation data to be effective when compared to MLLR, because MAP adaptation is specifically defined at the component level. When larger amounts of adaptation training data become available, MAP begins to perform better than MLLR, due to this detailed update of each component (rather than the pooled Gaussian transformation approach of MLLR). In fact the two adaptation processes can be combined to improve performance still further, by using the MLLR transformed means as the priors for MAP adaptation (by replacing μ_{jm} in equation (5) with the transformed mean of equation (1)). In this case components that have low occupation likelihood in the adaptation data, (and hence would not change much using MAP alone) have been adapted using a regression class transform in MLLR.

5 Experiments and Results

HTK provides two tools to adapt continuous density HMMs, offline supervised adaptation using MLLR and/or MAP. If MLLR and MAP adaptation is to be performed simultaneously using HTK in the same pass, then the restriction is that the entire adaptation must be performed statically. In this section, we describe the approaches that we tried and compare their performance.

5.1 MLLR

To evaluate the acoustic model adaptation performance, we carry out the supervised static MLLR experiments. All phones were classified into 65 regression classes. The tool HHEd was used to build a binary regression class tree, and to label each component with a base class number. Both diagonal matrix and bias offset were used in the MLLR transformation matrix. Adaptation set size ranging from 30 to 500 utterances for each speaker was tried. Results are shown in the Table 5. It is shown that when the number of adaptation utterances reaches 30, the all relative correct rates increase based on speaker independent (SI) system are more than 33%.

Table 5. Performance of MLLR with different adaptation sentences

Number of adaptation sentences		0	30	100	500
Naxi	Base syllable correct (%)	39.4	52.7	53.8	57.9
	Relative correct increase based on SI (%)	--	33.8	36.5	46.9
	Tonal syllable correct (%)	23.1	37.7	37.8	42.4
	Relative correct increase based on SI (%)	--	63.2	63.6	83.5
Lisu	Base syllable correct (%)	33.8	58.4	60.7	65.2
	Relative correct increase based on SI (%)	--	72.8	79.6	92.9
	Tonal syllable correct (%)	19.4	40.0	41.3	47.3
	Relative correct increase based on SI (%)	--	106	113	144

Table 6. Performance of combined MLLR and MAP with different adaptation sentences

Number of adaptation sentences		0	30	100	500
Naxi	Base syllable correct (%)	39.4	63.9	59.3	82.9
	Relative correct increase based on SI (%)	--	62.2	50.5	110
	Tonal syllable correct (%)	23.1	51.9	43.7	77.7
	Relative correct increase based on SI (%)	--	125	89.2	236
Lisu	Base syllable correct (%)	33.8	70.0	65.6	88.8
	Relative correct increase based on SI (%)	--	107	94.1	163
	Tonal syllable correct (%)	19.4	55.0	46.3	82.9
	Relative correct increase based on SI (%)	--	184	139	327

5.2 Combined MLLR and MAP

The results of combined MLLR and MAP are shown in the Table 6. It is shown that when the number of adaptation utterances reaches 500, the base syllable correct rates are more than 82% and the tonal syllable correct rates are more than 77% for both Naxi and Lisu accent speakers.

6 Summary

In this paper, a new non-native accents speech corpus of dictation Mandarin for LVCSR has been described. Based on the corpus, we explore how the acoustic models can be adapted to better recognize the non-native speech. The results show that there are many problems such as adaptation method, the non-native pronunciation patterns that remain to be investigated. While this speech appears significantly more difficultly to recognize than native Mandarin, we expect performance on this task to benefit from progress in speaker adaptation in general with more non-native accents, such as Bai, Yi, Zang, Dai etc. in Yunnan, China. In future, it will be necessary to improve speaker adaptation system by incorporating more extensive knowledge of speaker variation at both the acoustic and the pronunciation level.

Acknowledgements

The authors wish to acknowledge Eric Chang for providing the Microsoft Research Mandarin Speech Toolkit and the set of CD-ROMs of speech corpus.

References

[1] Tomokiyo, L.M.: Recognizing Non-native speech: Characterizing and Adapting to Non-native Usage in Speech Recognition. Ph.D. Thesis. Carnegie Mellon University (2001).

[2] Uebler, U., Boros, M.: Recognition of Non-native German Speech with Multilingual Recognizers. Proc. Eurospeech, Vol. 2. Budapest (1999), pp. 911-914.

[3] Wang, Z., Schultz, T., Waibel, A.: Comparison of Acoustic Model Adaptation Techniques on Non-Native Speech. Proc. ICASSP (2003), pp. 540-543.

[4] Woodland, P.C.: Speaker Adaptation for Continuous Density HMMs: A Review. ITRW on Adaptation Methods for Automatic Speech Recognition, Sophia Antipolis, France (2001), pp. 11-19.

[5] Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev,V., and Woodland, P.C.: The HTK Book. http://htk.eng.cam.ac.uk.

[6] Chang, E., Shi, Y., Zhou, J.L. and Huang, C.: Speech Lab in a Box: A Mandarin Speech Toolbox to Jumpstart Speech Related Research. Proc. Eurospeech, Aalborg, Denmark (2001), pp. 192-199.

[7] Leggetter, C.J., Woodland P.C.: Flexible Speaker Adaptation Using Maximum Likelihood-Linear Regression. Proc. ARPA Spoken Language Technology Workshop. Morgan Kaufmann (1995), pp. 104-109.

[8] Gauvain J.L., Lee C.H.: Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. IEEE Trans. SAP, Vol. 2 (1994), pp. 291-298.

Jian Yang received the B.Sc. degree from Yunnan University, Kunming, China, and the M.Eng. degree from University of Science and Technology of China (USTC), Hefei, in 1985 and 1989, respectively, all in electrical engineering.

He is now an associate professor with the School of Information Science and Technology, Yunnan University. His current research interests include all issues related to speech recognition and understanding, especially robust speech recognition for non-native speaker, adaptive modeling of speech, spoken language systems, and speaker recognition/verification.

Hong Wei received the B.Sc. degree in physical electronics and M.Sc. degree in electrical engineering from Yunnan University, Kunming, China, in 1995 and 2002 respectively.

Then he joined the School of Information Science and Technology, Yunnan University as a lecturer. His research interests include speech signal processing, speaker recognition/verification, and speaker adaptation.

Yuanyuan Pu received the B.Sc. and M.Sc. degree in electrical engineering from Yunnan University, Kunming, China, in 1995 and 1998, respectively.

She is now a lecturer at the School of Information Science and Technology, Yunnan University. Her research interests include speech recognition and understanding, non-native speech recognition and speaker adaptation.

Zhengpeng Zhao received the B.Sc. and M.Eng. degree in electrical engineering from Yunnan University, Kunming, China, in 2001 and 2004, respectively.

Then he joined the School of Information Science and Technology, Yunnan University as a lecturer. His research interests include speech recognition and understanding, language and accent recognition/verification.