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

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

Incorporating HMM-state Sequence Confusion for Rapid MLLR Adaptation to New Speakers

Bing Zhao, Bo Xu

National Lab of Pattern Recognition Institute of Automation, Chinese Academy of Sciences
{bzhao,xubo}@nlpr.ia.ac.cn

ABSTRACT

In this paper, we introduce the HMM-state sequence confusion characteristics as prior knowledge into the framework of MLLR to relax the transformation and reduce the risks of over-training when adaptation data size is small. There are two issues to be addressed as follows: first, how to estimate such confusion information reliably; second how to use the information in refining the estimation of MLLR adaptation. The pronunciation modeling technology was utilized to build the state sequence confusion table. Then the correlation of states is calculated according to the confusion table. Following proposed algorithm made a relaxation in the process of MLLR adaptation when the adaptation data is very small. Our experiment on a Mandarin state-tying triphone toneless LVCSR system showed that error rate reduction is 9.5% over standard MLLR with about 10 utterances (less than 30 seconds) of adaptation data.

1. INTRODUCTION

In adaptation techniques, MLLR is a popular choice for its effectiveness with small training data. But due to the limitations of the criteria of maximum likelihood, MLLR has two drawbacks. First, it is a global transformation and thus hard to refine the local behavior of the parameters; second it is difficult to introduce the prior knowledge of acoustic model parameters. The first limitation makes MLLR require more training data than MAP to approach to an SD model; the second one always leads to unreliable estimates and over training problems when the adaptation data size is very small It is not difficult to cope with the first limitation a very detailed regression class tree can be constructed, though the computation cost maybe expensive when the number of class is very large (i.e. 60 or more). But it is not easy to address the second one. Many previous works either severely restrict the space of model transforms [1] or use some tech. to reduce the parameters to be estimated [2]. This kind of approach always limits the power of the adaptation. Other works try to utilize the prior information from several sources such as the correlation of the Gaussian means [3] or the priors on the transforms [4, 5]. This induced unreliable assumptions on adaptation, and also requires extra estimation of the prior parameters.

In fact, incorporating prior knowledge into adaptation procedure can greatly reduce the risks of over-training. MAP, for instance, rarely has the problem of over-training, since it incorporated the prior assumed distribution of the Acoustic Model's parameters. But it is indeed very hard to incorporate prior knowledge into the framework of MLLR, as we stated above. Some work uses correlation between Gaussian mean parameters of acoustic model, which is somewhat limited, since the high static correlation between two parameters does not necessarily guarantee that the two parameters behave similarly when new training data is coming in. The similar problem occurs in the usual acoustic model re-training and adaptation that utilizes only phoneme-phoneme or state-state relation for updating of HMM parameters. A nature idea to use high reliably estimation of the correlation of HMM parameters, is to record the detail behavior of every parameter during the training process of acoustic model, and use this information to calculate the relationship. But this is so far a huge computation and hard to implement.

In this paper, we propose a more practical way to address the over-training issue by introducing prior information into the framework of MLLR. First, in section II, we estimated the local relationships between HMM states by using the pronunciation modeling technique. This relationship is expressed by HMM state sequence confusion characteristics in the state tying triphone LVCSR system. Then in section III, we use this information to give the MLLR procedure a relaxation to avoid the dangers of over-training when the adaptation data size is very small. In section IV, the experiment on a state tying triphone mandarin LVCSR system is presented and explained.

2.HMM STATE SEQUENCE CONFUSION

By the term “confusion”, we mean that a group of HMM parameters share something in common, which can be expressed by their correlations. The parameters move in the similar direction in model space when new training data come in. We will explain how to estimate this kind of local behavior relationships of HMM parameters via HMM state sequence confusion characteristics in this section.

2.1 Prior Knowledge of HMM-state Sequence

Here HMM state sequence is state sub-sequence of an utterance or word. It may consist of various numbers of state units. It can reflect the acoustic context in a level lower than word or sub-word unit. This kind of confusion statistics within various lengths of HMM state sequences is useful for speaker adaptation. It reflects useful informative factors such as speaker accent, acoustic context, and even the pronunciation variation caused by the phoneme's position in the phrase. As it is well known, these factors caused a large part of the mismatch between training and testing. In our previous work [8], we use the Mandarin syllable unit confusion in a context dependent/independent way to build a pronunciation model. Work [9] also showed the potential usefulness of the information represented by HMM state sequence confusion in building an acoustic model.

2.2 Estimation of Confusion Table

Our approach is an extension to our previous approaches described in [8]. Here we use unit of HMM state within various lengths of HMM state sequence context.

Since we are not aimed at pronunciation modeling, there are two important issues. First, the confusion is not used to correct the recognition output, but used as a hint of relationships of parameters' local behaviors. Second, the estimations are restricted in a tree structure. This tree structure was build to localize the transform of MLLR. In this tree, all parameters are divided into 2 groups according to Chinese phonology: initials and finals, and then we use LBG and K-means to develop the tree further. Thus every node has a centroid, and these centroids are used to calculate the distance between the nodes in the tree.

The estimation procedure can be briefly stated as follows:

a. First, we need to obtain a canonical pinyin transcription of the training data. A standard pinyin dictionary is used for our purpose. The training data is part of the training data of the acoustic model, which is non-accent. Here we use 863 training corpus which is widely used in testing/building a Chinese recognizer. The recognizer's output is the toneless syllable.
b. Feed the training data into the recognizer, and the beam search results N best state sequences. From the training data's label, we obtain the accurate surface-form of state sequence. From the output of the recognizer, we obtain the probably mis-recognized triphone state sequence.
c. After that, a dynamic programming (DP) tech. was used to align the pair-wise confusion sequences.
d. Count two kinds of occurrences: first, calculate the appearances of correct appearances in the aligned results; second, calculate the confusions between all the HMM state sequences in a pair-wise way. We save the results in a table. All of the two occurrences of HMM state sequence confusion characteristics are counted and normalized to 1.0. At this stage, we remove those unreliable estimates of two sequences whose average distance within the tree structure is beyond certain threshold, and thus ensure a table of reasonable estimates of HMM state confusion.

The confusion table saves the main confusion characteristics spanned across HMM states. In this table, the correlation information of the states local behaviors' relationship can be calculated according to this table. We give a simple example for illustration as follows:

Look at the simple Chinese sentence: (luo2 bo5 tu3 dou5 da4 bai2 cai4) toneless syllable recognized result:

Correct: luo BO tu DOU da bai cai
Recognized: luo BU tu DE da bai cai

There are two confusions: BO-BU and DOU-DE. Considering (BO-BU), the triphone level confusion is {(b-o+t, o-t+u), (b-u+t,u-t+u)}. And the states confusion sequence is {(st_1 , st_2 , st_3 , st_4), (st_1' , st_2' , st_3' , st_4')}. With (st_2 , st_2') is a confusion pair grouped from the sequence. This is showed in figure 1.

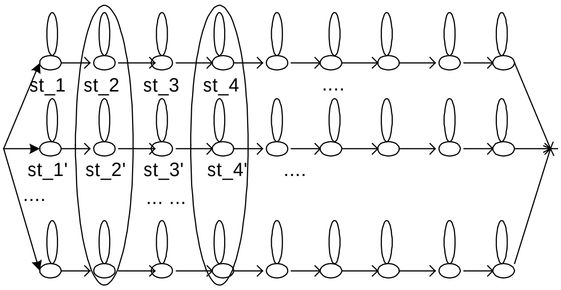

Fig. 1 The grouping of the state confusion

We count both the occurrences of the correct pair and confusion pair and normalized their appearance frequency to 1.0, and the probability of the confusion pair is calculated and saved in a confusion table like table 1.

States	Confusion	Triphone level confusion pairs	Prob.
st_2	st_2 (s_0)	{(b-o+t),(b-o+t)}	0.984
	st_2' (s_1)	{(b-u+t),(b-o+t)}	0.003
	st_2''(s_2)	{(b-ou+t),(b-o+t)}	0.001

Table 1. Example of state pair-wise confusion table: st_2 in triphone (b-o+t) and the corresponding confusion states.

For example, the output PDF of state st_2 in (b-u+t) is calculated as follows:

$$P^{ex}(o_t \mid st\_2) = \sum_i P((s_i, st\_2) \mid S^-, st\_2, S^+) * P(o_t \mid s_i) \qquad (1)$$

Where (s_i, st_2) denotes the confusion state pairs of s_i *and* st_2. S^- *and* S^+ denote the left and right states sequence context of state st_2. And $P((s_i, st\_2) \mid S^-, st\_2, S^+)$ is the normalized prob. of confusion pair (s_i, st_2) 's appearance. It should satisfy:

$$\sum_i P((s_i, st\_2) \mid S^-, st\_2, S^+) = 1.0$$

3. EM FRAMEWORK OF MLLR USING HMM-STATE SEQUENCE CONFUSION

We begin here with the standard MLLR within EM framework approach in a geometric information view, and then extend it by giving the procedure a relaxation in our approach to accommodate the HMM-state sequence confusion characteristics.

3.1 Standard MLLR Solutions

In [6], EM algorithm can be viewed as alternating minimization between a parameter set Θ and a family of desired parameter set. The alternating minimization is done under the information divergence defined as follows:

$$D(g_X, \theta) = \sum_x g_X(x) \log \frac{g_X(x)}{Q_X(x;\theta)} \tag{2}$$

Given the initial parameters, the E step and M step act interlacingly and the algorithm converges to the local optimal estimations [6][11].

Standard MLLR was can be solved under the EM framework. The regression class transformation W_s^{n+1} was re-estimated from the previous transformations W_s^{n}, and the transformation matrix satisfied the following equation:

$$\sum_{r=1}^{R}\sum_{\tau=1}^{T}\gamma_{s_r}(\tau)C_{s_r}^{-1}o_t\hat{\mu}_{s_r} = \sum_{r=1}^{R}\sum_{\tau=1}^{T}\gamma_{s_r}(\tau)C_{s_r}^{-1}W_s\hat{\mu}_{s_r}\hat{\mu}_{s_r}^{'} \tag{3}$$

Using Viterbi or Forward-backward algorithm to get the accumulators as follows:

$$v^{(r)} = \sum_{\tau=1}^{T}\gamma_{s_r}(\tau)C_{S_r}^{\ -1} \tag{4}$$

$$d^{(r)} = \sum_{\tau=1}^{T}\gamma_{s_r}(\tau)C_{S_r}^{\ -1}o_t \tag{5}$$

The solution to MLLR [10] is as follows

$$w'_i = G^{(i)-1}z'_i \tag{6}$$

3.2 Relaxation to MLLR Adaptation

In our approach, the objective function is extended to the following form:

$$\hat{D}(P_X, \theta) = \sum_x P_X^{ex}(x) \log \frac{P_x^{ex}(x)}{Q_X(x;\theta)} \tag{7}$$

Where $$p^{ex}(o_t \mid s) = \sum_{i \in Group\ (s)} \omega_i^{\ ex} \cdot p(o_t \mid \tilde{s}_i)$$

$$\omega_i^{ex} = prob(\tilde{s}_i \mid s) \tag{8}$$

$Group(s)$ denotes the confusion set including the state of S itself and all of its confusion counterpart $\tilde{S}_i$. (8) is calculated from the confusion table.

Thus $\hat{D}(P_X,\theta)$ now denotes the modified version of the standard object function. In fact it is the relaxed version, since it optimizes the relaxed PDF $P_X^{ex}(x)$ instead of the original PDF $P_X(x)$. During the accumulating stage of MLLR, the accumulators will be modified accordingly as follows:

$$v^{(r)} = \sum_{\tau=1}^{T}\Big(\sum_{i \in group(s)}\omega_i^{ex}.\gamma_{s_r}^{i}(\tau)\Big)C_{S_r}^{\ -1} \tag{9}$$

$$d^{(r)} = \sum_{\tau=1}^{T}\Big(\sum_{i \in group(s)}\omega_i^{ex}.\gamma_{s_r}^{i}(\tau)\Big)C_{S_r}^{\ -1}o_t \tag{10}$$

And $\gamma_{s_r}^{i}(\tau)$ is the state occupancy counter at time τ. It is clear that $p^{ex}(o_t \mid s)$ in (8) is still a Gaussian distribution, since it is the addition of several Gaussian PDFs. The difference from the original distribution is now it has more mixture components than original. Note it can also be written as M mixtures format as follows:

$$p^{ex}(o_t \mid S) = \sum_{i=1}^{M}\left\{\begin{array}{l} prob(s \mid s)w_i^s \cdot N^s(o_t;\mu_i^s,\eta_i^s) + \\ \displaystyle\sum_{j \in group(s)} prob(\tilde{s}_j \mid s)w_i^j \cdot N^j(o_t;\mu_i^j,\eta_i^j) \end{array}\right\} \tag{11}$$

From the equation above, we see that the additional part of the function is aimed to give the original PDF a relaxation as illustrated in fig. 2. It shifts a little bit from original one.

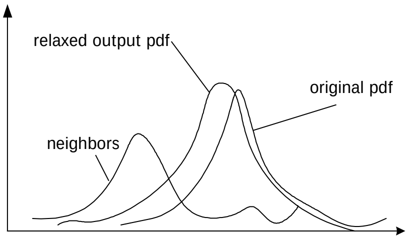

Fig 2. The relaxed PDF

In our case, It is easy to follow the same solving procedure as MLLR, the only difference is that we use a different way of calculating the accumulator for the transformation matrix, and the resulted transformation matrix W_i is aimed at transform the relaxed version of state output pdf as the objective function defined. In fact, it is utilized on the original Gaussian mean. In this way, the relationship between states is incorporated in the process of MLLR as a confidence weighing on the original pdf and its correlated counterparts. Our experiment showed this incorporating of prior knowledge of state confusion help to reduce the risks of over-training in the regime of small training data size. The experiment will be explained in the next section.

4. EXPERIMENT AND DISCUSSION

Our experiment is carried on our state-tying triphone system. There are totally 3062 shared states. Each state's output pdf has 16 mixture components. This acoustic model is trained from standard mandarin speech corpuses. The training corpuses are standard Chinese without accent. The recognizer's output is syllable without tone. We use gender dependent models. The testing data is recorded in an office environment, and there are totally 5 test speakers. Some of them have a little accent. Each speaker has 120 testing utterances. All experiments are done on this recognizer and the train/test corpuses described above. The recognizer's baseline is 69.19%. Part of the training corpus is used to extract the HMM state sequence confusion table. The size is about 1500 utterances from standard speakers.

When implementing MLLR, we first divided all the parameters into 2 classes according to Chinese phonology: initials and finials. And then develop the tree further by LBG and K-Means using KL distance. We give each node an occupancy threshold of 2000 frames. And the maximum number of confusion pairs is restricted to 8. When the adaptation data size is small (fewer than 40 utterances), the regression tree based MLLR showed better performance than the usual one-class (global) MLLR. And we use this MLLR as our baseline to evaluate the proposed algorithm.

The average adaptation performance of the 5 speakers testing results are used to compare standard MLLR with our extended case using HMM state sequence prior information with regard to different size of the training data. The size of the adaptation data varies from 5 utterances to 30 utterances. The result is as the following table 2:

Num of Utt.	MLLR	Relaxed-MLLR	Error reduction
5	69.76(1.9%)	71.32(6.9%)	5.16%
10	70.87(5.5%)	73.64(9.2%)	9.51%
15	71.39(7.1%)	73.48(13.9%)	7.31%
20	71.31(6.9%)	73.98(15.5%)	8.99%
30	72.45(10.6%)	73.26(13.2%)	2.94%

Table 2: Varying adaptation data size

The column of MLLR and Relaxed-MLLR indicates the recognition rate (the first number) and the error reduction over baseline (in the brackets). And the last column showed the error reduction of proposed relaxed method over standard MLLR. In the table, standard MLLR showed a slow increase over baseline when the data size is small. The proposed algorithm showed a little quicker performance improvement than the standard MLLR. When the adaptation data is very small, for example 10 utterances, we see an absolute 2.77 improvement or an error reduction of 9.5% over standard MLLR. On the other hand, when the data size increases, we see the power of the proposed method degenerated. When the data size is large enough (40 or more), the regression class transform can be robustly estimated, and the adding of the information of confusion becomes ineffective and even detrimental.

In general the proposed algorithm can capture the HMM parameters' acoustic behaviors. And this information is helpful in adaptation. Incorporating such information in the EM framework of MLLR can help to reduce the risks of over-training and localize the transform matrix. The total effects showed quicker adaptation rate than standard MLLR algorithm, when adaptation data size is small.

5. CONCLUSION

HMM-state sequence confusion is an informative source of the acoustic context and the parameters' local behaviors. Our study uses this information in relaxing the distribution of state's output PDF, and uses this relaxed version of PDF in the general EM framework of MLLR. This yields a moderate improvement and relieves the over-training problem of standard MLLR when the adaptation data is small (in our case less than 30 seconds). The key point is that the reliable estimation of HMM state sequence confusion, which requires quite a lot of recognition results. And how to use more suitable prior information in the framework of MLLR needs further study. And further more the confusion characteristics can also be used in a discriminative way in optimizing acoustic model, which worth more study.

6. REFERENCE

1. V.V. Digalakis, D. Rtishev, and L. G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures", IEEE T-SAP vol.3, pp. 357-366, Sept. 1995
2. Sam-Joo Doh and Richard M. Stern, "Weighted principal component MLLR for speeaker adaptation," Proc. of Automatic Speech Recognition and Understanding Workshop (ASRU 99), Colorado, USA, 1999
3. Scott ShaoBing Chen & Peter DeSouza, "Speaker Adaptation by correlation", Eurospeech, Volume 4 pages 2111 – 2114, 1997
4. W. Chou, "Maximun a posterior linear regression with elliptically symmetric matrix variate priors", Eurospeech, vol. 1, pp. 1-4, 1999
5. Constantinos Boulis, Vassilios Diagalakis, "Fast speaker adaptation of large vocabulary continuous density HMM speech recognizer using a basis transform approach", ICASSP, II-989-992, 2000.
6. I. Csiszar and G. Tusnady, "Information geometry and alternating minimization procedures", sta. & Dec., Supp. Iss. No. 1, 1977.
7. Yumi Wakita, Harald Singler, Yoshinori Sagisaka, "Speech recognition using HMM-state confusion characteristics", Eurospeech, Volume 1 pages 7 – 10, 1997
8. Mingkuan Liu, Bo Xu, Taiyi Huang, etc., "Mandarin accent adaptation based on context-independent/context-dependent pronunciation modeling", ICASSP, II-1025 – 1028, 2000
9. Murat Saraclar, Harriet Nock, Sanjeev Kludanpur, "Pronunciation modeling by sharing Gaussian Densities Across Phonetic Models", Eurospeech, Volume 1, Page 515-518, 1999
10. 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, p.171-185, 1995.
11. William Byrne, Asela Gunawardana, "Discounted Likelihood Linear Regression for Rapid Adaptation", Eurospeech, Volume 1, Page 203-206, 1999