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

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

AN INVESTIGATION INTO SUBSPACE RAPID SPEAKER ADAPTATION

Michael Zhang Jun Xu
R&D department, InfoTalk Technology
Singapore 609917
{michael.zhang, jun.xu}@infotalkcorp.com

ABSTRACT

Speaker adaptation is an essential part of any state-of-the-art Automatic Speech Recognizer (ASR). Recently, more and more application requirements appear for embedded ASR. For these cases, a more compact speech model, *Subspace Distribution Clustering Hidden Markov Model* (SDCHMM) is used instead of *Continuous Density Hidden Markov Model* (CDHMM). In previous studies on SDCHMM adaptation, the subspace Gaussian pools of SDCHMM are the parameters to be adjusted for speaker variations. Alternatively, we try to employ the link table parameters of SDCHMM, which defines the tying structure in subspaces, to model the inter-speaker mismatch, with the Gaussian parameters maintained. Since the variation range for the parameters is highly limited, this method is potentially faster than conventional Gaussian pools adaptation. Comparative study on *Continuous Digital Dialing* (CDD) task shows that when data is seriously insufficient, link table adaptation is more effective than conventional methods, with 17% relative improvement in utterance accuracy rate, compared to 14% improvement by previous Gaussian adaptation. However, further improvement with more data is limited. When data size doubled, this method gave 21% improvement, compared to 30% improvement by conventional method.

1. INTRODUCTION

Compact acoustic model and rapid speaker adaptation are two important features for emerging personal ASR applications where storage and computation are at a premium, (e.g. PDAs, hand phones). Model size can be resolved, based on SDCHMM by clustering Gaussians' projection in subspace over the whole CDHMM. Fast speaker adaptation technologies have also been thoroughly studied on CDHMM. Where SDCHMM is concerned, recent studies have already shown that adaptation on subspace Gaussian groups is more effective than on full-space Gaussian groups [2,3], since the distribution of data over the full-space in often uneven, group estimation in subspace is more efficient than in the full space. However, these previous studies assume that the Gaussians' clustering in subspaces is unchanged for SDCHMM. This assumption will inevitably lead to the loss of accuracy for adaptation. In this paper, we try to compare the performance of different parameters' adaptation for SDCHMM: the subspace tying structures (the link table parameters) estimation and the subspace prototypes (the Gaussian pools parameters) estimation.

In the next section, three different adaptation strategies for SDCHMM are presented. Both *Maximum A Posteriori* (MAP) estimation and Maximum Likelihood Linear Regression (MLLR) in subspace are discussed. Evaluations on CDD tasks will be presented in section 3. A discussion and a conclusion will be presented in section 4 and section 5, respectively.

2. SUBSPACE SPEAKER ADAPTATION

The theory of SDCHMM is based on tying the parameters of CDHMM at subspace distribution [1]. Parameters in SDCHMM are briefly introduced here for further discussion.

2.1 Structure of SDCHMM

Assuming that all Gaussians of CDHMM can be denoted as:

$$\left\{\mathcal{N}_{s_i,m}\right\} = \left\{\mathcal{N}\left(\boldsymbol{\mu}_{s_i,m}, \boldsymbol{\Sigma}_{s_i,m}\right)\right\} \quad 1 \le i \le I, 1 \le m \le M \quad (1)$$

Where I is the number of states across the model and M is the Gaussian number for each distribution. $\boldsymbol{\mu}_{s_t,m}$ and $\boldsymbol{\Sigma}_{s_t,m}$ are the mean vector and covariance matrix of the Gaussian $\mathcal{N}_{s_i,m}$.

The output probability of a given observation sequence $\mathbf{O} = \left\{\mathbf{o}_1, \cdots, \mathbf{o}_t, \cdots \mathbf{o}_T\right\}$ along hidden state transfer sequence $S = \left\{s_1, \cdots, s_t, \cdots s_T\right\}$ can be calculated as follows:

$$\begin{aligned} P(\mathbf{O}|S) &= \prod_{t=1}^{T} p(\mathbf{o}_t|s_t) \\ &= \prod_{t=1}^{T} \sum_{m=1}^{M} w_{s_t,m} \mathcal{N}\left(\mathbf{o}_t \middle| \boldsymbol{\mu}_{s_t,m}, \boldsymbol{\Sigma}_{s_t,m}\right) \end{aligned} \quad (2)$$

Where $w_{s_t,m}$ are the mixture weights and satisfy $\sum_{m=1}^{M} w_{s_t,m} = 1$.

If the correlation between dimensions can be simplified and modeled by a block diagonal matrix, the full observation space $\Re^D$ (where D is the observation dimension) can be decomposed into K orthogonal subspaces $\Re^{d_k}$ with dimension d_k. And $\sum_{k=1}^{K} d_k = D$.

0-7803-8678-7/04/$20.00 ©2004 IEEE

The probability in Eq.(2) can be deduced as:

$$p(\mathbf{o}_t|s_t) = \sum_{m=1}^{M} w_{s_t,m} \left(\prod_{k=1}^{K} \mathcal{N}\left(\mathbf{o}_{t,k} \middle| \boldsymbol{\mu}_{s_t,m,k}, \boldsymbol{\Sigma}_{s_t,m,k}\right) \right) \quad (3)$$

Where $\mathbf{o}_{t,k}$, $\boldsymbol{\mu}_{s_t,m,k}$, $\boldsymbol{\Sigma}_{s_t,m,k}$ are the projections of the full vector $\mathbf{o}_t$, $\boldsymbol{\mu}_{s_t,m}$, $\boldsymbol{\Sigma}_{s_t,m}$ into the subspace $\Re^{d_k}$, respectively.

The projections of CDHMM distributions over all models into each orthogonal subspace can be tied into a smaller number of Gaussian prototypes to form SDCHMM. Satisfied acoustic resolution can then be maintained. The probability calculation is approximated as:

$$P(\mathbf{o}_t|s_t) \approx \sum_{m=1}^{M} w_{s_t,m} \left(\prod_{k=1}^{K} \mathcal{N}\left(\mathbf{o}_{t,k} \middle| \boldsymbol{\mu}_k^{tied}, \boldsymbol{\Sigma}_k^{tied}\right) \right) \quad (4)$$

SDCHMM can be considered as a good approximation to conventional CDHMM. The subspace distribution tying structure of SDCHMM can be derived from tying CDHMM Gaussians into each subspace based on their similarities, or directly trained from speech data [1]. The parameters in SDCHMM can then be denoted with a sharing pool of these Gaussians prototypes:

$$\{\mathcal{N}^n\} = \{(\boldsymbol{\mu}^n, \boldsymbol{\Sigma}^n)\}, 1 \le n \le N \quad (5)$$

(Where N is the total number of Gaussian prototypes.)

Moreover, $Link(s,m,k)$ is employed here to denote the linkage relationship between these Gaussian projections into the orthogonal subspaces $\Re^{d_k}$, $\{\mathcal{N}_k^n\} = \{(\boldsymbol{\mu}_k^n, \boldsymbol{\Sigma}_k^n)\}, 1 \le n \le N$, and these full prototypes $\{\mathcal{N}^n\}$.

2.2 Adaptation for SDCHMM parameters

There are different alternatives for adaptation on SDCHMM:

1. Overall adaptation:

Gaussians of CDHMM can easily be restored from SDCHMM. All calculation can be done on this CDHMM following the conventional routine. The adapted SDCHMM can be converted from this intermediate CDHMM.

Since the conversion from CDHMM to SDCHMM includes an iterative clustering process, this scheme is not applicable in many cases. We only study it for comparison with other methods.

2. Gaussian Prototypes adaptation:

Both MAP and MLLR can be employed for adapting the Gaussian Prototypes in SDCHMM.

MAP adaptation:

The method discussed here could be viewed as a simplified version of the method proposed in [2]. For SDCHMM, the Gaussian occupation probability at time t is calculated as:

$$\gamma_{s_t,m} = \frac{w_{s_t,m}\mathcal{N}\left(\mathbf{o}_t \middle| \boldsymbol{\mu}_{s_t,m}^{SDCHMM}, \boldsymbol{\Sigma}_{s_t,m}^{SDCHMM}\right)}{\sum_{l=1}^{M} w_{s_t,l}\mathcal{N}\left(\mathbf{o}_t \middle| \boldsymbol{\mu}_{s_t,l}^{SDCHMM}, \boldsymbol{\Sigma}_{s_t,l}^{SDCHMM}\right)}$$

$$= \frac{w_{s_t,m}\prod_{k=1}^{K}\mathcal{N}\left(\mathbf{o}_{t,k} \middle| \boldsymbol{\mu}_k^{Link(s_t,m,k)}, \boldsymbol{\Sigma}_k^{Link(s_t,m,k)}\right)}{\sum_{l=1}^{M} w_{s_t,l}\prod_{k=1}^{K}\mathcal{N}\left(\mathbf{o}_{t,k} \middle| \boldsymbol{\mu}_k^{Link(s_t,l,k)}, \boldsymbol{\Sigma}_k^{Link(s_t,l,k)}\right)} \quad (6)$$

Only mean vector adaptation is considered in this study. The subspace MAP adaptation is calculated in the form of:

$$\hat{\boldsymbol{\mu}}_k^n = \frac{\tau\boldsymbol{\mu}_k^n + \sum_{t=1}^{T}\sum_{m=1}^{M}\delta(s_t,m,k)\gamma_{s_t,m}\mathbf{o}_{t,k}}{\tau + \sum_{t=1}^{T}\sum_{m=1}^{M}\delta(s_t,m,k)\gamma_{s_t,m}} \quad (7)$$

$$where\ \delta(s_t,m,k) = \begin{cases} 1 & when\ \ Link(s_t,m,k) = n \\ 0 & otherwise \end{cases}$$

Where τ is an empirical constant, which alters the contribution between prior mean and the MLE result.

MLLR adaptation

The subspace tied Gaussian prototype in SDCHMM has already provided some Gaussian grouping knowledge in each subspace. However, N is usually too large to gain a robust estimation in fast adaptation. Regression classes imposed on these Gaussian prototypes in each subspace are indispensable. Suppose there are C independent regression classes $\{\Phi_c\}, 1 \le c \le C$ for each subspace, MLLR transforms can be denoted as:

$$\begin{aligned} &\{T_k^c\} = \{(\mathbf{A}_k^c, \mathbf{B}_k^c)\}\ where\ 1 \le c \le C, 1 \le k \le K \\ &\hat{\boldsymbol{\mu}}_k^n = \hat{\boldsymbol{\mu}}_k^n\mathbf{A}_k^c + \mathbf{B}_k^c\ \ if\ \ \mathcal{N}_k^n \in c \end{aligned} \quad (8)$$

The Expectation-Maximization (EM) auxiliary function is:

$$\begin{aligned} &Q\left(\hat{\mathbf{A}}_k^c, \hat{\mathbf{B}}_k^c \middle| \mathbf{A}_k^c, \mathbf{B}_k^c\right) = E\left\{\log p\left(\mathbf{O}, S \middle| \hat{\mathbf{A}}_k^c, \hat{\mathbf{B}}_k^c\right) \middle| \mathbf{A}_k^c, \mathbf{B}_k^c\right\} \\ &= \sum_{t=1}^{T}\sum_{m=1}^{M}\theta(n,c)\gamma_{s_t,m}\log\mathcal{N}\left(\mathbf{o}_{t,k} \middle| \mu_k^n, \Sigma_k^n, \hat{\mathbf{A}}_k^c, \hat{\mathbf{B}}_k^c\right) \\ &n = Link(s_t,m,k) \\ &\theta(n,c) = \begin{cases} 1 & when\ \ \mathcal{N}_k^n \in \Phi_c \\ 0 & otherwise \end{cases} \end{aligned} \quad (9)$$

Where $\gamma_{s_t,m}$ is the occupation of Gaussian $\mathcal{N}_{s_i,m}$ at time t in full space. Its calculation follows Eq. (6) with transforms $\{\mathbf{A}_k^c, \mathbf{B}_k^c\}$ imposed in the corresponding subspaces.

By maximizing the function in Eq. (9), the subspace transform $\left\{\hat{\mathbf{A}}_k^c, \hat{\mathbf{B}}_k^c\right\}$ can be updated in a similar manner as conventional MLLR in full space.

3. Link-Table adaptation:

Since the sharing Gaussian pool in SDCHMM is converted from a speaker-independent CDHMM or from mass training data, we could consider it contains enough variations due to speaker differences. Then the mismatch introduced by different speakers only reflects on the index change in the SDCHMM link table with the original Gaussian pool unchanged. The link table of the result SDCHMM is reconstructed as follows:

$$Link(s,m,k) = \arg\min_{1\le i\le N} dist\left(\left(\boldsymbol{\mu}_{s,m,k}, \boldsymbol{\Sigma}_{s,m,k}\right), \left(\boldsymbol{\mu}_k^i, \boldsymbol{\Sigma}_k^i\right)\right)$$

$$\approx \arg\min_{1\le i\le N} \left\| \boldsymbol{\mu}_k^i - \frac{\tau\boldsymbol{\mu}_k^n + \sum_{t=1}^{T}\sum_{m=1}^{M}\delta(s_t,m,k)\gamma_{s_t,m}\mathbf{o}_{t,k}}{\tau + \sum_{t=1}^{T}\sum_{m=1}^{M}\delta(s_t,m,k)\gamma_{s_t,m}} \right\| \qquad (10)$$

The assumption that only link table parameters contain the inter-speaker information imposes a strong restriction for the estimation process, which leads to a faster convergence. We expect this method could be useful where the amount of data is very limited.

3. EXPERIMENTS

The schemes above were investigated on the Mandarin *Continuous Digital Dialing* recognition in telephone applications. More than 200,000 CDD utterances were collected from the telephone lines (on both fixed line and wireline) and were transcribed manually to serve as training data for the original acoustic model using HTK 3.0 utilities as the training tool. The front-end of the system was configured to extract *linear predictive cepstral coefficients* (LPCC) feature and its delta and delta-delta counterparts.

The test set was collected through several types of commonly used hand phone sets. Approximately 200 CDD strings of different lengths that correspond to typical application scenarios for voice dialing in a mobile environment were designed. The test set involved eight male and female speakers, each reading the same pre-designed CDD strings. Some utterances were discarded in the post transcription process due to mispronunciation and other errors during recording. With the range of 90 to 200 CDD strings from each speaker, 1408 testing utterances were obtained.

In order to focus on acoustic resolution, a grammar of free digit length from 1 to 12 was adopted for decoding. No additional rules of digital dialing were imposed on these experiments. For application reasons, system performance improvement was evaluated in utterance accuracy rate for CDD string instead of commonly used WER reduction.

The length of these testing CDD strings varied from 3 digits to 12 digits. The occurrences of different string lengths in the test set are shown in Table (1). Since eighty five percent of these utterances have more than 7 digits, 'open grammar' recognition on this test-set is a very difficult task indeed.

Table (1) Statistics of CDD string length in the test set

String Length (Digits)	3	4	5	7	8	10	11	12
Occurs In Test Set	126	21	74	143	145	214	554	131

To find the relationship between data size and the adaptation performance, we used the first 4, 6, and 8 utterances from each speaker to serve as different-sized adaptation sets. These sets contained 40, 49, and 70 digits, respectively. These utterances were removed from evaluation. All tests in this paper were conducted in supervised mode.

Table (2) is the result of conventional MAP adaptation on CDHMM parameters with varying data size. It served as a baseline for further comparison. It also gives us some guidance on the amount of data that needed for Gaussian adaptation.

Table (2) Conventional MAP adaptation on CDHMM.
(In Utterance Accuracy)

	Baseline	4Utt Adapt (40Digits)	6Utt Adapt (49Digits)	8Utt Adapt (70Digits)
LXL	59.38	65.31	68.56	67.19
DMH	35.42	71.94	75.26	79.69
WSP	53.13	55.10	57.22	59.90
XJ	60.12	65.45	72.26	77.98
CT	58.15	60.44	62.50	65.76
MB	72.61	73.32	75.12	75.80
PCL	46.59	47.58	48.96	51.70
ZJ	52.44	65.30	67.28	73.17
Average	54.41	62.77	63.23	68.38

Table (2) shows that the adaptation data set that consists of the first 8 utterances (70 digits) gave the best performance. The system performance improved absolutely by almost 14%. Experiments proved that more data would definitely result better performance. However, more than eight utterances for adaptation are not convenient for many practical applications. Therefore, we consider that 8Utt dataset is moderate for this task and 4Utt dataset is obviously insufficient for this task.

Table (3) gives the result of MAP adaptation on SDCHMM parameters. All the three adaptation strategies discussed in the 2nd part were tested here. Among these methods, the overall adaptation is not applicable in real fields due to its computation complexity. We only tested it for comparison with other methods. Provided with enough data, the overall method should give the best performance theoretically. Compared to adaptations on CDHMM, the overall method should provide a similar performance on SDCHMM. This can be proved by comparing the results from Table (2) and Table (3). Conventional MAP on CDHMM and the overall method on SDCHMM made about 25.7% and 28.1% relative improvements on average, respectively, with the 8Utt data set. This performance can be viewed as an upper limit for other two methods tested on SDCHMM.

However, with limited data size in our tests, Gaussian prototype adaptation in subspaces has shown to be better than the over all methods. Results in Table (3) showed that about 30.5% relative improvement was obtained by Gaussian prototypes adaptation while 28% improvement was made by the overall method on the same 8Utt dataset. We consider that this occurred due to the sparse data problem in estimations for these experiments.

Table (3) Different Strategies for subspace MAP adaptation on SDCHMM. (In Utterances Accuracy)

	Base-line	Overall Adaptation (Mean)		LinkTable Adaptation Gaussian		Prototype Adaptation (Mean)	
		4Utt Adapt	8Utt Adapt	4Utt Adapt	8Utt Adapt	4Utt Adapt	8Utt Adapt
LXL	55.21	55.10	64.58	56.63	57.29	57.65	63.54
DMH	40.63	59.18	79.17	72.45	76.04	67.86	81.77
WSP	48.44	48.47	54.17	50.00	52.08	52.04	57.81
XJ	55.36	63.37	76.33	69.19	72.62	65.12	75.60
CT	59.78	59.04	65.80	63.30	65.76	61.17	66.85
MB	71.34	71.43	77.80	74.53	75.80	72.05	78.98
PCL	40.34	41.11	48.50	43.89	45.45	43.33	49.43
ZJ	40.24	43.02	65.54	53.49	54.88	48.84	69.51
Average	51.77	55.64	66.31	60.65	62.72	58.98	67.56

Obviously, the link-table adaptation outperforms the other two methods when data is highly insufficient. From table (3), it provides 17.15% relative improvements on 4Utt dataset, compared to 7.50% and 13.93% improvements by Overall and Gaussian adaptation respectively. However, when more data available, this method cannot describe the variations between different speakers as effectively as the other two methods. With 8Utt dataset, link-table method could only achieve 21.15% relative improvements, compared to 30.50% by prototype adaptation and 28.09% by overall method.

4. DISCUSSION

Comparing the results from Table (2) and Table (3) on the same 8Utt adaptation dataset, MAP for Gaussians provides 13.97% and 15.79% improvement absolutely (or 25.68% and 30.50% relatively) on CDHMM and SDCHMM respectively, which means subspace MAP adaptation on Gaussian prototypes provides better performance than full space MAP adaptation on Gaussians. This is probably because SDCHMM provides a subspace tying-structure among all Gaussians across the model. This structure makes parameter estimation more reliable in MAP, especially for fast speaker adaptation.

From Table (3), we found that link table adaptation is suitable for occasions where data is seriously insufficient. When more data available, subspace Gaussian adaptation obviously gives better performance than the link table adaptation. This result implies that the acoustic variation between speakers is quite complex that only changing the link table index in subspaces is not precise enough to reflect this mismatch. At the same time, overall adaptation also gives poorer performance than Gaussian prototype adaptation. This probably suggests that tying structures adaptation have a minor contribution for SDCHMM adaptation.

For the CDD task, we can design the corpus carefully to guarantee that all models can be seen in the adaptation process. So MAP is adopted here. As far as MLLR is concerned, previous study has shown that MLLR with Subspace Regression Class (SSRC) achieves better performance than conventional MLLR with Full Space Regression Class (FSRC) [3]. With SSRC, transform parameters in different subspace are estimated with different neighbor Gaussians depending on their acoustic similarity in the subspace. We can expect SSRC to also give good performance for Gaussian pools adaptation on SDCHMM.

5. CONCLUSION

In this paper, we implemented 3 adaptation strategies with SDCHMM parameters. With moderate data size, experimental results show that the previously proposed MAP estimation in subspaces achieves the best performance. And it outperforms MAP in conventional full space. The introduction of subspace tying relation of SDCHMM Gaussians reduces the number of parameters for estimation so that the results becomes more reliable in fast speaker adaptation. When available data is obviously insufficient, the new link table adaptation behaves better than the other two methods, since the range of the estimation result is highly restricted in this method. However, this restriction also brings a fast saturate shortcoming for this method. In practice, we have to make balance between these factors: accuracy, computational complexity, and the available data size. For CDD adaptation task, we found that Gaussian prototypes' adaptation with fixed tying structure in subspaces gives a satisfied performance. This method can work effectively with SDCHMM where speed and memory are mostly concerned.

6. EFERENCES

[1] B. Mak**,** *"Towards A Compact Speech Recognizer: Subspace Distribution Clustering Hidden Markov Model"*, PhD Thesis, OGI, April 1998.

[2] K. M. Wong, B. Mak, "*MAP Adaptation with Subspace Regression Classes and Tying",* in Proc.IEEE on Acoustics, Speech, and Signal Processing, 2000.

[3] K.Wong and B.Mak, *"Rapid Speaker Adaptation Using MLLR and Subspace Regression Classes"*, Eurospeech'2001, Aalborg, Denmark, Sept. 2001.

[4] Gauvain, J.-L., Lee, C.-H., 1994. *"Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains"*, IEEE Trans. Speech Audio Process. 2, 291-298.

[5] K.Shinoda and C-H.Lee, *"Structural MAP speaker adaptation using hierarchical priors"*, in Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, 1997.

[6] C. J. Leggetter and P. C. Woodland, *"Speaker adaptation of continuous density HMMs using multivariate linear regression"*, in Proc. of ICSLP 94, Japan, Sep.1994,

[7] M. Gales, "*The generation and use of regression class trees for MLLR adaptation"*, Technical Report. TR263, Cambridge University Engineering Department, 1996.