PREDICTIVE SPEAKER ADAPTATION AND ITS PRIOR TRAINING

Dieu Tran and Ken-ichi Iso

C & C Media Research Laboratories, NEC Corporation
4-1-1 Miyazaki, Miyamae-ku, Kawasaki 216-8555, JAPAN
E-mail: {tran, iso}@ccm.cl.nec.co.jp

ABSTRACT

In this paper, we propose a speaker adaptation technique called *Predictive Speaker Adaptation* (PSA) in which speaker dependent HMM(SD-HMM) for a new speaker is predicted using adaptation utterances and a speaker independent HMM(SI-HMM). The method requires a prior training, during which parameters of the prediction function are optimally estimated with many speaker's SD-HMMs and their adaptation utterances as examples. This method revealed to be efficient for small amount of adaptation data. 60,000-word recognition experiments reported a word error-rate reduction of 16% when only 10 adaptation words were used.

1. INTRODUCTION

Speaker adaptation techniques such as MAP [1], while approaching performance of a SD-HMM provided enough speaker-specific adaptation data are available, may not be suited for some practical applications in which very small amount of adaptation data is available. These conventional methods only adapt gaussians in the SI-HMM which have observations in the adaptation utterances. For small amount of adaptation data, only few gaussians are observed while most remain unseen. Even seen ones are poorly trained since the number of observations for these gaussians is usually small.

Recently, some methods have come up, in which many speaker's SD-HMMs and multiple linear regression are used to find speaker-independent correlation among speech units in order to update unseen parameters in a predictive manner [2]-[5].

Our approach is separated into two parts, a prior training session and adaptation session. In the prior training session, parameters of the predictive adaptation method are robustly estimated using many speaker's SD-HMMs and their utterances for adaptation words. In the adaptation session, the pre-calculated parameters are used for adapting SI-HMM to the new speaker (see fig.1). In the following, algorithms in both sessions are explained in detail.

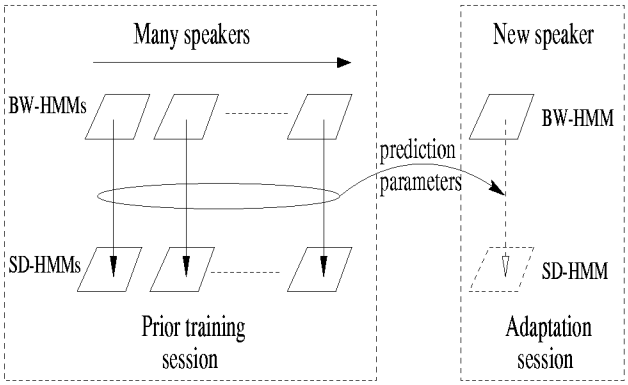

Figure 1: Basic idea of PSA.

2. ADAPTATION

The adaptation session consists of three steps:

1.Mean Training
2.Centroid Adaptation
3.Prediction with predetermined coefficients

First Mean Training is carried out. This is a simple Baum-Welch training of SI-HMM using the new speaker's adaptation utterances. In the next step, Centroid Adaptation is applied as a kind of preprocessing to increase the baseline performance of SI-HMM. The real part of adaptation is done at last, using the output models of the two previous stages and the prediction coefficients from the prior training. Each of these three steps will be explained in the following subsections.

2.1. Mean training

A SI-HMM is first retrained using Baum-Welch reestimation on the adaptation data provided by the new speaker. Due to the limited amount of training material, only the mean of gaussians are updated. The resulting HMM is referred to BW-HMM in the following explanations.

2.2. Centroid Adaptation

To remove the global offset existing between the speaker-independent acoustic space and the new speaker's one, Centroid Adaptation has been introduced. Fig.2 illustrates the idea. All means in the SI-HMM are "moved" to the direction of the new speaker using a global displacement vector $\Delta\overline{\mu}$. The shifted means are called *centroid adapted mean* $\mu_i^{(CA)}$:

$$\mu_i^{(CA)} = \mu_i^{(SI)} + \Delta\overline{\mu}, \qquad i = 1..N \tag{1}$$

where

$$\Delta\overline{\mu} = \frac{1}{M}\sum_{m=1}^{M}(\mu_m^{(BW)} - \mu_m^{(SI)}) \tag{2}$$

N is the total number of gaussians and $\mu_i^{(SI)}$ denotes the mean of the i-th gaussian in the SI-HMM. $\{\mu_m^{(BW)} | m = 1..M\}$ is the set of retrained means in BW-HMM e.g. the set of means of gaussians for which there are observations in the adaptation words (seen gaussians).

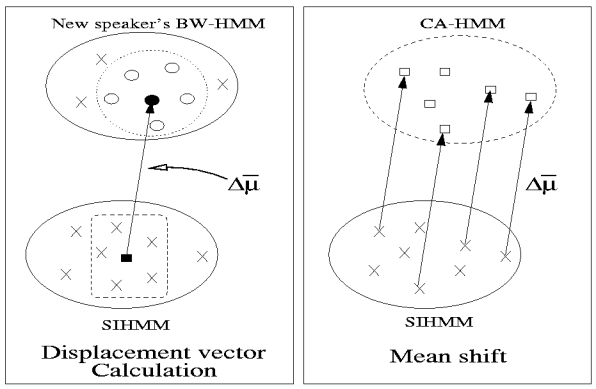

Figure 2: Geometric representation of Centroid Adaptation. Crosses represent gaussian means in SI-HMM, circles represent the retrained means. The filled rectangle represents the average of the means in SI-HMM for which there are observations in the adaptation words, the filled circle the average of retrained means.

2.3. Prediction with predetermined coefficients

The adapted means for the new speaker are obtained with the following equation:

$$\hat{\mu}_i = \mu_i^{(CA)} + \Delta\hat{\mu}_i \tag{3}$$

$\Delta\hat{\mu}_i$ is an adaptation vector. It is linearly predicted as follows:

$$\Delta\hat{\mu}_i = \sum_{m\in N(i)} w_{im}\Delta\mu_m \tag{4}$$

$\Delta\mu_m$ are the *difference vectors* for seen gaussians. They are defined as the difference between a retrained mean in BW-HMM and its corresponding centroid adapted mean:

$$\Delta\mu_m = \mu_m^{(BW)} - \mu_m^{(CA)}, \qquad m = 1..M \tag{5}$$

N(i) is a small subset of all difference vectors called *neighborhood set* and w_{im} the associated *prediction coefficients*. The unknown parameters $\{w_{im}, N(i) | i = 1..N, m \in N(i)\}$ are predetermined in the training session described in the next section.

3. PRIOR TRAINING

The aim of the prior training is to determine the parameters $\{w_{im}, N(i) | i = 1..N, m \in N(i)\}$ before adaptation (see eq.4). To obtain speaker-independent parameters, a large set of S training speakers is needed for estimation. We first prepare these S speakers' fully trained SD-HMM using Baum-Welch training on a large amount of speaker-specific data. In addition, using each speaker's adaptation utterances (same for all speakers), a population of speaker-specific BW-HMM is built (by means of Baum-Welch re-estimation) from the initial SI-HMM. Since the SD-HMM for each training speaker is available, eq.4 can be rewritten as:

$$\Delta\mu_i{}^{(s)} = \sum_{m\in N(i)} w_{im}\Delta\mu_m^{(s)}, \qquad s = 1..S \tag{6}$$

where $\Delta\mu_i{}^{(s)}$ is computed from the SD-HMM and $\Delta\mu_m^{(s)}$ from the BW-HMM of training speaker s:

$$\Delta\mu_i{}^{(s)} = \mu_i^{(SD)} - \mu_i^{(CA)}$$

$$\Delta\mu_m^{(s)} = \mu_m^{(BW)} - \mu_m^{(CA)}$$

Making use of all S pairs of $(\Delta\mu_i{}^{(s)}, \Delta\mu_m^{(s)})$ as training examples for eq.6, a reliable estimate of the prediction parameters can be obtained by minimizing the mean square error:

$$L_i = \frac{1}{2}\sum_{s=1}^{S}\|\Delta\mu_i^{(s)} - \sum_{m\in N(i)} w_{im}\Delta\mu_m^{(s)}\|^2 \tag{7}$$

with respect to $\{w_{im}, N(i)\}$.
We solved eq.7 in two passes. First the neighborhood set N(i) is determined for each i. $N(i)$ denotes the group of seen gaussians which are most likely to contribute to the prediction of gaussian i in the SD-HMM. Gaussians which are strongly correlated with i are the most appropriate for prediction [2]-[4]. The correlation coefficients c_{im} between $\Delta\mu_i$ and $\Delta\mu_m$ are computed according to a cosine metric:

$$c_{im} = \{\frac{1}{S}\sum_{s=1}^{S}\frac{\Delta\mu_i^{(s)}\Delta\mu_m^{(s)}}{\|\Delta\mu_i^{(s)}\|\ \|\Delta\mu_m^{(s)}\|}\} \tag{8}$$

The K ($K \ll M$) $\Delta\mu_m$ which have the largest correlation coefficient value are selected as the neighbors of i. Once $N(i) = \{m_1, m_2, .., m_K\}$ is set, w_{im} are obtained by solving eq.7 using the following iteration:
For $i = 1..N$,

1. Residual initialization: $\delta\mu_i^{(s)} = \Delta\mu_i^{(s)}$
2. Contribution of each neighbor in reducing $\delta\mu_i^{(s)}$:
   For $l = 1..K$,

$$\hat{w}_{im_l} = \frac{\sum_{s=1}^{S} \delta\mu_i^{(s)} \Delta\mu_{m_l}^{(s)}}{\sum_{s=1}^{S} \left\| \Delta\mu_{m_l}^{(s)} \right\|^2}$$

$$\delta\mu_i^{(s)} = \delta\mu_i^{(s)} - \hat{w}_{im_l} \Delta\mu_{m_l}^{(s)}$$

4. EXPERIMENTAL EVALUATION

PSA was evaluated on 60,000 isolated word recognition experiments. Table 1 contains the experimental settings. For prior training, we used DB1. Using all utterances, a reliable male SI-HMM was trained. Each speaker's SD-HMM was also obtained with Baum-Welch training on his 2064 word utterances. Furthermore, for each speaker, using 10 to 250 words from his 2064 word utterances as adaptation utterances, his BW-HMM was trained from the male SI-HMM. We estimated the prediction function parameters $\{w_{im}, N(i)\}$ based on these SD-HMM and BW-HMM pairs for 59 speakers. For speaker adaptation experiments, 5 male speakers different from the 59 training speakers were used as new speakers. These 5 test speakers uttered adaptation utterances (DB2) which were identical to the ones used in the prior training. They also provided utterances (DB3) which were different from DB2 for testing.

Table 2 gives information about the number of seen gaussians for various amount of adaptation data. Because of the large number of seen gaussians, using all prediction coefficients for adaptation requires a too large amount of storage capacity. For 20 adaptation words for example, one matrix of size $2052 * 908$ for the prediction coefficients ($\{w_{im}/i = 1..2052, m = 1..908\}$) would be needed. Moreover, as can be seen in fig.3, the prediction coefficients' values decrease rapidly for every amount of adaptation words which indicates that only the contribution of the first coefficients are significant. In all our experiments we set the number of neighbors to 10.

From fig.4 we can conclude that the proposed algorithm works properly. Recognition accuracy increases steadily with the augmentation of the amount of adaptation words. In addition, the effectiveness of Centroid Adaptation (CA) to move the SI-HMM closer to the new speaker is confirmed. Best results are obtained when PSA is applied with CA. To compare PSA with other speaker adaptation methods,

Feature vector (21 dim):	ΔPow, Cep, ΔCep
Sampling rate:	11.025kHz
Frame size:	23.2 ms
Frame shift:	11.6 ms
Preemphasis filter:	$1 - z^{-1}$
Speech database	
DB1 (59 male speakers):	2064 words/speaker
DB2 (5 male speakers):	10-250 words / speaker
DB3 (5 male speakers):	250 words / speaker
HMM	256 demi-syllables, 4 states/demi-syllable, 2 gaussians/state

Table 1: Experimental conditions

Nb Adaptation words	Seen/Total nb gaussians
10	588/2052 (28.6 %)
20	908/2052 (44.2 %)
50	1328/2052 (64.7 %)
100	1548/2052 (75.4 %)
250	1692/2052 (82.4 %)

Table 2: Percentage of seen gaussians.

we chose the Spectral Interpolation with MAP adaptation (SPINT+MAP) [6] and the Speaker Adaptation with autonomous control using tree structure [7] (TREE). In SPINT+MAP, seen gaussians are adapted with MAP. Parameters of unseen ones are estimated from the interpolation of neighboring seen gaussians. In TREE, spatial relations between gaussians are captured using tree clustering. The leaf nodes of the tree correspond to all gaussians. For each node, an adaptation shift shared by all leaf nodes that fall below it is calculated. Depending on the amount of adaptation utterances, appropriate nodes in the tree are selected autonomously and their shifts are used to adapt all gaussians which reside below those nodes. To further improve the first method, CA was introduced into SPINT+MAP. Amelioration of recognition accuracy for small amount of adaptation data is achieved with this combination (see fig.5). Finally, results in fig.6 show that PSA outperforms both methods for any number of adaptation words. This proves that the use of additional prior knowledge improves consistently the recognition performance.

5. CONCLUSION

In this paper, we have presented a new approach of adaptation. Experimental results confirm its efficiency for small adaptation data. Unlike existing predictive adaptation techniques [2]-[4] which use many SD-HMMs only, the proposed method includes adaptation utterances from many training speakers as additional

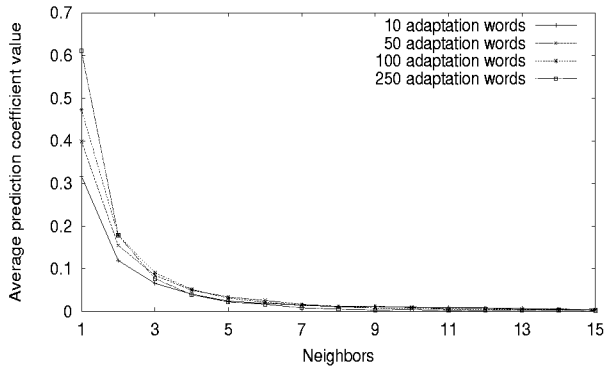

Figure 3: Prediction coefficient value for the first 15 closest neighbors (averaged over all gaussians in the model)

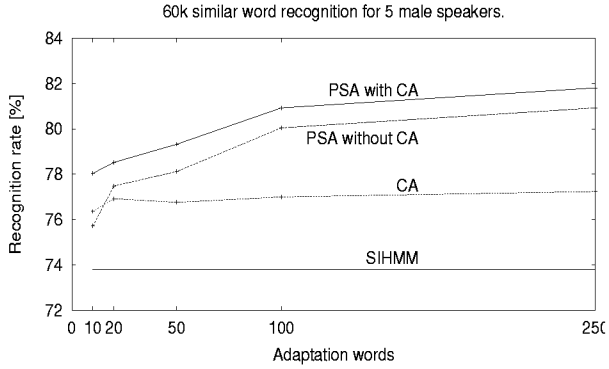

Figure 4: Effect of CA in PSA.

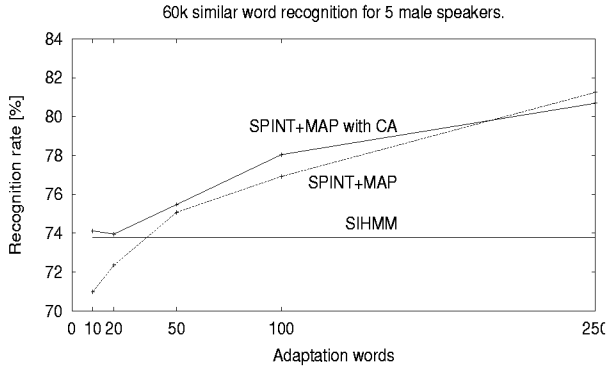

Figure 5: Effect of combining CA with SPINT+MAP.

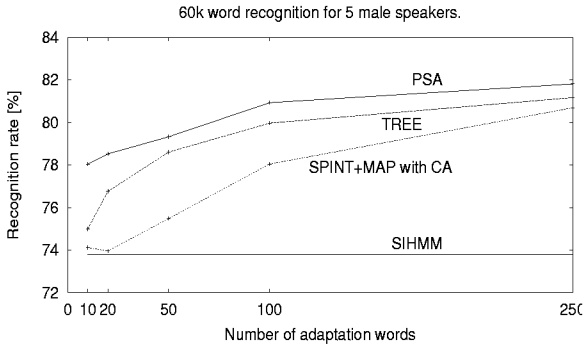

Figure 6: Recognition results.

prior knowledge. In this way, more robust and stable estimation of the prediction coefficients are obtained. This in turn means that to change the adaptation vocabulary, new utterances from the training speakers have to be collected.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Takao Watanabe for his continuous encouragement. Many thanks as well goes to Mr. Koichi Shinoda for the useful discussions and for providing the tools for 'Spectral Interpolation' and 'Tree Adaptation'.

REFERENCES

[1]J-L.Gauvain, G-H.Lee,"Maximum *a Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains"*IEEE Transactions on Speech and Audio Processing*, Vol.2, pp.291-298,1994.

[2]Y.Obuchi, A.Amano, N.Hataoka, "A Novel Speaker Adaptation Algorithm and its Implementation on a Risc Microprocessor"*1997 IEEE Workshop on Automatic Speech Recognition And Understanding Proceedings*, pp.442-449,1997.

[3]A.M.Ahadi, P.C.Woodland, "Rapid Speaker Adaptation Using Model Prediction"*Proc. ICASSP-95*,pp.684-687,1995.

[4]S.S.Chen and P.DeSouza, "Speaker Adaptation by Correlation"*Eurospeech 97*,pp.2111-2114, 1997.

[5]S.Furui, "A Training Procedure for Isolated Word Recognition Systems"*IEEE Transactions on Acoustic, Speech, Signal Processing*, ASSP-28, pp.129-136, 1980.

[6]K.Shinoda, K.Iso, T.Watanabe, "Speaker Adaptation for Demi-syllable Based Continuous Density HMM"*Proc. ICASSP-91*,pp.857-860,1991.

[7]K.Shinoda, T.Watanabe, "Speaker Adaptation with Autonomous Control using Tree Structure", *Eurospeech 95*,pp.1143-1146,1995.