UNSUPERVISED CROSS-VALIDATION ADAPTATION ALGORITHMS FOR IMPROVED ADAPTATION PERFORMANCE

Takahiro Shinozaki, Yu Kubota, Sadaoki Furui

Department of Computer Science, Tokyo Institute of Technology, Japan

ABSTRACT

An unsupervised cross-validation adaptation algorithm and its variation are proposed that introduce the idea of cross-validation in the unsupervised batch-mode adaptation framework to improve the adaptation performance. The first algorithm is constructed on a general adaptation technique such as MLLR and can be used in combination with any adaptation method. The second algorithm is a modified version of the first algorithm and works with lower computational cost by assuming MLLR. These algorithms are extensions of our previously proposed CV training methods and are useful to suppress the negative effect of the conventional unsupervised batch-mode adaptation process that reinforces the errors included in automatic transcriptions. The proposed algorithms were evaluated in domain adaptation, speaker adaptation, and in their combination for large vocabulary spontaneous speech recognition. When the domain and speaker adaptations were combined using a read speech initial model, the relative word error rate reduction by the proposed method was 29% whereas the reduction by the conventional approach was 23%.

Index Terms— Unsupervised adaptation, cross-validation, MLLR, MAP, computational cost

1. INTRODUCTION

Unsupervised adaptation is a useful technique to achieve high recognition performance without requiring a reference for adaptation. Usually, unsupervised adaptation is performed by first running a speech recognizer for input utterances to obtain a recognition hypothesis and then applying a supervised adaptation technique such as MLLR [1] to update the model parameters using the recognition hypothesis as a transcript. This strategy is used both for on-line adaptation and for off-line adaptation. For the off-line adaptation, a common strategy is the batch-mode adaptation where a set of input utterances are decoded and then a model is updated using all the hypotheses. The process is iterated several times for higher recognition performance [2].

While the batch-mode strategy is effective, a problem is that the hypothesis includes errors and the model parameters are adjusted using not only the correct labels but also the errors. When the adaptation is iterated, the errors are reinforced during the iteration since the decoding step and the update step are repeated using the same data. Even for correct labels, the over fitting problems are unavoidable since a bias in the parameter estimation is intensified during the iteration. Because of these, the adaptation performance is limited.

To address these problems, we propose two unsupervised cross-validation (CV) adaptation schemes that introduce the idea of CV into the iterative unsupervised adaptation framework to efficiently avoid the data overlap between the decoding step and the model update step. The first version is constructed on a general adaptation technique and has an advantage that it is independent from the details of the underlying adaptation method and can be used in combination with any adaptation method. The second algorithm is a variation of the first algorithm that is specially designed for the MLLR method and has an advantage that it can largely reduce the computational cost. In this paper, the first algorithm is referred to as a general version and the second as an efficient version.

The proposed adaptation algorithms are similar to the CV based gradient estimation method for MMI training [3] and to our previously proposed CV-EM supervised training method [4] in that CV is introduced in the iterative model estimation framework. The differences are that the proposed algorithms are unsupervised adaptation schemes and the CV technique is used to compute hypotheses rather than the gradients as in the MMI method or the sufficient statistics for true transcripts as in CV-EM.

In experiments, the proposed CV methods are applied to speaker adaptations for large vocabulary spontaneous speech recognition. In addition, they are also evaluated in the context of unsupervised domain adaptation in which a speaker independent spontaneous speech model is made from a read speech model using spontaneous utterances from multiple speakers as the domain adaptation data.

The organization of this paper is as follows. The general version of the unsupervised CV adaptation scheme is proposed in Section 2. The efficient version for the CV MLLR adaptation is described in Section 3. Experimental conditions are described in Section 4 and the results are shown in Section 5. Conclusions and future works are given in Section 6.

2. UNSUPERVISED CROSS-VALIDATION (CV) ADAPTATION ALGORITHM

Figure 1 shows the procedure of the proposed K-fold unsupervised cross-validation (CV) adaptation method. In the CV adaptation, the input speech data is first randomly partitioned into K exclusive subsets so that each subset has roughly the same size. As in the conventional batch-mode adaptation, CV adaptation repeats the decoding step and the model update step. The first decoding step is basically the same as the batch-mode adaptation and the K subsets are processed using the same initial model. Then, given the recognition hypotheses, K CV models are made by excluding a transcript from one subset instead of making a single model using an adaptation method such as MLLR and MAP [5]. As an initial model to estimate the k-th CV model, the k-th CV model of the previous epoch is used. Each CV model is used in the next decoding step to make a new hypothesis for the data subset whose hypothesis has been excluded from the parameter estimation of that model.

After several iterations, a final recognition hypothesis is obtained by gathering the hypotheses of the K subsets made in the last decoding step. With this procedure, the data used for the decoding and for the model parameter estimation are effectively separated minimizing the undesired effects of reinforcing the errors and bias. The fragmentation problem is minimal for large K since

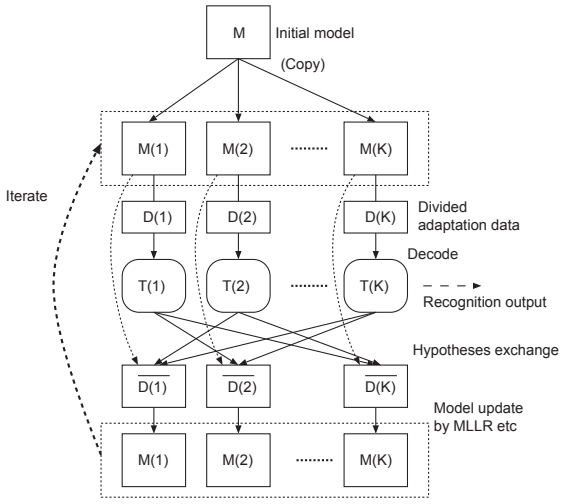

Fig. 1. Unsupervised cross-validation (CV) adaptation. M is a model, $D(k)$ is the k-th Data subset, T is a transcript, $\bar{D_k}$ is a data subset excluding $D(k)$.

$(K-1)/K$ of the data is used for the parameter estimation of each CV model. The computational cost of the decoding step is mostly constant for K but the cost for the update step is linear in K.

CV-adaptation is somewhat similar to cross-adaptation [6] when $K = 2$. The difference is that while cross-adaptation uses transcripts from different models representing different views of the same data, CV-adaptation uses the different data of the same view. That is, CV-adaptation is performed on a single recognition system whereas cross-adaptation requires two.

3. A VARIATION FOR EFFICIENT MLLR ADAPTATION

While the general version of the CV adaptation method is independent from the underlying adaptation method, the efficient algorithm is based on utilizing the details of the MLLR algorithm. In this section, we first overview the MLLR algorithm for mean transformation [1] and then propose an efficient version of unsupervised CV MLLR adaptation.

3.1. MLLR algorithm

In the MLLR adaptation, mean vectors of a set of Gaussian mixture HMMs are classified into M classes and a transformation shown in Equation (1) is estimated for each class so as to maximize the likelihood by the adapted HMM, where m is a class index, m_r is a Gaussian component index belonging to m-th class, $\boldsymbol{W_m}$ is a transformation matrix, $\boldsymbol{\xi_{m_r}} = [1,\ \boldsymbol{\mu}_{\boldsymbol{m_r}}^T]^T$ is an extended mean vector consisting of a constant term and an original mean vector $\boldsymbol{\mu_{m_r}}$, and $\boldsymbol{\mu'_{m_r}}$ is a transformed mean vector.

$$\boldsymbol{\mu'_{m_r}} = \boldsymbol{W_m}\boldsymbol{\xi_{m_r}}. \tag{1}$$

Given a set of adaptation utterances, the optimal transformation $\boldsymbol{W_m}$ is obtained by solving Equation (2).

$$\begin{aligned} &\sum_t \sum_{m_r} \gamma_{m_r}(t)\, \boldsymbol{\Sigma_{m_r}^{-1}}\boldsymbol{o}(t)\, \boldsymbol{\xi_{m_r}^T} \\ = &\sum_t \sum_{m_r} \gamma_{m_r}(t)\, \boldsymbol{\Sigma_{m_r}^{-1}}\boldsymbol{W_m}\boldsymbol{\xi_{m_r}}\boldsymbol{\xi_{m_r}^T}, \end{aligned} \tag{2}$$

where $\boldsymbol{\Sigma_{m_r}}$ is a covariance matrix of the m_r-th Gaussian component of the original model, $\boldsymbol{o}(t)$ is an observation vector at time t, and $\gamma_{m_r}(t) = P(q_{m_r}(t)\,|\lambda, \boldsymbol{O})$ is an occupation count of being at Gaussian mixture component q_{m_r} at time t given HMM model parameters λ and the observation sequence $\boldsymbol{O}$.

The transformation estimation using Equation (2) can be divided into two steps. The first step is an accumulation step expressed in Equation (3) and the second step is a transformation estimation step that solves Equation (4).

$$\begin{aligned} A^0_{m_r} &= \sum_t \gamma_{m_r}(t), \\ \boldsymbol{A^1_{m_r}} &= \sum_t \{\gamma_{m_r}(t)\, \boldsymbol{o}(t)\}. \end{aligned} \tag{3}$$

$$\begin{aligned} &\sum_{m_r} \boldsymbol{\Sigma_{m_r}^{-1}} \boldsymbol{A^1_{m_r}} \boldsymbol{\xi_{m_r}^T} \\ = &\sum_{m_r} A^0_{m_r} \boldsymbol{\Sigma_{m_r}^{-1}} \boldsymbol{W_m} \boldsymbol{\xi_{m_r}} \boldsymbol{\xi_{m_r}^T}. \end{aligned} \tag{4}$$

While the accumulation step requires summation over observation sequences and the computational cost is linear in the amount of data, the cost for transformation estimation step is constant. Therefore, for a large amount of data, the computational cost is dominated by the accumulation step.

3.2. Efficient unsupervised CV MLLR adaptation

Figure 2 shows the procedure of the proposed efficient unsupervised CV adaptation algorithm for MLLR. The differences from the general CV algorithm are that the MLLR model update procedure is split into two steps and the data exchange for the CV operation is performed between the two steps. That is, the MLLR statistics defined by Equation (3) are accumulated in the accumulation step for each CV subset using the recognition hypothesis of that subset and a corresponding CV model. Then, MLLR transforms for the k-th CV model are estimated in the estimation step described by Equation (4) by gathering all the statistics excluding the one for the k-th subset. The new k-th CV model is made by applying the estimated transforms to the k-th CV model of the previous epoch.

In this procedure, the computational cost for the MLLR accumulation step is constant for the number of CV folds K excepting the overhead of reading multiple models since each input utterance is processed only once while it is processed $K-1$ times using different CV models in the general unsupervised CV algorithm. Therefore, when the computational cost of MLLR is dominated by the accumulation step, the model update step of this efficient version works with only $1/(K-1)$ of the original cost.

4. EXPERIMENTAL SETUPS

Test set was the evaluation set of the Corpus of Spontaneous Japanese (CSJ) [7] that consisted of 10 academic presentations given by different male speakers. The length of each presentation is about 10 to 20 minutes and the total duration is 2.3 hours. The unsupervised speaker adaptations were performed for each of these presentations. Two types of initial models were used for the speaker adaptation experiments. One was an in-domain model that was trained using spontaneous speech from the CSJ corpus and the other was a cross-domain model that was trained using read speech from the Japanese News Article Sentences (JNAS) corpus [8]. Both of them were tied-state Gaussian mixture triphone HMM with 32 mixtures. The CSJ model had 3000 states and trained using 254

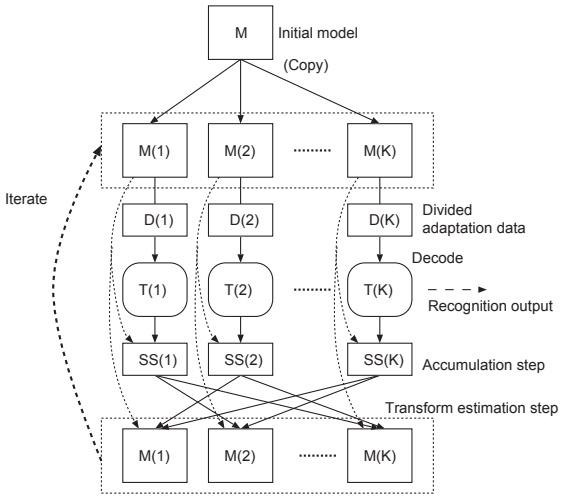

Fig. 2. Efficient unsupervised cross-validation (CV) adaptation for MLLR. M is a model, $D(k)$ is the k-th Data subset, T is a transcript, $SS(k)$ is a set of MLLR sufficient statistics.

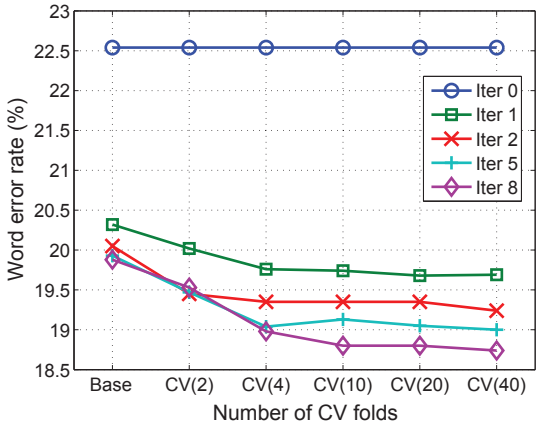

Fig. 3. Number of cross-validation folds (K) and recognition performance. The CV adaptation is by the general version of CV MLLR adaptation with the CSJ initial model. The zero-th iteration is the result of the speaker independent model. The batch-mode baseline adaptation result is denoted as Base.

hours of academic oral presentations whereas the JNAS model had 2000 states and trained from 52 hours of gender independent data. In addition to directly use these initial models for the unsupervised speaker adaptation, a combination of unsupervised domain and speaker adaptations was also evaluated in which the JNAS model was first adapted in an unsupervised manner using two hours of CSJ training data from 24 speakers and then unsupervised speaker adaptation was performed. The MLLR method was used for the speaker adaptation and the MAP method was used for the domain adaptation.

Feature vectors had 39 elements comprising 12 MFCCs and log energy, their delta, and delta delta values. The proposed efficient CV MLLR algorithm was implemented by extending the HTK toolkit [9]. In our current implementation, a common set of transformation classes were used over the CV models. This behavior is different from the general version of the CV adaptation where each CV model has its own transformation classes. The MLLR adaptation was performed using a regression class tree with 32 leaf nodes. The language model was a trigram model trained from 6.8M words of academic and extemporaneous presentations from CSJ and the dictionary size was 30k. The recognizer was the T^3 WFST decoder [10].

5. EXPERIMENTAL RESULTS

Figure 3 shows the unsupervised speaker adaptation results by the general CV adaptation method with the different number of CV folds K. The CSJ model was used as the initial model. The CV-adaptation gave lower word error rates than the batch-mode baseline adaptation with all the CV folds K. The best results were obtained when K was greater than around 10 to 20. This is because when K is small, the effective adaptation data is reduced for the model parameter estimation. As the value of K increases, stable results are obtained since $(K-1)/K$ of the data is used in the model parameter estimation. When K was equal or greater than four, the improvement by the CV adaptation from the baseline batch adaptation was statistically significant for all the iterations 1 to 8, respectively. In the following experiments, $K = 20$ was used.

Table 1 shows the results of unsupervised speaker adaptation using the CSJ and JNAS initial models and Table 2 summarizes their computational costs. When the CSJ model was used as an initial model, both the general and efficient version of the unsupervised CV MLLR adaptation gave mostly the same word error rates that were significantly lower than the baseline batch-mode adaptation. The computational cost consumed for the MLLR model update step in the efficient version of the CV method was about $1/3$ of that of the general CV method while their decoding cost was mostly the same. As the total, the computational cost of the efficient version of the CV method was only about 60% of that of the general version. The relative word error rate reductions by the baseline, general CV, and efficient CV adaptation methods after eight iterations were 12%, 17%, and 16%, respectively.

When the JNAS initial model was used, the initial word error rate was much higher than when the CSJ model was used. This was because of the mismatch in the training and test domain. Both the general and efficient version of the CV algorithms gave significantly better performance than the batch-mode baseline. When the general and efficient version of the CV algorithms are compared, the general version gave slightly better performance than the efficient version for the number of iterations larger than two. This was probably because the approximation introduced in the efficient version of the CV algorithm to compute the sufficient statistics and the different treatment of the transformation classes in the current implementation. The relative word error rate reductions by the baseline, general CV, and efficient CV adaptation methods after eight iterations were 18%, 26%, and 25%, respectively.

The unsupervised domain adaptation was performed by the conventional batch-mode MAP adaptation and the general version of the CV adaptation with the MAP method. The word error rates by the batch-mode and CV domain adaptations after five iterations were 31.2% and 30.4%, respectively. Table 3 shows the word error rates when the domain adapted speaker independent models were used as an initial model for the unsupervised speaker adaptations. The zero-th iteration is the results by the domain adapted speaker independent models. Both general and efficient CV methods significantly outperformed the baseline. The general CV method gave slightly better

Table 1. Word error rates of the CSJ test set. In the table, "CSJ" and "JNAS" indicate the CSJ and JNAS initial models, respectively. "Base" is the baseline conventional batch mode MLLR adaptation, "CV" is the general CV MLLR adaptation, and "ECV" is the efficient CV MLLR adaptation. The zero-th iteration is the result by the speaker independent initial model

Condition		# of iterations					
Init	Adpt	0	1	2	3	5	8
CSJ	Base	22.5	20.3	20.1	19.9	19.9	19.9
	CV	22.5	19.7	19.4	19.2	19.1	18.8
	ECV	22.5	19.7	19.3	19.2	19.1	19.0
JNAS	Base	34.7	30.0	29.3	29.0	28.7	28.5
	CV	34.7	28.5	27.1	26.4	26.1	25.6
	ECV	34.7	28.5	27.1	26.8	26.4	26.2

Table 2. Averaged computational costs of the decoding and model update steps in each adaptation epoch to process the CSJ test set. The cost is measured in hours

Init	Step	Base	CV	ECV
CSJ	Decode	3.9	4.6	4.8
	Update	0.4	7.8	2.6
	Total	4.4	12.4	7.4
JNAS	Decode	3.9	4.5	4.9
	Update	0.4	7.1	1.9
	Total	4.3	11.6	6.8

performance than the efficient version when the number of iterations was large. When the CV-MAP adapted initial model was used, the relative word error rate reductions by the unsupervised speaker adaptations using the batch-mode, general CV, and efficient CV methods were 13%, 19%, 18%, respectivelly. The relative word error rate reduction from the result of the domain independent JNAS model was 23% when the batch-mode domain and batch-mode speaker adaptations were combined, whereas the reductions by the combinations of CV domain and CV speaker adaptations were 29% and 28%, respectivelly, when the general and efficient versions of the CV speaker adaptations were used.

6. CONCLUSION

An unsupervised CV adaptation algorithm and its variation have been proposed that can be used as a substitute of the conventional batch-mode adaptation framework. The first algorithm is referred to as a general version and it can be used in combination with any adaptation method. The second algorithm is referred to as an efficient version and it works with lower computational cost than the general version by assuming MLLR. These CV algorithms have an ability to suppress the negative effect of the batch-mode process that reinforces the errors included in automatic transcription. Experimental results showed that the general version of the CV methods gives the highest recognition performance whereas the efficient version is advantageous for quick adaptation with reduced computational cost especially when the adaptation is iterated only once or twice.

Future work includes improving the performance of the efficient CV algorithm for the larger adaptation iterations and implementing efficient version of the CV MAP adaptation. Combinations with other acoustic model adaptation techniques and applications to other adaptation problems not limited to speech recognition are also interesting.

Table 3. Word error rates of the CSJ test set using the domain adapted model as an initial model. "Base" is the baseline conventional batch mode adaptation, "CV" is the general CV adaptation, and "ECV" is the efficient CV adaptation. The domain adaptation was performed by MAP and the speaker adaptation was performed by MLLR. The zero-th iteration is the results by the domain adapted initial models

Adaptation		# of iterations					
Domain	Spkr	0	1	2	3	5	8
Base	Base	31.2	27.6	27.1	27.0	26.8	26.7
CV	Base	30.4	27.1	26.6	26.5	26.3	26.3
CV	CV	30.4	26.1	25.4	25.3	24.9	24.7
CV	ECV	30.4	26.2	25.3	25.3	25.2	24.9

7. ACKNOWLEDGMENTS

This work was supported by KAKENHI (19700167).

8. REFERENCES

[1] C. J. Leggetter and P. C. Woodland, "Flexible speaker adaptation using maximum likelihood linear regression," in *Proc. Eurospeech*, 1995, pp. 1155–1158.

[2] M. Gales and S. Young, "The application of hidden Markov models in speech recognition," *Foundations and Trends in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2008.

[3] N. S. Kim and C. K. Un, "Deleted strategy for MMI-based HMM training," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 299–303, 1998.

[4] T. Shinozaki and M. Ostendorf, "Cross-validation and aggregated EM training for robust parameter estimation," *Computer speech and language*, vol. 22, no. 2, pp. 185–195, 2008.

[5] J.L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[6] H. Soltau, B. Kingsbury, L. Mangu, D. Povery, G. Saon, and G. Zweig, "The IBM 2004 conversational telephony system for rich transcription," in *Proc. ICASSP*, 2005, vol. I, pp. 205–208.

[7] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui, "Benchmark test for speech recognition using the Corpus of Spontaneous Japanese," in *Proc. SSPR2003*, 2003, pp. 135–138.

[8] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Acoust Soc Jpn E*, vol. 20, no. 3, pp. 199–206, 1999.

[9] S. Young *et al.*, *The HTK Book*, Cambridge University Engineering Department, 2005.

[10] P. R. Dixon, D. A. Caseiro, T. Oonishi, and S. Furui, "The TITech large vocabulary WFST speech recognition system," in *Proc. IEEE ASRU*, 2007, pp. 443–448.