論文 / 著書情報
Article / Book Information

Title	Investigations on Ensemble Based Unsupervised Adaptation Methods
Author	Yu Kubota, Takahiro Shinozaki, Sadaoki Furui
Journal/Book name	IEEE ICASSP2010, , , pp. 4874-4877
発行日 / Issue date	2010, 3
権利情報 / Copyright	(c)2010 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

INVESTIGATIONS ON ENSEMBLE BASED UNSUPERVISED ADAPTATION METHODS

Yu Kubota, Takahiro Shinozaki, Sadaoki Furui

Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Tokyo, Japan

ABSTRACT

We have previously proposed unsupervised cross-validation (CV) adaptation that introduces CV into an iterative unsupervised batch mode adaptation framework to suppress the influence of errors in an internally generated recognition hypothesis and have shown that it improves recognition performance. However, a limitation was that the experiments were performed using only a clean speech recognition task with a ML trained initial acoustic model. Another limitation was that only the CV method was investigated while there was a possibility of using other ensemble methods. In this study, we evaluate the CV method using a discriminatively trained baseline and a noisy speech recognition task. As an alternative to CV adaptation, unsupervised aggregated (Ag) adaptation is proposed and investigated that introduces a bagging like idea instead of CV. Experimental results show that CV and Ag adaptations consistently give larger improvements than the conventional batch adaptation but the former is more advantageous in terms of computational cost.

Index Terms— Cross-validation, bagging, machine learning ensemble, unsupervised adaptation, acoustic model

1. INTRODUCTION

Batch-type unsupervised adaptation is a useful technique to achieve high recognition performance without requiring any human transcribed adaptation data. It is generally performed by first running an automatic recognizer to derive a hypothesis for the target utterances and then a parameter estimation algorithm such as MLLR [1] is applied to update the model using that hypothesis. Based on the adapted model, this process is iterated for lower recognition error rates.

While it is effective, a problem is that the hypothesis made by the recognizer always includes errors. Since the model is updated using the hypothesis, it is likely that the same recognition error occurs in the next decoding step. Moreover, the negative effect is reinforced through the iteration as the same data is used for the model update and for the decoding.

Based on these observations, we previously proposed an unsupervised cross-validation (CV) adaptation algorithm to improve the generalization performance of the adaptation process [2]. The idea was to separate the data used in the decoding step and in the model update step by introducing the CV technique. In this way, the repetition of the same recognition error can be avoided since utterances used to estimate a model are not recognized by the decoder using that model in the next decoding step.

The unsupervised batch mode adaptation can be seen as a kind of EM algorithm with the viterbi approximation and parameter constraints. The expectation and the maximization steps of EM respectively correspond to the decoding and the model update steps of the unsupervised batch adaptation. Therefore, the unsupervised CV adaptation can be regarded as an extension of our previously proposed CV-EM that introduces CV into the iterative EM framework [3]. As an alternative for CV-EM, we have proposed Ag-EM that introduces a bagging like idea into the EM framework instead of CV and have shown that it gives better performance than CV-EM [4]. Similar to the relationship between CV-EM and unsupervised CV adaptation, it is possible to extend the Ag-EM to unsupervised adaptation. This results in an unsupervised aggregated (Ag) adaptation that we propose and investigate in this paper. Together with the CV adaptation method, we refer to these adaptation techniques as unsupervised ensemble adaptations. Compared to the traditional use of ensemble methods that directly improves evaluation performance, these methods differ in that the ensemble scheme is integrated inside of iterative parameter estimation process.

In the previous study, the unsupervised CV adaptation method was evaluated by a clean speech task using an EM trained maximum likelihood (ML) initial model. Here, the CV and Ag adaptation methods are evaluated using both ML and MPE [5] based systems. Moreover, they are evaluated for a noisy speech recognition task recorded in real car environments to investigate their performance on different tasks.

The organization of this paper is as follows. In section 2, we first briefly review the CV adaptation method and then describe the Ag adaptation algorithm. Experimental conditions are described in Section 3 and the results are shown in Section 4. Finally, conclusions and future works are given in Section 5.

2. UNSUPERVISED ENSEMBLE ADAPTATION ALGORITHMS

In this section we first review the CV adaptation algorithm and then explain the Ag adaptation method.

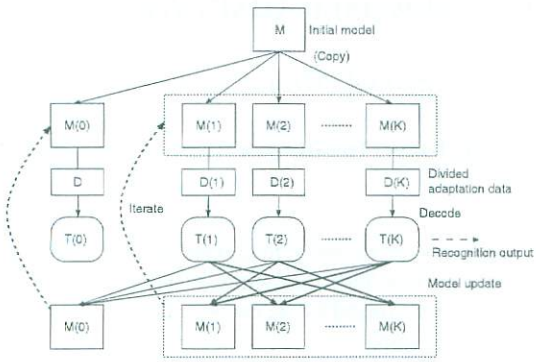

Fig. 1. Unsupervised cross-validation (CV) adaptation algorithm. $M(k)$ is the k-th CV model, $D(k)$ is the k-th exclusive subset of adaptation data, and $T(k)$ is a recognized hypothesis for $D(k)$ decoded by using $M(k)$. $M(0)$ is a global CV model made by gathering all the K hypotheses.

2.1. Cross-validation (CV) adaptation

Figure 1 shows the procedure of the unsupervised cross-validation (CV) adaptation method. In this procedure, the target utterances are divided into K exclusive subsets $(D(1), D(2), \cdots, D(K))$ so that each subset has roughly the same size. The first decoding step is basically the same as the conventional batch mode adaptation and the K subsets are processed using the same initial model. Then, given the K recognition hypotheses $(T(1), T(2) \cdots, T(K))$, K cross-validation models $(M(1), M(2), \cdots, M(K))$ are made by excluding one of the recognition hypotheses, instead of making a single model. As an initial model to estimate the k-th CV model, the k-th CV model of the previous stage is used. Each model is used in the next decoding step to make a new hypothesis for the data subset that has been excluded from the parameter estimation of that model. The decoding step and the model update step are repeated as in conventional batch mode adaptation and the final recognition hypothesis is obtained by gathering the hypotheses of the K subsets made in the last decoding step. If a single adapted model is required as an output of the adaptation process, a global CV model $(M(0))$ may be made in the last update step by using all recognition hypotheses.

With this procedure, the data used for the decoding and for the model parameter estimation are effectively separated. The data fragmentation problem is minimal for large K, since $(K-1)/K$ of the data is used for the parameter estimation of each CV model. The parameter update can be performed by using any kind of adaptation methods and not limited to a specific algorithm. The computational cost for the decoding step is constant for K excepting the overhead due to reading K different models. The computational cost for the update step is proportional to K.

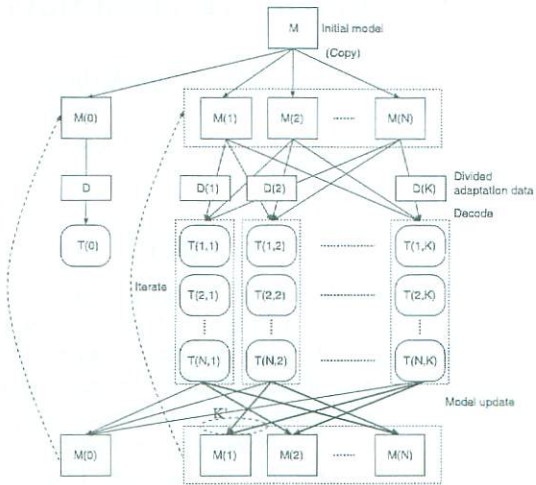

Fig. 2. Unsupervised aggregated (Ag) adaptation algorithm. $T(n,k)$ is a recognition hypothesis for the k-th subset made by using n-th Ag model $M(n)$. $M(0)$ is a global Ag model.

2.2. Aggregated (Ag) adaptation

Figure 2 describes the unsupervised aggregated (Ag) adaptation algorithm. Unlike the CV adaptation, Ag adaptation allows overlap between the data used in the decoding step and the update step. Instead, the generalization ability is obtained through aggregating N models as in the bagging method. More specifically, the target utterances are first divided into K exclusive subsets $(D(1), D(2), \cdots, D(K))$. Then, each data subset is repeatedly decoded using N models $(M(1), M(2), \cdots, M(N))$. Initially, these N models are prepared just by copying an initial model. In the update step, N models are made using NK' hypotheses from $K'(<K)$ subsets that are randomly selected. Depending on the underlying adaptation method, the observation counts may be normalized by N in the parameter estimation since N hypotheses from the same utterance are simultaneously used. The N models are then used in the next decoding step. The computational cost for the decoding step is $O(N)$ and the cost for the update step is $O\left(N^2K'/K\right)$.

In this procedure, N recognition hypotheses are generated for each utterance. In order to make a single output, the hypotheses were integrated by word level voting based on an alignment by the progressive multiple alignment method with UPGMA [6]. If necessary, a global Ag model $(M(0))$ may be made in the last update step by using all recognition hypotheses.

3. EXPERIMENTAL SETUPS

The unsupervised CV and Ag adaptation algorithms were evaluated as speaker adaptation methods using clean and noisy speech recognition tasks. The test set for the clean

speech task was the evaluation set of the Corpus of Spontaneous Japanese (CSJ) [7] that consisted of 10 academic presentations given by 10 different male speakers. The length of each presentation was about 10 to 20 minutes and the total duration was 2.3 hours. Unsupervised speaker adaptation was performed for each of these presentations and their word error rates were averaged. The acoustic model was a tied-state Gaussian mixture triphone HMM that was trained from academic oral presentations from the CSJ using the ML and MPE methods. The total length of the training set was 254 hours. The HMM had 3000 states and Gaussian mixture with 32 components per state. Feature vectors had 39 elements comprising 12 MFCCs and log energy, and their delta, and delta delta values. The language model was a trigram trained from 6.8M words of academic and extemporaneous presentations from CSJ. The dictionary size was 30k.

Noisy speech recognition was performed using speech FctCHom tJ G"DtivGU' CPCPGG5PGEJ CotpuUP CCCTEPvitpmCpt" EptpuU[8] givCP b[ptpHGUjopCNFtvGUCUCtGU set. St consisted of 20 male speakers and 20 female speakers. The utterances were voice commands to a car navigation system and there was a total of 108 utterances per speaker. They were recorded inside a car in idling mode, running in a city, or running on a highway. The total amount of data per speaker was six minutes. The speaker-independent initial acoustic model was a tied-state Gaussian mixture triphone HMM. St was first trained on 52 hours of clean speech data from the Japanese News Article Sentences (JNAS) corpus [9] that included both genders and then adapted to noisy speech conditions by using 1795 CSJ utterances that were randomly mixed with 28 types of noise from the JEgb A-NmgSE corpus that included car noises and seven different SNPs. The HMM had 2000 states and Gaussian mixture with 16 components per state. The dictionary size was 300 and a network grammar based language model was used. Feature vectors had 38 elements consisting of 12 MFCCs, their delta plus delta log energy, and delta delta values. Spectral subtraction was performed both in the estimation of the initial speaker independent noisy speech model and recognition of the evaluation data. The adaptation was performed for each speaker, and their word error rates were averaged.

Sn both of the conditions, the HTi toolkit [10] was used for the MLLP adaptation. The MLLP was based on mean transformations and was performed using regression class trees with 32 leaf nodes. For the Ct adaptation, the default threshold value of the toolkit was used to determine the number of transforms. For the Ag adaptation, the threshold was multiplied by N for the normalization purpose. The decoding was performed using the T^3 decoder [11].

4. EXPERIMENTAL RESULTS

Figure 3 shows the word error rates when the academic presentations were used as the test set. The Ct adaptation used

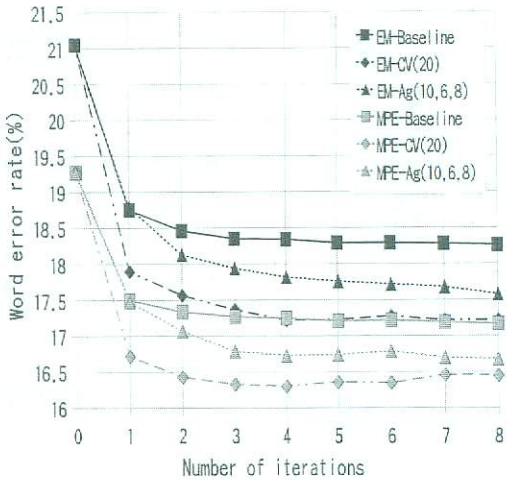

Fig. 3. Number of adaptation iterations and averaged word error rates for the CSJ test set using ML and MPE based initiCNmoFGU "BCUGNPG" iUbCUGNPGEPPvGPtioPCNbCtEJ-moFG adaptation. The zero-th iteration corresponds to the results of the speaker independent model.

$K = 20$, and the Ag adaptation used $K = 10, K' = 6, N = 8$ following the parameter settings of the Ag-EM experiments [4]. As a speaker independent initial model, ML and MPE trained acoustic models were used. Both Ct and Ag methods gave improvement over the conventional adaptation for both of the initial models. Among them, Ct adaptation gave better performance than Ag adaptation in this task. The word error rate by the speaker independent ML initial model was 21.1C and the relative word error reductions by the baseline batch mode, Ct, and Ag adaptations were 13C, 18C, and 16C, respectively. Similarly, the word error rate by the MPE initial model was 19.3C and the relative word error reductions by the baseline, Ct, and Ag adaptations were 11C, 15C, and 13C, respectively. The improvements by the Ct and Ag adaptations from the baseline adaptation were both statistically significant for both of the ML and MPE conditions.

Figure 4 is the result of the speaker adaptation using the utterances from the real car environment. St can be seen that Ct and Ag adaptations gave similar improvements from the conventional batch mode baseline adaptation. A slight increase of the error rate was observed for the Ct adaptation when the number of iterations was larger than four, which was probably due to an over-training. This is because while Ct adaptation separates the data used for the decoding and model update steps, small dependencies still remain between the i-th model update step and the $(i+2)$-th decoding step through recognition hypotheses of the $(i+1)$-th step. The initial word error rate was 13.4C, and the relative word error rate reductions by the baseline, Ct, and Ag adaptations after eight iterations were 4.5C, 8.3C, and 9.3C, respectively.

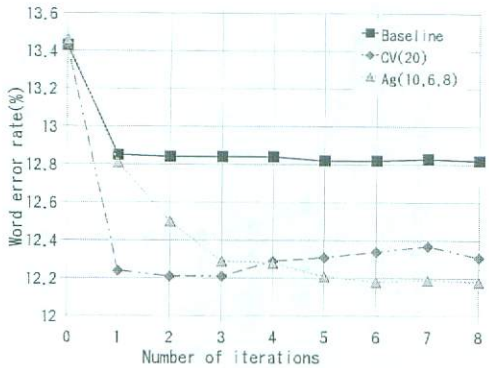

Fig. 4. Adaptation results using noisy speech from real car environments.

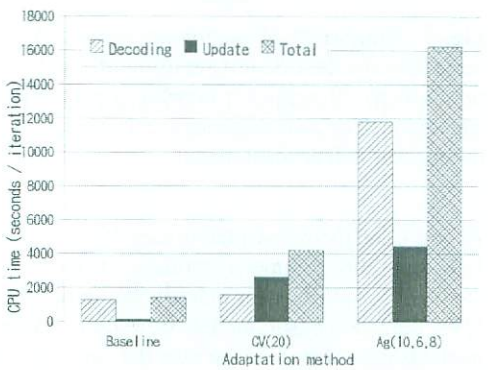

Fig. 5. CPU time in each adaptation method.

Figure 5 shows CPU time observed in the academic presentation recognition using the ML initial model. The CPU times are per speaker and are averaged over all the decoding steps and the update steps. As mentioned in Section 2, CV adaptation has roughly the same computational cost for the decoding step as the baseline conventional batch mode adaptation. The cost for the update step is proportional to K but because adaptation is cheaper than decoding, the total cost of the 20-fold CV adaptation was about only three times of the baseline adaptation. On the other hand, the computational cost of the Ag adaptation is generally higher than the baseline adaptation in both of the decoding and update steps.

5. CONCLUSION

We have evaluated the unsupervised CV adaptation method together with newly proposed unsupervised Ag adaptation using both ML and MPE trained initial models for the clean speech recognition task form the Corpus of Spontaneous Japanese (CSJ) and using a noisy speech recognition task recorded in real car environments. Experimental results showed that both of the ensemble adaptation methods give consistently higher recognition performance than the conventional batch mode adaptation method. Among them, CV adaptation was more advantageous than Ag adaptation giving similar or better improvements with smaller computational cost. Future work includes improving the ensemble adaptation methods utilizing confidence measures, and their applications to other problems not limited to speech recognition.

6. ACKNOWLEDGMENTS

This study was partly supported by a Microsoft CORE5 project selected by the MS Institute for Japanese Academic Research Collaboration. The "Drivers' Japanese Speech Corpus in Car Environment" was used by Asahi Kasei Corp. under the "Development of Fundamental Speech Recognition Technology" project supported by the Japanese Ministry of Economy, Trade and Industry. We would like to thank Asahi Kasei Corp. for letting us use the corpus.

7. 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] T. Kinoshita, Y. Tsubota and T. Kawahara, "Unsupervised cross-validation adaptation algorithms for improved adaptation performance," *ICASSP2009*, P00W

[3] T. Kinoshita and M. Ostendorf, "Cross-validation and hierarchical EM training for robust recognition of street speech," *Computer speech and language*, vol. 22, no. 2, pp. 185–195, 2008.

[4] T. Kinoshita and T. Shinozaki, "GMM and HMM training by aggregated EM algorithm with increased ensemble sizes for robust recognition of street speech," in *Proc. ICASSP*, 2008, pp. 4405–4408.

[5] D. Povey and P.C. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *Proc. ICASSP*, 2002, vol. I, pp. 105–108.

[6] D.F. Feng and R. F. Doolittle, "Progressive sequence alignment as a prerequisite to correct phylogenetic trees," *Journal of Molecular Evolution*, 1987.

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

[8] T. Kato, J. Ogata and M. Kozaki, "Analysis of speech at automobile driving and analysis of interaction between the recognizer and the user," in *Autumn Meeting of the Acoustical Society of Japan*, 2000, vol. 3-Q-25, pp. 267–268, in Japanese.

[9] A. Ito, M. Yamamoto, A. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JEIDA-50: Japanese speech corpus for large vocabulary continuous speech recognition research," *Acoust Soc Jpn E*, vol. 20, no. 3, pp. 199–206, 1999.

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

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